SIGIR 2026 · Conference Program

The 49th International ACM SIGIR Conference · Melbourne, 20–24 July 2026

Schedule Overview

Monday — 20 July · Tutorials and Doctoral Colloquium

Time	Exhibition 21B/22B	Eureka Foyer	Eureka 1	Eureka 2	Eureka 3	Room 111	Room 112	Courtyard 1	Courtyard 2
09:00 – 10:30			Bridging Personalization and AI: From RAG to Agent	LLM Personalization: Foundations, Breakthroughs, and Frontiers	Multi-Agentic Recommender Systems: Foundations, Perspectives, and Lessons from Large Scale Deployments in eCommerce	Retrieve, Rerank, Answer, Experiment: Hands-On IR Research with PyTerrier	Temporal Information Retrieval and Extraction: From Foundations to RAG	MANILA26: SIGIR 2026 Tutorial on Information Retrieval for Climate Change Impact	Doctoral Colloquium
10:30 – 11:00		Coffee Break							
11:00 – 12:30			Bridging Personalization and AI: From RAG to Agent	LLM Personalization: Foundations, Breakthroughs, and Frontiers	Multi-Agentic Recommender Systems: Foundations, Perspectives, and Lessons from Large Scale Deployments in eCommerce	Retrieve, Rerank, Answer, Experiment: Hands-On IR Research with PyTerrier	Temporal Information Retrieval and Extraction: From Foundations to RAG	MANILA26: SIGIR 2026 Tutorial on Information Retrieval for Climate Change Impact	Doctoral Colloquium
12:30 – 13:30		Lunch							
13:30 – 15:00			Bridging Personalization and AI: From RAG to Agent	Reasoning for IR & IR for Reasoning	Beyond Relevance: Utility-Centric Retrieval in the LLM Era	Verifiable User Simulation for Search and Recommendation Systems	Trustworthy Personalized Retrieval in the LLM Era: Fairness, Bias Awareness, Privacy, and Multimodality	Transdisciplinarity for Human Information Interaction and Retrieval: Strategies for Collaborating Across Research Disciplines to Foster Societal Impact	Doctoral Colloquium
15:00 – 15:30		Coffee Break							
15:30 – 17:00			Bridging Personalization and AI: From RAG to Agent	Reasoning for IR & IR for Reasoning	Beyond Relevance: Utility-Centric Retrieval in the LLM Era	Verifiable User Simulation for Search and Recommendation Systems	Trustworthy Personalized Retrieval in the LLM Era: Fairness, Bias Awareness, Privacy, and Multimodality	Transdisciplinarity for Human Information Interaction and Retrieval: Strategies for Collaborating Across Research Disciplines to Foster Societal Impact	Doctoral Colloquium
17:30 – 19:30	Welcome Reception								

Tuesday — 21 July (Day 1)

Time	Exhibition 21B/22B	Goldfields	Room 110	Eureka 1	Eureka 2	Eureka 3	Courtyard 1+2
09:00 – 09:30		Opening · Welcome to Country					
09:30 – 10:30		Keynote — Bhaskar Mitra					
10:30 – 11:00	Coffee Break						

Time	Exhibition 21B/22B	Goldfields	Room 110	Eureka 1	Eureka 2	Eureka 3	Courtyard 1+2
11:00 – 12:30		Diffusion- and Flow-based Recommendation	Personalized Recommendation and Preference Modeling	Domain-Specific and Structured IR Tasks	Fact Checking and Entity Resolution	Reranking	RAG Systems
12:30 – 13:30	Lunch						
13:30 – 15:00		Multimodal Recommendation	Reasoning for Recommendation	Privacy, Security, and Robustness in IR	Video Retrieval	Search Behavior	Multi-Vector Retrieval
15:00 – 16:30	Poster Session 1 · Coffee Break						
16:30 – 17:45		LLMs for Recommendation	Graph Representation Learning	Dense Retrieval	Social Media Analysis	Fairness and Responsible IR	Faithfulness and Robustness in RAG
18:00 – 20:00	Student Reception (Students Only) · The General Assembly						

Wednesday — 22 July (Day 2)

Time	Exhibition 21B/22B	Goldfields	Room 110	Eureka 1	Eureka 2	Eureka 3	Courtyard 1+2
09:00 – 10:00		Keynote — Cecile Paris					
10:00 – 10:30	Coffee Break						
10:30 – 12:00		Context-Aware Recommendation	CTR Prediction	Item Tokenization	Misinformation and Bias in Search	LLM-based Evaluation and Relevance Assessment	Efficient Retrieval, Training, and Indexing
12:00 – 13:00	Lunch · Women in IR Posters						
13:00 – 14:30		Graph Collaborative Filtering	Sequential Recommendation	Complex Reasoning and QA	Conversational and Personalized Information Seeking	Cross-Modal Retrieval	Generative and LLM-based Retrieval
14:30 – 16:00	Poster Session 2 · Coffee Break						
16:00 – 17:30		LLMs for Sequential Recommendation	Agentic Search	Computational Advertising and Large-Scale Serving	Adaptive Retrieval for Reasoning	Learning to Rank	Low-Resource IR
18:30 – 21:30	Gala Dinner · MCEC Level 2, Melbourne Rooms						

Thursday — 23 July (Day 3)

Time	Exhibition 21B/22B	Goldfields	Room 110	Eureka 1	Eureka 2	Eureka 3	Courtyard 1+2
09:00 – 10:00		Keynote — Ji-Rong Wen					
10:00 – 10:30	Coffee Break						
10:30 – 12:00		Industry Session I	Cross-Domain Recommendation	Multimodal Representation Learning	Explainability, Alignment, and Debiasing	Query Reformulation and Tool Use	Evaluation Methodology
12:00 – 13:30	Lunch — Grab a Box Lunch	SIGIR Business Meeting — Bring a Box Lunch					
13:30 – 15:00		Industry Session II	Contrastive Collaborative Filtering	Knowledge Graph Reasoning	Retrieval and Understanding of Visual Documents and Video	Applications and Low-Resource IR	Sparse Retrieval
15:00 – 16:15	Industry Poster Session · Coffee Break						

Time	Exhibition 21B/22B	Goldfields	Room 110	Eureka 1	Eureka 2	Eureka 3	Courtyard 1+2
16:15 – 17:30		**Industry Panel**	Generative Recommendation	Community and Subgraph Search	Multimodal RAG	Query Performance and User Interaction Models	Low-Resource and Educational IR
17:30 – 17:45		**Conference Close**					

Friday — 24 July · Workshops

Time	Exhibition 21B/22B	Eureka 1	Eureka 2	Eureka 3	Room 101	Room 102	Room 111	Room 112	Courtyard 1	Courtyard 2
09:00 – 10:30		AgentSearch: Indexing, Retrieval, and Ranking of AI Agents	Second Workshop on Explainability in Information Retrieval	The 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'26)	ReNeuIR at SIGIR 2026: The Fifth Workshop on Reaching Efficiency in Neural Information Retrieval	SIGIR 2026 Workshop on eCommerce (ECOM26)	The Second Workshop on Evaluation of Multimodal Generation	LLM-UP: SIGIR 2026 Workshop on LLM-powered User Profiling for Search and Recommendation	International Workshop on Vulnerabilities in Generative Systems for Information Retrieval (VulGen'26)	Justice, Emancipation, Democracy, and Information Access (JEDI): The SIGIR Workshop on Resisting Corporate and Authoritarian Capture of Information Access Platforms
10:30 – 11:00	**Coffee Break**									
11:00 – 12:30		AgentSearch: Indexing, Retrieval, and Ranking of AI Agents	Second Workshop on Explainability in Information Retrieval	The 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'26)	ReNeuIR at SIGIR 2026: The Fifth Workshop on Reaching Efficiency in Neural Information Retrieval	SIGIR 2026 Workshop on eCommerce (ECOM26)	The Second Workshop on Evaluation of Multimodal Generation	LLM-UP: SIGIR 2026 Workshop on LLM-powered User Profiling for Search and Recommendation	International Workshop on Vulnerabilities in Generative Systems for Information Retrieval (VulGen'26)	Justice, Emancipation, Democracy, and Information Access (JEDI): The SIGIR Workshop on Resisting Corporate and Authoritarian Capture of Information Access Platforms
12:30 – 13:30	**Lunch**									
13:30 – 15:00		AgentSearch: Indexing, Retrieval, and Ranking of AI Agents	Second Workshop on Explainability in Information Retrieval	The 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'26)	ReNeuIR at SIGIR 2026: The Fifth Workshop on Reaching Efficiency in Neural Information Retrieval	SIGIR 2026 Workshop on eCommerce (ECOM26)	The Second Workshop on Evaluation of Multimodal Generation	SynthIR: The First Workshop on Synthetic Content in Information Retrieval Ecosystems	International Workshop on Vulnerabilities in Generative Systems for Information Retrieval (VulGen'26)	Justice, Emancipation, Democracy, and Information Access (JEDI): The SIGIR Workshop on Resisting Corporate and Authoritarian Capture of Information Access Platforms
15:00 – 15:30	**Coffee Break**									
15:30 – 17:00		AgentSearch: Indexing, Retrieval, and Ranking of AI Agents	Second Workshop on Explainability in Information Retrieval	The 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'26)	ReNeuIR at SIGIR 2026: The Fifth Workshop on Reaching Efficiency in Neural Information Retrieval	SIGIR 2026 Workshop on eCommerce (ECOM26)	The Second Workshop on Evaluation of Multimodal Generation	SynthIR: The First Workshop on Synthetic Content in Information Retrieval Ecosystems	International Workshop on Vulnerabilities in Generative Systems for Information Retrieval (VulGen'26)	Justice, Emancipation, Democracy, and Information Access (JEDI): The SIGIR Workshop on Resisting Corporate and Authoritarian Capture of Information Access Platforms

Time	Exhibition 21B/22B	Eureka 1	Eureka 2	Eureka 3	Room 101	Room 102	Room 111	Room 112	Courtyard 1	Courtyard 2
18:30 – 21:30	ICTIR Dinner (ICTIR Attendees Only) · Rivers Edge Events									

Oral Sessions

Diffusion- and Flow-based Recommendation ·

Goldfields · Tuesday 11:00–12:30

Dual-Diffusional Generative Fashion Recommendation [Full]

Mingzhe Yu (Singapore Management University); Lei Wu (Shandong University); Qianru Sun (Singapore Management University); Yunshan Ma (Singapore Management University)

Personalized generative recommender systems have emerged as a promising solution for fashion recommendation. However, existing methods primarily rely on implicit visual embeddings from historical interactions, which often contain preference-irrelevant information and result in insufficient user behavior modeling. Moreover, these models typically generate only item images, providing limited interpretability. To address these limitations, we propose DualFashion, a Dual-Diffusional Generative Fashion Recommendation Architecture that jointly models image and text modalities for personalized and explainable recommendation. DualFashion adopts a dual-diffusion Transformer with image and text branches, where structured attribute-level captions and visual outfit information are jointly used as conditioning signals to model user behavior. The proposed architecture produces both fashion item images and textual descriptions, ensuring visual compatibility while providing explicit semantic interpretability. Furthermore, we introduce a text-augmented fine-tuning strategy that enhances generation diversity and enables effective cross-modal knowledge transfer without incurring heavy computational costs. Extensive experiments on iFashion and Polyvore-U across Personalized Fill-in-the-Blank and Generative Outfit Recommendation tasks demonstrate that DualFashion achieves strong performance in behavior modeling, interpretability, and efficiency compared to state-of-the-art methods. Our code and model checkpoints are available at https://github.com/LinkMingzhe/DualFashion.

Beyond Static Diffusion: Explicitly Modeling Temporal Patterns in Sequential Recommendation [Full]

Yao Wu (The Hong Kong Polytechnic University); Chengyi Liu (The Hong Kong Polytechnic University); Wenqi Fan (The Hong Kong Polytechnic University); Rui Zhang (Huazhong University of Science and Technology)

Sequential recommendation predicts the next items a user will interact with by modeling evolving preferences over time. Recent diffusion-based generative recommenders show promise in capturing complex dependencies, but they typically treat temporal context as an external conditioning signal rather than integrating temporal transitions into the diffusion dynamics. In this paper, we introduce TDRec (Temporally-aware Diffusion for sequential Recommendation), a novel framework that integrates temporal progression into both forward and reverse processes: at each diffusion step, a position's latent is updated by noise injection and by mixing with its preceding latent. We derive a closed-form solution for this temporal mixing process, proving that it allows for efficient parallel training with O(1) complexity relative to sequence length. Furthermore, we establish the existence of a corresponding DDPM-like reverse process and a reparameterized objective, ensuring efficient optimization and sampling without incurring extra computational overhead. Empirical results on three public datasets demonstrate that TDRec consistently outperforms state-of-the-art baselines, including recent diffusion models. Ablation studies confirm the effectiveness of the temporal scheduler and sequence-reduction module in generating coherent, context-aware predictions. Code is available at https://github.com/wuyaoericyy/TDRec.

Countering Interest Over-Smoothing: Distilling Latent Factors via Diffusion for Multi-Interest Retrieval [Full]

Yankun Le (Northeastern University); Fu Zhang (Northeastern University); Haoran Li (Xiaohongshu Inc.); Baoyuan Ou (Xiaohongshu Inc.); Yingjie Qin (Xiaohongshu Inc.); Zhixuan Yang (Northeastern University); Ruilong Su (Xiaohongshu Inc.)

Multi-interest recommendation is essential for the matching stage. By generating multiple user representations, it can better cover the diverse interests derived from user interaction history. Ideally, multi-interest models should effectively identify the underlying latent factors — the specific themes, intents, or preferences — within historical behaviors. However, conventional methods typically rely on weighted aggregation (e.g., Attention), which we argue leads to over-smoothed representations. This aggregation dilutes the intensity of significant patterns that appear only locally, blending them into a blurry average. To address this, we propose DMI, a model-agnostic diffusion framework that distills precise interests by amplifying co-occurring latent factors across behaviors. Distinct from prior diffusion works that reconstruct the single next item—which risks collapsing diverse interests—DMI reconstructs the interest vectors themselves to preserve their distributional independence. To support this, we introduce a cross-transformer module that adaptively extracts interest-specific information from designated historical interactions, transforming the diffusion process from an unconditional one into a guided, interest-disentangled pathway. In addition, we design a gradient back-propagation strategy to decouple the joint optimization of the reconstruction and recommendation losses, thereby improving training stability. Extensive offline experiments demonstrate DMI's superiority over existing methods, achieving an average relative improvement of 11.2% across all metrics on Amazon Books datasets while increasing recommendation diversity by 11.8%. Successfully deployed in a real-world recommender system, DMI effectively enhances user satisfaction and system performance at scale, serving the major traffic of hundreds of millions of daily active users.

GFlowGR: Fine-tuning Generative Recommendation Frameworks with Generative Flow Networks [Full]

Yejing Wang (City University of Hong Kong); Shengyu Zhou (Alibaba Group); Jinyu Lu (Alibaba Group); Qidong Liu (City University of Hong Kong); Xinhang Li (Alibaba Group); Wenlin Zhang (City University of Hong Kong); Feng Li (Alibaba Group); Pengjie Wang (Alibaba Group); Chuan Yu (Alibaba Group); Jian Xu (Alibaba Group); Bo Zheng (Alibaba Group); Xiangyu Zhao (City University of Hong Kong)

Generative recommendation (GR) has shown great promise in industrial applications, particularly for candidate generation and end-to-end recommendations. However, existing GR training paradigms suffer from two fundamental mismatches with real-world deployment requirements. First, they optimize for point-wise prediction of a single ground-truth item, whereas practical systems must produce a diverse, high-value set of candidates. Second, they treat all user interactions as equally informative, ignoring their inherent differences in utility. Although reward-based fine-tuning offers a partial remedy, it often lacks token-level supervision. To address these challenges, we reformulate GR as a sequential set-generation problem and propose GFlowGR, a GFlowNet-based fine-tuning framework that explicitly aligns generation probabilities with item-level utilities. GFlowGR comprises three tightly integrated components, each addressing a key limitation of conventional fine-tuning: a trajectory sampler that constructs training trajectories from candidate sets to enable set-wise learning, a behavior-aware reward model that quantifies item utility to support value-aware optimization, and a GFlowNet objective that provides token-level supervision. Extensive experiments on three real-world datasets with two representative LLM-based GR backbones show consistent and significant improvements over strong baselines, validating the effectiveness of our approach. For real-world deployment, GFlowGR has been integrated into Taobao 's search advertising businesses, delivering a 0.4% relative improvement in annual revenue since its launch in mid-2025, corresponding to billion-level monetary gains. Code is available at https://github.com/Applied-Machine-Learning-Lab/SIGIR26_GFlowGR.

FAVE: Flow-based Average Velocity Establishment for Sequential Recommendation [Full]

Ke Shi (University of Electronic Science and Technology of China); Yao Zhang (University of Electronic Science and Technology of China); Feng Guo (University of Electronic Science and Technology of China); Jinyuan Zhang (University of Electronic Science and Technology of China); JunShuo Zhang (University of Electronic Science and Technology of China); Shen Gao (University of Electronic Science and Technology of China); Shuo Shang (University of Electronic Science and Technology of China)

Generative recommendation has emerged as a transformative paradigm for capturing the dynamic evolution of user intents in sequential recommendation. While flow-based methods improve the efficiency of diffusion models, they remain hindered by the "Noise-to-Data" paradigm, which introduces two critical inefficiencies: prior mismatch, where generation starts from uninformative noise, forcing a lengthy recovery trajectory; and linear redundancy, where iterative solvers waste computation on modeling deterministic preference transitions. To address these limitations, we propose a Flow-based Average Velocity Establishment (Fave) framework for one-step generation recommendation that learns a direct trajectory from an informative prior to the target distribution. Fave is structured via a progressive two-stage training strategy. In Stage 1, we establish a stable preference space through dual-end semantic alignment, applying constraints at both the source (user history) and target (next item) to prevent representation collapse. In Stage 2, we directly resolve the efficiency bottlenecks by introducing a semantic anchor prior, which initializes the flow with a masked embedding from the user's interaction history, providing an informative starting point. Then we learn a global average velocity, consolidating the multi-step trajectory into a single displacement vector, and enforce trajectory straightness via a JVP-based consistency constraint to ensure one-step generation. Extensive experiments on three benchmarks demonstrate that Fave not only achieves state-of-the-art recommendation performance but also delivers an order-of-magnitude improvement in inference efficiency, making it practical for latency-sensitive scenarios. Code is available at https://github.com/Blue130/Fave

Advantage-Conditioned Flow Policy for Offline Reinforcement Learning in Recommendation [Full]

Xiaocong Chen (Data 61, CSIRO); Siyu Wang (Macquarie University); Lina Yao (The University of New South Wales)

Offline reinforcement learning (RL) is a useful approach for recommender systems because it can optimize long-term user feedback from logged interaction data without online exploration. A key challenge is the multi-modal nature of user preferences: a user may like several unrelated item types, so a unimodal policy (for example, a Gaussian) tends to average across modes and generate actions that do not match any interest. Recent diffusion-based policies can model complex preference distributions, but they often require many denoising steps. We propose PerfRec (Preference-aware Flow for Recommendation), a flow-matching offline RL framework that learns an expressive behavioral policy and distills it into an efficient one-step policy. PerfRec (i) trains a conditional flow model to clone the logged action distribution, (ii) trains twin Q-networks using next actions sampled from the learned flow policy, and (iii) trains an advantage-conditioned one-step policy with Q-guidance for improvement and a distillation loss that keeps the policy close to the flow policy. We use binary advantage conditioning to separate high-advantage and low-advantage regions of the flow-induced action

distribution, so that at inference we can sample from the high-advantage mode with a single forward pass. Experiments on five benchmark datasets and one online simulation platform show that PerfRec improves recommendation performance over strong offline RL baselines.

Personalized Recommendation and Preference Modeling · *Room 110 · Tuesday 11:00–12:30*

When and How to Ask: Dynamic Preference Elicitation Strategies for Conversational Recommendation [Full]

Feng Xia (University of Sheffield); Shuo Zhang (Bloomberg); Xi Wang (University of Sheffield)

Conversational Recommender Systems (CRSs) are interactive systems that use multi-turn natural language dialogue to understand evolving user preferences and provide personalized recommendations. To achieve this goal, CRSs rely on preference elicitation strategies to actively gather informative preference cues from users; however, the timing and selection of these strategies during a conversation remain largely unexplored. While many existing studies emphasize eliciting explicit item attributes and tend to adopt relatively static elicitation strategies, the use of item-based preference elicitation and how it varies across different dialogue stages remains less explored. In this work, we conduct a systematic investigation of preference elicitation strategies from a stage-aware perspective. We provide empirical evidence that optimal preference elicitation strategies are stage-dependent and context-sensitive: attribute-based inquiries are effective in early stages, while item-based strategies become superior as preferences refine. To support this paradigm, we introduce InPE, a dataset enriched with fine-grained annotations for elicitation necessity and strategy selection. With this dataset, we propose COPE (COnversational Preference Elicitation via Mixture of Experts), a novel architecture for strategy modeling. Extensive offline evaluation on our dataset indicates that context-aware preference elicitation strategies are beneficial for conversational recommendation. In addition, the analysis of the predicted strategies uncovers consistent stage-wise tendencies in dialogue progression, providing empirical evidence of common interaction patterns in conversational recommendation systems. Our dataset is available at https://github.com/juanfacabian/InPE.

A Standardized Re-evaluation of Conversational Recommender Systems on the ReDial Dataset [Reproducibility]

Ivica Kostric (University of Stavanger); Krisztian Balog (University of Stavanger)

Recent years have seen a surge of research into conversational recommender systems (CRS). Among existing datasets, ReDial is the most widely used benchmark, cited in hundreds of studies. However, variations in how the dataset is preprocessed and used in experiments, particularly in the definition of ground-truth items, make it difficult to compare results across studies. These comparisons are further complicated by confounding factors such as the choice of the underlying large language model (LLM) and the use of external data sources. In this work, we revisit seven prominent CRS methods across three architectural families and evaluate them under standardized conditions. Our reproducibility study reveals a "granularity gap," where fine-grained ranking (Recall@1) is highly sensitive to implementation details, while our replicability analysis shows that nearly 50% of reported accuracy stems from "repetition shortcuts" that are absent in novelty-focused evaluation. Furthermore, we find that performance gains are often driven more by the capacity of the LLM backbone than by specific architectural innovations. Finally, by applying user-centric utility metrics, we demonstrate that traditional recall frequently overstates a system's actual conversational effectiveness. This work establishes a transparent, controlled baseline and promotes evaluation practices that prioritize novelty and interaction efficiency.

Learning from Natural Language Feedback for Personalized Question Answering [Full]

Alireza Salemi (University of Massachusetts Amherst); Hamed Zamani (University of Massachusetts Amherst)

Personalization is crucial for enhancing both the effectiveness and user satisfaction of language technologies, particularly in information-seeking tasks like question answering. Current approaches for personalizing large language models (LLMs) often rely on retrieval-augmented generation (RAG), followed by reinforcement learning with scalar reward signals to teach models how to use retrieved personal context. We believe that these scalar rewards sometimes provide weak, non-instructive feedback, limiting learning efficiency and personalization quality. We introduce VAC, a novel framework for personalized response generation that replaces scalar rewards with natural language feedback (NLF) that are generated conditioned on the user profiles and the question narratives. NLF serves as a rich and actionable supervision signal, allowing the policy model to iteratively refine its outputs and internalize effective personalization strategies. Training alternates between optimizing the feedback model and fine-tuning the policy model on the improved responses, resulting in a policy model that no longer requires feedback at inference. Evaluation on the LaMP-QA benchmark that consists of three diverse domains demonstrates consistent and significant improvements over the state-of-the-art results. Human evaluations further confirm the superior quality of the generated responses. These results demonstrate that NLF provides more effective signals for optimizing personalized question answering.

One Adapts to Any: Meta Reward Modeling for Personalized LLM Alignment [Full]

Hongru Cai (The Hong Kong Polytechnic University); Yongqi Li (The Hong Kong Polytechnic University); Tiezheng Yu (Huawei Technologies Ltd.); Fengbin Zhu (National University of Singapore); Wenjie Wang (University of Science and Technology of China); Fuli Feng (University of Science and Technology of China); Wenjie Li (The Hong Kong Polytechnic University)

Alignment of Large Language Models (LLMs) aims to align outputs with human preferences, and personalized alignment further adapts models to individual users. This relies on personalized reward models that capture user-specific preferences and automatically provide individualized feedback. However, developing these models faces two critical challenges: the scarcity of feedback from individual users and the need for efficient adaptation to unseen users. We argue that addressing these constraints requires a paradigm shift from fitting static user models to "learning to learn" adaptation. To realize this, we propose Meta Reward Modeling (MRM), which reformulates personalized reward modeling as a meta-learning problem. Specifically, we represent each user's reward model as a weighted combination of base reward functions, and optimize the initialization of these weights using a Model-Agnostic Meta-Learning (MAML)-style framework to support fast adaptation under limited feedback. To ensure robustness, we introduce the Robust Personalization Objective (RPO), which places greater emphasis on hard-to-learn users during meta optimization. Extensive experiments on personalized preference datasets validate that MRM enhances few-shot personalization, improves user robustness, and consistently outperforms baselines. We release code at https://github.com/ModalityDance/MRM.

ProMax: Exploring the Potential of LLM-derived Profiles with Distribution Shaping for Recommender Systems [Full]

Yi Zhang (Anhui University); Yiwen Zhang (Anhui University); Kai Zheng (University of Electronic Science and Technology of China); Tong Chen (The University of Queensland); Hongzhi Yin (University of Queensland)

The remarkable text understanding and generation capabilities of large language models (LLMs) have revitalized the field of general recommendation based on implicit user feedback. Rather than deploying LLMs directly as recommendation models, a more flexible paradigm leverages their ability to interpret users' historical interactions and semantic contexts to extract structured profiles that characterize user preferences. These profiles can be further transformed into actionable high-dimensional representations, serving as powerful signals to augment and strengthen recommendation models. However, the mechanism by which such profiles enhance recommendation performance within the feature space remains insufficiently understood. Moreover, existing studies predominantly rely on nonlinear alignment and fusion strategies to incorporate these profiles, which often lead to semantic loss and fail to fully exploit their potential. To address these limitations, we revisit profiles from a retrieval perspective and propose a simple yet effective recommendation framework built upon distribution shaping (ProMax) in this paper. We begin by employing dense retrieval to uncover the collaborative relationships between user and item profiles within the feature space. Based on this insight, we introduce a dual distribution-reshaping process, in which the profile distribution acts as a guiding signal to steer the recommendation model toward learning user preferences for unseen items beyond the scope of observed interactions. We apply ProMax to four classic recommendation methods on three public datasets. The results indicate that ProMax substantially improves base model performance and outperforms existing LLM-based recommendation approaches.

A Reproducibility Study of Bundle Editing and Bundle Recommendation [Reproducibility]

Yiran An (Wuhan University of Technology); Lin Li (Wuhan University of Technology); Ming Li (York University); Wenxin Ye (Wuhan University of Technology); Qing Xie (Wuhan University of Technology); Jimmy Xiangji Huang (York University)

\beginabstract Bundle recommender system is divided into two main stages: bundle editing and bundle recommendation. While substantial research progress has been made in each stage, in practical application scenarios, bundle compositions and the final recommended bundles mutually influence each other: the continuously editing bundle compositions affect the recommendation results, while user feedback on recommended bundles in turn guides the refinement of bundle compositions. This paper presents the first comprehensive reproducibility study of the complete bundle recommendation pipeline. We implement eight bundle-level editing methods, nine item-level editing methods, and seven state-of-the-art bundle recommendation models, and evaluate their performance across six real-world datasets. Our empirical analysis reveals several key findings. First, bundle-level editing faces the challenge of generating high-quality bundles. Second, in the item-level editing, the replacement operation emerges as a universal bottleneck across all methods. Third, in the recommendation stage, recommendation models exhibit varying performance across different interaction density scenarios (e.g., cold-start). Finally, bundle recommendation suffers degraded performance when integrating item-level editing and bundle recommendation within a unified pipeline. Overall, there is the systemic limitation of bundle recommendation: prior work has focused on optimizing individual stages independently, disregarding the interdependencies throughout the entire recommendation system. These findings highlight the urgent need to develop end-to-end solutions that can holistically address the bundle editing and recommendation workflow. Our repository is now available for public access via https://github.com/anyr123/Bundle_Edit_Rec_SIGIR26. \endabstract

Domain-Specific and Structured IR Tasks · *Eureka 1* ·

Tuesday 11:00–12:30

Looking for the Bottleneck in Fine-grained Temporal Relation Classification [Full]

Hugo Sousa (University of Porto); Ricardo Campos (University of Beira Interior); Alipio Jorge (University of Porto)

Temporal relation classification is the task of determining the temporal relation between pairs of temporal entities in a text. Despite recent advancements in natural language processing, temporal relation classification remains a considerable challenge. Early attempts framed this task using a comprehensive set of temporal relations between events and temporal expressions. However, due to the task complexity, datasets have been progressively simplified, leading recent approaches to focus on the relations between event pairs and to use only a subset of relations. In this work, we revisit the broader goal of classifying interval relations between temporal entities by considering the full set of relations that can hold between two time intervals. The proposed approach, Interval from Point, involves first classifying the point relations between the endpoints of the temporal entities and then decoding these point relations into an interval relation. Evaluation on the TempEval-3 dataset shows that this approach can yield effective results, achieving a temporal awareness score of 70.1 percent, a new state-of-the-art on this benchmark.

RULER: Robust Unified LLM-based Efficient Retrieval for Legal Information [Full]

Chenyu Hou (College of Computer Science and Technology, Zhejiang University of Technology); Ziyang Wang (College of Computer Science and Technology, Zhejiang University of Technology); Bin Cao (College of Computer Science and Technology, Zhejiang University of Technology); Jiaxing Wang (College of Computer Science and Technology, Zhejiang University of Technology); Tianming Zhang (College of Computer Science and Technology, Zhejiang University of Technology); Tiantian Li (College of Computer Science and Technology, Zhejiang University of Technology)

Legal information retrieval demands high precision, yet traditional "Retrieve-then-Rerank" pipelines with two separate models suffer from cascading error propagation and knowledge disconnects between stages. To address these issues, we propose RULER, a Robust Unified LLM-based Efficient Retrieval that integrates efficient Bi-Encoder retrieval and high-precision Cross-Encoder reranking within a parameter-sharing architecture. To mitigate the Phantom Hits problem that irrelevant documents are assigned unreasonably high confidence, we introduce a Distribution-Robust Data Construction strategy that explicitly simulates pure-negative candidate groups. This is coupled with a Dynamic Margin Ranking Objective and Maximum Entropy Regularization, which collectively enforce uncertainty on irrelevant samples and enhance robustness. Extensive experiments on the JuDGE and LeCaRDv2 benchmarks demonstrate that RULER achieves state-of-the-art performance, outperforming all independent retrievers in retrieval tasks and surpassing competing unified architectures—where retriever and reranker parameters are shared—in high-precision reranking.

FollowTable: A Benchmark for Instruction-Following Table Retrieval [Full]

Rihui Jin (Southeast University); Yuchen Lu (Southeast University); Ting Zhang (Southeast University); Jun Wang (Southeast University); Kuicai Dong (Huawei Noah's Ark Lab); Zhaocheng Du (Huawei Noah's Ark Lab); Dongping Liu (Southeast University); Gang Wang (Huawei Noah's Ark Lab); Yong Liu (Huawei Noah's Ark Lab); Guilin Qi (Southeast University)

Table Retrieval (TR) has traditionally been formulated as an ad-hoc retrieval problem, where relevance is primarily determined by topical semantic similarity. With the growing adoption of LLMs-based agentic systems, access to structured data is increasingly instruction-driven, where relevance is conditional on explicit content and schema constraints rather than topical similarity alone. We therefore formalize Instruction-Following Table Retrieval (IFTR), a new task that requires models to jointly satisfy topical relevance and fine-grained instruction constraints. We identify two core challenges in IFTR: (i) sensitivity to content scope, such as inclusion and exclusion constraints, and (ii) awareness of schema-grounded requirements, including column semantics and representation granularity—capabilities largely absent in existing retrievers. To support systematic evaluation, we introduce FollowTable, the first large-scale benchmark for IFTR, constructed via a taxonomy-driven annotation pipeline. We further propose a new metric, termed the Instruction Responsiveness Score , to evaluate whether retrieval rankings consistently adapt to user instructions relative to a topic-only baseline. % Our results indicate that existing retrieval models struggle to follow fine-grained instructions over tabular data. In particular, they exhibit systematic biases toward surface-level semantic cues and remain limited in handling schema-grounded constraints, highlighting substantial room for future improvements. Our benchmark is released publicly. https://github.com/AriKing11/FollowTable_Benchmark

Corpus-Centric Learning for Zero-Shot Table Retrieval [Full]

Zhou He (Zhejiang University); Zhifei Pang (Zhejiang University); Xiu Tang (Zhejiang University); Sai Wu (Zhejiang University); Gang Chen (Zhejiang University)

Tabular data represents a major source of structured knowledge for open-domain question answering (QA) and enterprise data lakes, yet effective table retrieval remains challenging due to the structure–semantics gap imposed by tabular layouts. Heuristic table linearization often leads to semantic loss, particularly for implicit queries, while recent supervised retrieval models (e.g., Birdie, Contr) rely heavily on large-scale query logs and labeled QA pairs, limiting their applicability in Day-0 cold-start scenarios. We propose GeCo-TR (Generative Schema and Contrastive Table Retrieval), a zero-shot table retrieval framework that eliminates the need for supervised QA data by shifting from direct query-to-table learning to modeling the intrinsic structural semantics of the table corpus. GeCo-TR introduces UHMI, a unified hybrid representation that integrates table structure with linked knowledge graph entities, and employs a hybrid neural–symbolic retrieval mechanism that dynamically combines dense semantic retrieval, symbolic graph traversal, and sparse lexical matching. This design enables robust semantic generalization while enforcing explicit structural constraints, resulting in high-precision and high-recall retrieval for implicit queries in a zero-shot setting. Extensive experiments on public benchmarks demonstrate the effectiveness of GeCo-TR; notably, on Open-WikiTable, it achieves 97.5% Recall@5 in the zero-shot setting, ranking second among all evaluated methods despite requiring no query-level supervision.

Aspect-Aware Content-Based Recommendations for Mathematical Research Papers [Full]

Ankit Satpute (FIZ Karlsruhe); André Greiner-Petter (University of Göttingen); Noah Gießing (FIZ Karlsruhe); Olaf Teschke (FIZ Karlsruhe); Moritz Schubotz (FIZ Karlsruhe); Akiko Aizawa (National Institute of Informatics); Bela Gipp (University of Göttingen)

Content-based research paper recommendation (CbRPR) has seen advances in computer science and biomedicine, but remains unexplored for mathematics, where paper relatedness is more conceptual than explicit textual or citation-based similarity. Mathematics papers may be connected through shared proof techniques, logical implications, or natural generalizations, yet exhibit minimal textual or citation overlap, rendering existing CbRPR ineffective. To address this gap, we first conduct an expert-driven study characterizing mathematical recommendations, revealing that relevance is inherently aspect -driven. Grounded in this insight, we introduce GoldRiM (small, expert-annotated) and SilverRiM (large, automatically derived), the first datasets for aspect -aware CbRPR in mathematics. Recognizing that LLM embeddings of mathematical content alone yield suboptimal representation, we propose AchGNN, an aspect -conditioned heterogeneous GNN that jointly models textual semantics, citation structure, and author lineage. Across GoldRiM and SilverRiM, AchGNN consistently outperforms prior aspect -based CbRPR methods, achieving substantial gains across all evaluated aspects. We conduct ablation studies to analyze the contributions of individual aspect supervision, authorship lineage, and graph-structural signals to AchGNN's performance. To assess domain generality, we further evaluate AchGNN on the Papers with Code dataset of machine learning publications, demonstrating that our aspect -aware approach effectively transfers beyond mathematics. We deploy our system on the MaRDI platform to help mathematicians with recommendations and release datasets and code publicly for reproducibility: github.com/gipplab/MathAspectRecSys.

MUDY: Multi-Granular Dynamic Candidate Contextualization for Unsupervised Keyphrase Extraction [Full]

Hyeongu Kang (Korea University); Susik Yoon (Korea University)

Keyphrase extraction aims to automatically identify concise phrases that effectively represent the content of a document. While recent methods leveraging pre-trained language models (PLMs) have significantly improved the extraction of keyphrases with strong global semantic relevance, they often fall short in capturing the local contextual importance of keyphrases tied to specific subtopics dispersed in a document. In this paper, we propose a novel context-centric framework, MUDY, that effectively captures multi-granular contextual salience of candidate keyphrases. MUDY employs two complementary components: (1) a prompt-based scoring that estimates the generation likelihood of each candidate keyphrase, augmented with candidate-aware weighting to better reflect its local contextual importance, and (2) a self-attention-based scoring that utilizes multi-granular attention patterns from PLMs to assess candidate significance at both the document-wide and segment-specific levels. Evaluations on four real-world datasets demonstrate that MUDY outperforms state-of-the-art baselines in top-k accuracy at various cutoff thresholds. In-depth quantitative and qualitative analyses further highlight the efficacy of context-centric keyphrase extraction with multi-granular saliency. For reproducibility, the source code of MUDY is available at https://github.com/HgKang1/MUDY.

Fact Checking and Entity Resolution · *Eureka 2* ·

Tuesday 11:00–12:30

R3Check: Reinforcement Learning for Iterative Retrieval and Structured Reasoning in Complex Fact Checking [Full]

Peng Qi (National University of Singapore); Yuyang Zhao (National University of Singapore); Wynne Hsu (National University of Singapore); Mong Li Lee (National University of Singapore)

Automated fact-checking aims to verify the veracity of claims based on related evidence, and has become increasingly important as large language models (LLMs) make it easier to generate and disseminate misinformation at scale. In open settings, effective fact-checking requires models to iteratively retrieve relevant evidence and reason over noisy and incomplete information. While recent LLM-based approaches have shown promising reasoning capabilities, prompt-based methods remain limited by the inherent behaviors of base LLMs, and supervised fine-tuning methods typically require costly annotated reasoning trajectories. In this paper, we propose R3Check, a

rule-guided reinforcement learning framework that enables LLMs to perform iterative retrieval–reasoning for multi-hop fact-checking. R3Check formulates the retriever as an external environment and optimizes the model using Group Relative Policy Optimization, relying only on final veracity labels and format-based rewards rather than explicit reasoning annotations. To mitigate the mutual interference between retrieval and reasoning that arises under joint training, we introduce a two-stage curriculum that first trains structured reasoning under closed fact-checking with gold evidence, and then jointly optimizes retrieval and reasoning with real-time retrieval. An importance-based sampling strategy further strengthens effective supervision signals during training. Despite using only a 7B backbone, R3Check outperforms existing baselines and even powerful reasoning LLMs, under both given-evidence and real-time retrieval settings, while producing interpretable reasoning chains. This work demonstrates the potential of pure reinforcement learning to induce effective retrieval–reasoning behaviors for fact-checking under weak supervision.

Multi-Sourced, Multi-Agent Evidence Retrieval for Fact-Checking [Full]

Shuzhi Gong (University of Melbourne); Richard Sinnott (University of Melbourne); Jianzhong Qi (University of Melbourne); Cecile Paris (CSIRO); Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence); Zhuohan Xie (Mohamed bin Zayed University of Artificial Intelligence)

Misinformation spreading over the Internet poses a significant threat to both societies and individuals, necessitating robust and scalable fact-checking that relies on retrieving accurate and trustworthy evidence. Previous methods rely on semantic and social-contextual patterns learned from training data, which limits their generalization to new data distributions. Recently, Retrieval Augmented Generation (RAG) based methods have been proposed to utilize the reasoning capability of LLMs with retrieved grounding evidence documents. However, these methods largely rely on textual similarity for evidence retrieval and struggle to retrieve evidence that captures multi-hop semantic relations within rich document contents. These limitations lead to overlooking subtle factual correlations between the evidence and the claims to be fact-checked during evidence retrieval, thus causing inaccurate veracity predictions. To address these issues, we propose a Web-enhanced Knowledge Graph retrieval Fact-Checking agentic framework (WKGFC), which exploits authorized open knowledge graph as a core resource of evidence. LLM-enabled retrieval is designed to assess the claims and retrieve the most relevant knowledge subgraphs, forming structured evidence for fact verification. To augment the knowledge graph evidence, we retrieve web contents for completion. The above process is implemented as an automatic Markov Decision Process (MDP): A reasoning LLM agent decides what actions to take according to the current evidence and the claims. To adapt the MDP for fact-checking, we use prompt optimization to fine-tune the agentic LLM. Our extensive experiments over datasets in three categories (Wikipedia, websites, and article summaries) show that WKGFC outperform several advanced state-of-the-art fact-checking methods in balanced accuracy score by over 5% absolute. These results highlight the effectiveness of knowledge-centric evidence retrieval for fact-checking under open-world settings.

CoMCo: Consistency-Aware Multi-Agent Coordination for Zero-Shot Cross-Modal Entity Matching [Full]

Shiqi Zhang (National University of Defense Technology); Weixin Zeng (National University of Defense Technology); Ziheng Zhang (National University of Defense Technology); Wenzhe Hou (National University of Defense Technology); Weidong Xiao (National University of Defense Technology); Xiang Zhao (National University of Defense Technology)

Entity Matching (EM) is a fundamental task in data integration, traditionally studied over structured data such as tables and knowledge graphs. In modern repositories, real-world objects are often represented across multiple modalities, including \emph{structured entities} with symbolic attributes and \emph{visual entities} in image-centric collections. This motivates cross-modal entity matching, which aims to identify visual-structured entity pairs that refer to the same object. Existing methods typically rely on pretrained vision-language models to compute entity pair similarity and derive correspondence via local ranking, which, however, can be unreliable given noisy and ambiguous cross-modal data and may produce globally inconsistent correspondences across related entities. Multimodal large language models (MLLMs) offer richer cross-modal cues for matching, but exhaustive MLLM reasoning over large candidate spaces is prohibitively expensive. To address these limitations, in this work, we propose \ourq, a blackboard-based multi-agent framework for zero-shot cross-modal entity matching that \emph{performs iterative, self-correcting refinement by fusing multiple matching signals, explicitly regulating global consistency, and selectively invoking an MLLM only for hard cases}. We further construct two new benchmarks from real-world visual and structured data. Extensive experiments show that \ourq consistently outperforms competitive baselines, providing an effective solution for zero-shot cross-modal entity matching. We release our data and code at \url{https://github.com/Q-17/CoMCo}.

Unified Entity Matching under Scarce Supervision via Meta-Rule Induction and Retrieval [Full]

Ziheng Zhang (National University of Defense Technology); Weixin Zeng (National University of Defense Technology); Jiuyang Tang (National University of Defense Technology); Xiang Zhao (National University of Defense Technology)

Entity matching is a fundamental task in a wide range of retrieval and knowledge applications, aiming to identify whether two objects correspond to the same real-world entity across heterogeneous sources. Typical variants include entity resolution (ER), entity linking (EL), and entity alignment (EA). While recent unified matchers have made progress through multi-task training with comprehensive annotations, real-world pipelines often operate under scarce supervision, where labeled data is incomplete and fails to cover the full spectrum of matching scenarios. In this regime, supervised unified models degrade substantially, and deployable compact LLMs remain unreliable: lightweight fine-tuning and in-context learning yield inconsistent behavior and can even exhibit negative effects under scenario shifts. To fill in this gap, we propose øurs, a meta-rule induction and retrieval framework for unified entity matching under scarce supervision. Instead of relying on parametric adaptation, øurs converts limited supervision into explicit natural-language rules, abstracts them into reusable meta-rules via hierarchical clustering, and retrieves the most relevant meta-rules to guide the LLM's inference for each input instance. This design improves robustness by grounding decisions on explicit and reusable evidence, instead of relying solely on implicit adaptation or prompt demonstrations. Extensive experiments show that øurs achieves state-of-the-art performance on unified entity matching under scarce supervision.

SciCheck: Reasoning Distillation for Biomedical Claim Verification [Full]

Gabriel Pereira (Universidade Federal de Pernambuco); Luciano Barbosa (CESAR School)

False claims about medical information can have damaging consequences. Although LLMs have been used to verify such claims, their success depends on access to accurate content and robust reasoning capabilities. Prior research assumes access to gold evidence during inference or relies on expensive, compute-intensive scaling strategies. However, these approaches fall short when applied to real-world verification tasks, where relevant evidence is hard to obtain and efficiency is crucial. To address this, we introduce SciCheck, a novel solution that integrates web evidence retrieval with a process of reasoning distillation. Our approach fine-tunes a small language model using reasoning traces generated by an LLM. The distillation process involves a data preparation pipeline that avoids data leakage and filters reasoning traces to retain only those leading to correct answers. It also combines web and gold evidence during training to improve robustness, while evaluation is performed with web retrieval only. We performed an extensive experimental evaluation on different claim verification datasets. The results demonstrate that SciCheck outperforms competing approaches and proprietary LLMs such as Gemini 2.5 Flash in most scenarios with lower computational cost.

OntoEL: Neuro-Symbolic Biomedical Entity Linking with Differentiable Fuzzy EL⊥ Reasoning [Full]

Chang Lu (Yale University); Yizheng Zhao (Nanjing University)

Current neural biomedical entity linking (BioEL) models treat ontologies as flat dictionaries, ignoring the rich terminological knowledge (TBox) that defines concept boundaries. Consequently, they struggle with contextual ambiguity, often retrieving logically inconsistent candidates based solely on surface similarity. We present OntoEL, a neuro-symbolic framework that shifts BioEL from surface-level matching to logic-grounded reasoning. OntoEL integrates differentiable fuzzy EL⊥ reasoning into the retrieval pipeline as a consistency-aware re-ranker, employing a hybrid strategy: structural TBox reasoning is delegated to classical polynomial-time reasoners, while the sigmoidal Reichenbach implication performs soft type-consistency evaluation, effectively resolving the "implication bias" gradient pathology in previous neuro-symbolic methods. By enforcing ontological axioms as differentiable soft constraints, OntoEL aligns neural representations with logical truth. Comprehensive experiments on three benchmarks (MedMentions, BC5CDR, and NCBI Disease) demonstrate state-of-the-art performance, surpassing strong baselines by up to 4.2% in Accuracy@1. On highly ambiguous mentions requiring ontological reasoning, our method corrects 71.2% of retrieval errors, proving the efficacy of incorporating logical semantics into neural retrieval.

Reranking · *Eureka 3 · Tuesday 11:00–12:30*

When Vision Meets Texts in Listwise Reranking [Full]

Hongyi Cai (Universiti Malaya)

Recent advancements in information retrieval have highlighted the potential of integrating visual and textual information, yet effective reranking for image-text documents remains challenging due to the modality gap and scarcity of aligned datasets. Meanwhile, existing approaches often rely on large models (7B--32B parameters) with reasoning-based distillation, incurring unnecessary computational overhead while primarily focusing on textual modalities. In this paper, we propose Rank-Nexus, a multimodal image-text document reranker that performs listwise qualitative reranking on retrieved lists incorporating both images and texts. To bridge the modality gap, we introduce a progressive cross-modal training strategy. We first train each modality separately: leveraging abundant text reranking data, we distill ranking knowledge into the text branch using GPT-4o as the teacher model. For images, where labeled data is scarce, we construct distilled pairs from multimodal large language model (MLLM) captions on image retrieval benchmarks. Subsequently, we train on a joint image-text reranking dataset. Rank-Nexus achieves outstanding performance on text reranking benchmarks (TREC, BEIR) and the challenging image reranking benchmarks (INQUIRE, MMDocIR), using only a lightweight 2B pretrained visual-language model. This efficient design ensures strong generalization across diverse multimodal scenarios without excessive parameters or reasoning overhead.

Consensus-Anchored Expansion for Noise-Resilient Re-ranking [Full]

Nischal Subedi (University of Delaware); Cencheng Shen (Microsoft)

Dense retrievers bridge vocabulary gaps but suffer from semantic drift, ranking topically similar passages that are not actually relevant to the query. Sparse retrievers like BM25 are generally less prone to such false positives but miss paraphrases. We observe that these failure modes are complementary, with relevant passages concentrated in regions where both retrievers score highly, while single-retriever confidence proves unreliable. Building on this insight, we propose AgreRank, a re-ranking method that leverages sparse-dense consensus for query expansion. AgreRank identifies anchor documents where both retrievers agree, expands the query through these verified anchors, and applies geometric gating to suppress drift from the original intent. On TREC Deep Learning 2021 and 2022, AgreRank improves over Dense+CE by 3.4--5.7% and over RRF+CE by 3.7--5.6% under identical retrieval and reranking components, achieving nDCG@10 of 0.706 and 0.627 with a lightweight 33M-parameter cross-encoder. These gains require no additional parameters or training, demonstrating that more effective use of retrieval signals offers a practical alternative to model scaling.

Where Relevance Emerges: A Layer-Wise Study of Internal Attention for Zero-Shot Re-Ranking [Reproducibility]

Haodong Chen (The University of Queensland); Guido Zuccon (The University of Queensland); Shengyao Zhuang (The University of Queensland); Zheng Yao (The University of Queensland); Teerapong Leelanupab (The University of Queensland)

Zero-shot document re-ranking with Large Language Models (LLMs) has evolved from Pointwise methods to Listwise and Setwise approaches that optimize computational efficiency. Despite their success, these methods predominantly rely on generative scoring or output logits, which face bottlenecks in inference latency and result consistency. In-Context Re-ranking (ICR) has recently been proposed as an O(1) alternative method. ICR extracts internal attention signals directly, avoiding the overhead of text generation. However, existing ICR methods simply aggregate signals across all layers; layer-wise contributions and their consistency across architectures have been left unexplored. Furthermore, no unified study has compared internal attention with traditional generative and likelihood-based mechanisms across diverse ranking frameworks under consistent conditions. In this paper, we conduct an orthogonal evaluation of generation, likelihood, and internal attention mechanisms across multiple ranking frameworks. We further identify a universal "bell-curve" distribution of relevance signals across transformer layers, which motivates the proposed Selective-ICR strategy that reduces inference latency by 30%--50% without compromising effectiveness. Finally, evaluation on the reasoning-intensive BRIGHT benchmark shows that precisely capturing high-quality in-context attention signals fundamentally reduces the need for model scaling and reinforcement learning: a zero-shot 8B model matches the performance of 14B reinforcement-learned re-rankers, while even a 0.6B model outperforms state-of-the-art generation-based approaches. These findings redefine the efficiency-effectiveness frontier for LLM-based re-ranking and highlight the latent potential of internal signals for complex reasoning ranking tasks. Our code and results are publicly available at https://github.com/ielab/Selective-ICR.

Think When Needed: Model-Aware Reasoning Routing for LLM-based Ranking [Full]

Huizhong Guo (Zhejiang University); Tianjun Wei (Nanyang Technological University); Dongxia Wang (Zhejiang University); Yingpeng Du (Nanyang Technological University); Ziyan Wang (Nanyang Technological University); Jie Zhang (Nanyang Technological University); Zhu Sun (Singapore University of Technology and Design)

Large language models (LLMs) are increasingly applied to ranking tasks in retrieval and recommendation. Although reasoning prompting can enhance ranking utility, our preliminary exploration reveals that its benefits are inconsistent and come at a substantial computational cost, suggesting that when to reason is as crucial as how to reason. To address this issue, we propose a reasoning routing framework that employs a lightweight, plug-and-play router head to decide whether to use direct inference (Non-Think) or reasoning (Think) for each instance before generation. The router head relies solely on pre-generation signals: i) compact ranking-aware features (e.g., candidate dispersion) and ii) model-aware difficulty signals derived from a diagnostic checklist reflecting the model's estimated need for reasoning. By leveraging these features before generation, the router outputs a controllable token that determines whether to apply the Think mode. Furthermore, the router can adaptively select its operating policy along the validation Pareto frontier at deployment time, enabling dynamic allocation of computational resources toward instances most likely to benefit from Think under varying system constraints. Experiments on three public ranking datasets with different scales of open-source LLMs show consistent improvements in ranking utility with reduced token consumption (e.g., +6.3% NDCG@10 with –49.5% tokens on MovieLens with Qwen3-4B), demonstrating reasoning routing as a practical solution to the accuracy-efficiency trade-off.

Revisiting Text Ranking in Deep Research [Reproducibility]

Chuan Meng (The University of Edinburgh); Litu Ou (The University of Edinburgh); Sean MacAvaney (University of Glasgow); Jeff Dalton (The University of Edinburgh)

Deep research has emerged as an important task that aims to address hard queries that need extensive open-web exploration. To tackle it, most prior work equips large language model (LLM)-based agents with opaque web search APIs, enabling agents to iteratively issue search queries, retrieve external evidence, and reason over it. Despite search's essential role in deep research, black-box web search APIs leave the behaviour of established text ranking methods in deep research largely unclear. To fill this gap, we reproduce key findings and best practices for text ranking methods in deep research. We examine their effectiveness from three perspectives: (i) retrieval units (documents vs. passages), (ii) pipeline configurations (different retrievers, re-rankers, and re-ranking depths), and (iii) query characteristics (the mismatch between agent-issued queries and the training queries of text rankers). We perform experiments on BrowseComp-Plus, a deep research dataset with a fixed corpus, evaluating 2 open-source agents, 5 retrievers, and 3 re-rankers. We find that agent-issued queries typically follow web-search-style syntax (e.g., quoted exact matches), favouring lexical, learned sparse, and multi-vector retrievers; passage-level units are more efficient under limited context windows, and avoid the difficulties of document length normalisation in lexical retrieval; re-ranking is highly effective. We further propose a query-to-question (Q2Q) method that translates agent-issued queries into natural-language questions, significantly reducing the query mismatch.

Reproducing Adaptive Reranking for Reasoning-Intensive IR [Reproducibility]

Mandeep Rathee (L3S Research Center); Venktesh V (Stockholm University); Sean MacAvaney (University of Glasgow); Avishek Anand (Delft University of Technology (TU Delft))

The classical cascading pipeline of retrieve--rerank suffers from a bounded recall problem, stemming from limitations of the first-stage retriever. Most current approaches address the bounded recall problem by improving the first-stage retriever, but this incurs substantial training and inference costs, especially to handle queries that require substantial reasoning. To circumvent the computational costs of reasoning-based retrievers, we replicate the findings of GAR, Graph-based Adaptive Reranking, on the BRIGHT reasoning-intensive retrieval benchmark. GAR addresses the bounded recall problem by modifying the reranking process itself through iterative exploration of a corpus graph, but it was previously only tested on models designed for topical and question-answering-style queries. Hence, reproduce GAR in reasoning-intensive settings with reasoning and non-reasoning reranking models. We observe that the quality of the reranker's signal plays an important role in identifying additional relevant documents within the corpus graph. Overall, we find that GAR boosts the effectiveness of reasoning-intensive retrieval across a variety of models while contributing minimally to computational overheads. Ultimately, this work enables more practical deployment of retrieval systems that can address reasoning-intensive queries.

RAG Systems · *Courtyard 1+2 · Tuesday 11:00–12:30*

SRAG: A Lightweight and Specialized Retrieval-augmented Generation System at the Edge [Full]

Ruikun Luo (Huazhong University of Science and Technology); Zihan Xing (Huazhong University of Science and Technology); Lin Gu (Huazhong University of Science and Technology); Song Wu (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology); Xiaoyu Xia (RMIT University)

Retrieval-augmented generation (RAG) has shown strong potential for deploying large language models at the edge, yet existing designs largely rely on generic and monolithic knowledge bases that are poorly matched to the heterogeneous queries and resource-constrained edge computing environments. Through extensive empirical analysis, we find that domain-specialized knowledge bases, when deployed on individual edge servers, deliver substantially higher retrieval accuracy and generation quality than generic knowledge bases under identical resource budgets. Based on this, we propose SRAG, a distributed RAG system that enforces knowledge specialization at the edge. Each edge server maintains a domain-aware specialized knowledge base by retaining domain-aligned knowledge and decoupling out-of-domain content. SRAG uses a buffer-based knowledge migration mechanism to redistribute out-of-domain content to better-matched edge servers, enabling efficient global knowledge utilization without central coordination. To handle domain-mismatched queries, SRAG employs lightweight cross-node routing guided by compact metadata summaries, avoiding full knowledge replication. Together, these mechanisms form an end-to-end workflow for decentralized edge RAG. Experiments show that SRAG improves retrieval relevance, generation quality, and storage efficiency, while reducing end-to-end latency.

Incentivizing Retrieval-Augmented Generation via Inner Adaptive Context Selection [Full]

Chenxu Cui (Institute of Information Engineering, Chinese Academy of Sciences); Lin Shen (Institute of Information Engineering, Chinese Academy of Sciences); Haihui Fan (Institute of Information Engineering, Chinese Academy of Sciences); Sa Zhu (Institute of Information Engineering, Chinese Academy of Sciences); Feifei Dai (Institute of Information Engineering, Chinese Academy of Sciences); Bo Li (Institute of Information Engineering, Chinese Academy of Sciences)

Retrieval-Augmented Generation (RAG) techniques have emerged as a promising direction to merge the non-parametric knowledge into Large Language Models (LLMs), thereby alleviating factual errors, hallucinations and outdated knowledge. Existing RAG methods, which append multiple retrieved documents or passages to the input of LLMs, will inevitably increase the context length, resulting in not only significant computational

overhead and inference latency, but also performance degradation. Although reranking or compression modules have been introduced to address these challenges, they overlook the contextual preferences of the generative LLMs itself and may inadvertently discard information that is crucial for generation accuracy. To this end, we introduce InnerRAG, which incentivizes RAG via Inner Adaptive Context Selection. InnerRAG is a novel paradigm that empowers LLMs to autonomously select relevant context during generation. Our proposed InnerRAG endows the model to accurately identify the documents that are most helpful for generation from long contexts. By endowing the model with this capability, InnerRAG facilitates more effective exploitation of external knowledge without being misled by disturbed information, leading to substantial improvements in generation quality while maintaining computational efficiency. Extensive experiments across multiple benchmarks and human evaluations demonstrate that our method consistently outperforms state-of-the-art RAG baselines. Moreover, our framework is orthogonal and complementary to in-context RAG approaches, offering further performance improvements when combined.

The Powerless Noise: How Experimental Settings Shape the Reported Power of Noise [Reproducibility]

Micha■ Mazuryk (University of Amsterdam); Fleur Dolmans (University of Amsterdam); Louis Gehringer (University of Amsterdam); Ina Klaric (University of Amsterdam); Jia-Huei Ju (University of Amsterdam); Mohammad Aliannejadi (University of Amsterdam)

Recent work has suggested that adding irrelevant documents to the input of retrieval-augmented generation (RAG) systems can improve question-answering performance, a phenomenon referred to as the " Power of Noise. " This motivated investigations into the role of noise in information retrieval. In this paper, we reproduce the main findings of Cuconasu et al. \citecuconasu2024power and evaluate the robustness of the effect under extended experimental settings. We first confirm that the phenomenon holds under the original setup, which uses earlier-generation LLMs, restrictive prompting and constrained decoding settings. We subsequently introduce a series of extensions to investigate the underlying causes of the noise effect, examining the authors' original design choices including the use of different models, instruction prompting, and relaxed output length constraints. Across these ablations, the Power-of-Noise pattern proves highly sensitive to inference configuration: it can appear, weaken, or disappear under small changes to prompt formulation and decoding limits. Combined with our error analysis, which shows substantial contributions from truncation and malformed generations, this variance indicates that the original effect cannot be robustly confirmed as a general benefit of noisy retrieval under these experimental conditions. More broadly, our work highlights the importance of carefully scrutinizing inference design in retrieval-augmented generation systems. Our code is available at https://github.com/ina0105/The-Power-of-Noise-Reproduction.

NuggetIndex: Governed Atomic Retrieval for Maintainable RAG [Full]

Saber Zerhoudi (University of Passau); Michael Granitzer (Interdisciplinary Transformation University Austria); Jelena Mitrovi■ (University of Passau)

Retrieval-augmented generation (RAG) systems are frequently evaluated via fact-based metrics, yet standard implementations retrieve passages or static propositions. This unit mismatch between evaluation and retrieval objects hinders maintenance when corpora evolve and fails to capture superseded facts or source disagreements. We propose NuggetIndex, a retrieval system that stores atomic information units as managed records, so called nuggets. Each record maintains links to evidence, a temporal validity interval, and a lifecycle state. By filtering invalid or deprecated nuggets prior to ranking, the system prevents the inclusion of outdated information. We evaluate the approach using a nuggetized MS MARCO subset, a temporal Wikipedia QA dataset, and a multi-hop QA task. Against passage and unmanaged proposition retrieval baselines, NuggetIndex improves nugget recall by 42%, increases temporal correctness by 9 percentage points without the recall collapse observed in time-filtered baselines, and reduces conflict rates by 55%. The compact nugget format reduces generator input length by 64% while enabling lightweight index structures suitable for browser-based and resource-constrained deployment. We release our implementation, datasets, and evaluation scripts.

FedMosaic: Federated Retrieval-Augmented Generation via Parametric Adapters [Full]

Zhilin Liang (Beihang University); Yuxiang Wang (Beihang University); Zimu Zhou (City University of Hong Kong); Hainan Zhang (Beihang University); Boyi Liu (Beihang University); Yongxin Tong (Beihang University)

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by grounding generation in external knowledge to improve factuality and reduce hallucinations. Yet most deployments assume a centralized corpus, which is infeasible in privacy-aware domains where knowledge remains siloed. This motivates federated RAG (FedRAG), where a central LLM server collaborates with distributed silos without sharing raw documents. In-context RAG violates this requirement by transmitting verbatim documents, whereas parametric RAG encodes documents into light weight adapters that merge with a frozen LLM at inference, avoiding raw-text exchange. We adopt the parametric approach but face two unique challenges induced by FedRAG: high storage and communication from per-document adapters, and destructive aggregation caused by indiscriminately merging multiple adapters. We present FedMosaic, the first federated RAG framework built on parametric adapters. FedMosaic clusters semantically related documents into multi-document adapters with document-specific masks to reduce overhead while preserving specificity, and performs selective adapter aggregation to combine only relevance-aligned, non-conflicting adapters. Experiments show that FedMosaic achieves an average 10.9% higher accuracy than state-of-the-art methods in four categories, while lowering storage costs by 78.8% to 86.3% and communication costs by 91.4%, and never sharing raw documents.

Deletion Isn't Enough: Auditing RAG for Selective Forgetting [Perspective]

Leila Tavakoli (Independent Researcher); Mark Sanderson (RMIT University)

\begin{abstract}For information access systems, it is not enough for outputs or answers to be relevant or correct; they must also be \emph{permitted}. This paper highlights a research gap: non-disclosure obligations often concern \emph{propositions} that must not be stated, while deployed Retrieval-Augmented Generation (RAG) systems enforce restrictions through record-level handling such as document removal or access-control lists. Such systems can appear compliant while still disclosing revoked facts in generated answers. Most RAG evaluations fail to assess this important aspect of compliance. We explore the nature of this gap and introduce \emph{Forgetting-by-Design (FBD)}, a mechanism-agnostic audit that runs probes across paired system states before and after revocation. \emph{FBD} separates compliance into two observable channels---retrieval/citation \emph{exposure} and answer-level \emph{disclosure or abstention}---and reports the cost of compliance using matched lawful controls and substitution-aware utility signals. We instantiate \emph{FBD} in a reproducible RAG setting and show how the resulting report card reveals failures that single metrics miss: retrieval exposure can be suppressed while answer-level leakage persists, and interventions that reduce disclosure can still degrade lawful utility or collapse citation coverage.\end{abstract}

Multimodal Recommendation · *Goldfields* · *Tuesday*

13:30–15:00

Same Frequency Begets Shared Interests: Popular-Niche Wavelet Graph Learning for Multimodal Recommendation [Full]

Yue He (University of Electronic Science and Technology of China); Hongbo Chen (University of Electronic Science and Technology of China); Jingxi Xie (University of Electronic Science and Technology of China); Fengling Li (University of Technology Sydney); Jingjing Li (University of Electronic Science and Technology of China)

Multimodal recommendation systems have achieved success in capturing user interests, yet user interests are inherently diverse. Existing GCN-based methods adopt a uniform approach to model user-item interaction graphs, failing to differentiate between "popular interests" and "niche interests". This deficiency leads to problems such as the information of popular items overshadowing that of niche items during message passing and multimodal fusion, thereby compromising the accuracy and diversity of recommendations. To address these problems, we propose a novel Popular-Niche Graph Wavelet Learning Framework for Multimodal Recommendation (PNGRec). PNGRec first extracts user behavior signals and item popularity signals from the user-item interaction graph, then decomposes the frequency-domain signals of interests into a Popular User-Item Graph and a Niche Graph via frequency-domain signal decomposition, aiming to capture users' popular and niche interests respectively. Furthermore, we introduce a User Interests Four Quadrants Graph Learning mechanism, which enhances the diversity of users' interests under multimodal perception by finely dividing user behavior representations and item popularity representations. Finally, the model is optimized through a self-supervised multi-task joint optimization loss. Extensive experiments on four real-world industrial datasets demonstrate the effectiveness of our proposed interest-divided modeling approach. The source code is publicly available at https://github.com/orangeheyue/PNGRec.

DIGEST: Dynamic Graph Refinement with Dual Contrastive Semantic Transfer for Multimodal Recommendation [Full]

Xiangyu Sai (South China University of Technology); Meysam Madadi (Universitat de Barcelona); Sergio Escalera (Universitat de Barcelona); Yong Xu (South China University of Technology)

Multimodal recommendation benefits from leveraging rich content signals such as images and texts to alleviate interaction sparsity, yet existing graph-based approaches are still hindered by (i) noisy user—item edges that are treated as static during training and (ii) inconsistent representation spaces across interaction-driven and modality-induced graph views. To address these issues, we propose DIGEST, a multi-graph framework that propagates trainable ID embeddings on a denoised user—item graph and a fused modality-induced item—item graph, and interleaves message passing with dynamic graph refinement that iteratively reweights existing edges to suppress noisy connections. To enable reliable semantic transfer across views, DIGEST further introduces a dual contrastive alignment that (i) aligns the collaborative and semantic item views and (ii) constrains the semantic graph representations to projected multimodal features, together with a lightweight dimension decorrelation regularizer and adaptive gated fusion to reduce redundancy and stabilize multi-view learning. Extensive experiments on three Amazon benchmark datasets demonstrate that DIGEST consistently outperforms state-of-the-art multimodal recommenders, achieving up to 8.43% relative improvement on NDCG@20 and 7.66% on Recall@20 over the strongest baselines.

MLLMRec: A Preference Reasoning Paradigm with Graph Refinement for Multimodal Recommendation [Full]

Yuzhuo Dang (National University of Defense Technology); Xin Zhang (National University of Defense Technology); Zhiqiang Pan (National University of Defense Technology); Yuxiao Duan (National University of Defense Technology); Wanyu Chen (National University of Defense Technology); Fei Cai (National University of Defense Technology); Honghui Chen (National University of Defense Technology)

Multimodal recommendation combines the user historical behaviors with the modal features of items to capture the tangible user preferences, presenting superior performance compared to the conventional ID-based recommender systems. However, existing methods still encounter two key problems in the representation learning of users and items, respectively: (1) the initialization of multimodal user representations is either agnostic to historical behaviors or contaminated by irrelevant modal noise, and (2) the widely used KNN-based item-item graph contains noisy edges with low similarities and lacks audience co-occurrence relationships. To address such issues, we propose MLLMRec, a novel preference reasoning paradigm with graph refinement for multimodal recommendation. Specifically, on the one hand, the item images are first converted into high-quality semantic descriptions using a multimodal large language model (MLLM), thereby bridging the semantic gap between visual and textual modalities. Then, we construct a behavioral description list for each user and feed it into the MLLM to reason about the purified user preference profiles that contain the latent interaction intents. On the other hand, we develop the threshold-controlled denoising and topology-aware enhancement strategies to refine the suboptimal item-item graph, thereby improving the accuracy of item representation learning. Extensive experiments on three publicly available datasets demonstrate that MLLMRec achieves the state-of-the-art performance with an average improvement of 21.48% over the optimal baselines. The source code is provided at https://github.com/Yuzhuo-Dang/MLLMRec.

Multi-modal Relational Item Representation Learning for Inferring Substitutable and Complementary Items [Full]

Junting Wang (University of Illinois Urbana-Champaign); Chenghuan Guo (Amazon); Yang Jiao (Amazon); Yanhui Guo (Amazon); Hari Sundaram (University of Illinois Urbana-Champaign); Yan Gao (Amazon)

We study the problem of inferring substitutable and complementary items, which underpins applications such as alternative and follow-up purchase suggestions. Existing approaches typically learn from behavior-derived item-item associations using GNNs or leverage item content alone. However, these methods often overlook two key challenges: (i) user behaviors (e.g., co-view/co-purchase) only provide noisy weak supervision, and (ii) behavior signals are long-tailed, leaving many items with sparse associations. We propose MMSC, a self-supervised multi-modal relational representation learning framework that combines a multi-modal foundation model adapted to encode item metadata and a self-supervised denoising module that learns relationship-aware representations from noisy user behaviors, unified by a hierarchical aggregation mechanism. We further use LLM-assisted supervision to mitigate noise in behavior-derived supervision during training. Experiments on five real-world datasets show that MMSC consistently outperforms existing baselines by 26.1% for substitutable and 39.2% for complementary item inference, while remaining effective for cold-start items.

Well Begun is Half Done: Training-Free and Model-Agnostic Semantically Guaranteed User Representation Initialization for Multimodal Recommendation [Full]

Jinfeng Xu (The University of Hong Kong); Zheyu Chen (Beijing Institute of Technology); Shuo Yang (The University of Hong Kong); Jinze Li (The University of Hong Kong); Hewei Wang (Carnegie Mellon University); Jianheng Tang (Peking University); Wei Wang (Macao Polytechnic University); Xiping Hu (Beijing Institute of Technology); Edith C. H. Ngai (University of Hong Kong)

Recent advancements in multimodal recommendations, which leverage diverse modality information to mitigate data sparsity and improve recommendation accuracy, have gained significant attention. However, existing multimodal recommendations overlook the critical role of user representation initialization. Unlike items, which are naturally associated with rich modality information, users lack such inherent information. Consequently, item representations initialized based on meaningful modality information and user representations initialized randomly exhibit a significant semantic gap. To this end, we propose a Semantically Guaranteed User Representation Initialization (SG-URInit). SG-URInit constructs the initial representation for each user by integrating both the modality features of the items they have interacted with and the global features of their corresponding clusters. SG-URInit enables the initialization of semantically enriched user representations that effectively capture both local (item-level) and global (cluster-level) semantics. Our SG-URInit is training-free and model-agnostic, meaning it can be seamlessly integrated into existing multimodal recommendation models without incurring any additional computational overhead during training. Extensive experiments on multiple real-world datasets demonstrate that incorporating SG-URInit into advanced multimodal recommendation models significantly enhances recommendation performance. Furthermore, the results show that SG-URInit can further alleviate the item cold-start problem and also accelerate model convergence, making it an efficient and practical solution for multimodal recommendations.

TimeMM: Time-as-Operator Spectral Filtering for Dynamic Multimodal Recommendation [Full]

Wei Yang (Xiaohongshu Inc.); Rui Zhong (Xiaohongshu Inc.); Zihan Lin (Xiaohongshu Inc.); Xiaodan Wang (Xiaohongshu Inc.); Cheng Chen (Xiaohongshu Inc.); Huan Ren (Xiaohongshu Inc.); Yao Hu (Xiaohongshu Inc.)

Multimodal recommendation improves user modeling by integrating collaborative signals with heterogeneous item content. In real applications, user interests evolve over time and exhibit nonstationary dynamics, where different preference factors change at different rates. This challenge is amplified in multimodal settings because visual and textual cues can dominate decisions under different temporal regimes. Despite strong progress, most multimodal recommenders still rely on static interaction graphs or coarse temporal heuristics, which limits their ability to model continuous preference evolution with fine-grained temporal adaptation. To address these limitations, we propose TimeMM, a time-conditioned spectral filtering framework for dynamic multimodal recommendation. TimeMM instantiates Time-as-Operator by mapping interaction recency to a family of parametric temporal kernels that reweight edges on the user--item graph, producing component-specific representations without explicit eigendecomposition. To capture non-stationary interests, we introduce Adaptive Spectral Filtering that mixes the operator bank according to temporal context, yielding prediction-specific effective spectral responses. To account for modality-specific temporal sensitivity, we further propose Spectral-Aware Modality Routing that calibrates visual and textual contributions conditioned on the same temporal context. Finally, a ranking-space Spectral Diversity Regularization encourages complementary expert behaviors and prevents filter-bank collapse. Extensive experiments on real-world benchmarks demonstrate that TimeMM consistently outperforms state-of-the-art multimodal recommenders while maintaining linear-time scalability.

Reasoning for Recommendation · *Room 110 · Tuesday 13:30–15:00*

Verifiable Reasoning for LLM-based Generative Recommendation [Full]

Xinyu Lin (National University of Singapore); Hanqing Zeng (Meta AI); Hanchao Yu (Meta AI); Yinglong Xia (Meta AI); Jiang Zhang (Meta AI); Aashu Singh (Meta AI); Fei Liu (Meta AI); Wenjie Wang (National University of Singapore); Fuli Feng (National University of Singapore); Tat-Seng Chua (National University of Singapore); Qifan Wang (Meta AI)

Reasoning in Large Language Models (LLMs) has recently shown strong potential in enhancing generative recommendation through deep understanding of complex user preference. Existing approaches follow a reason-then-recommend paradigm, where LLMs perform step-by-step reasoning before item generation. However, this paradigm inevitably suffers from reasoning degradation (i.e., homogeneous or error-accumulated reasoning) due to the lack of intermediate verification, thus undermining the recommendation. To bridge this gap, we propose a novel reason-verify-recommend paradigm, which interleaves reasoning with verification to provide reliable feedback, guiding the reasoning process toward more faithful user preference understanding. To enable effective verification, we establish two key principles for verifier design: 1) reliability ensures accurate evaluation of reasoning correctness and informative guidance generation; and2) multi-dimensionality emphasizes comprehensive verification across multi-dimensional user preferences. Accordingly, we propose an effective implementation called VRec. It employs a mixture of verifiers to ensure multi-dimensionality, while leveraging a proxy prediction objective to pursue reliability. Experiments on four real-world datasets demonstrate that VRec substantially enhances recommendation effectiveness and scalability without compromising efficiency.

Factorized Latent Reasoning for LLM-based Recommendation [Full]

Tianqi Gao (Independent Researcher); Chengkai Huang (University of New South Wales); Zihan Wang (Meituan LongCat Interaction Team); Cao Liu (Meituan LongCat Interaction Team); Ke Zeng (Meituan LongCat Interaction Team); Lina Yao (University of New South Wales)

Large language models (LLMs) have recently been adopted for recommendation by framing user preference modeling as a language generation problem. However, existing latent reasoning approaches typically represent user intent with a single latent vector, which struggles to capture the inherently multi-faceted nature of user preferences. We propose Factorized Latent Reasoning (FLR), a novel framework for LLM-based sequential recommendation that decomposes latent reasoning into multiple disentangled preference factors. FLR introduces a lightweight multi-factor attention module that iteratively refines a latent thought representation, where each factor attends to distinct aspects of the user's interaction history. To encourage diversity and specialization, we design orthogonality, attention diversity, and sparsity regularization objectives, and dynamically aggregate factor contributions for the final prediction. We further integrate FLR with an efficient reinforcement learning strategy based on group-relative policy optimization, enabling stable alignment directly in the latent reasoning space. Experiments on multiple benchmarks show that FLR consistently outperforms strong baselines while improving robustness and interpretability. Our data and code are available at https://github.com/ToAdventure/FLR.

Beyond the Single Path: Divergent Reasoning for LLM-based Recommendation [Full]

Guojia An (University of Electronic Science and Technology of China); Jie Zou (University of Electronic Science and Technology of China); Yuhan Yang (University of Electronic Science and Technology of China); Shuai Qin (University of Electronic Science and Technology of China); Weikang Guo (Southwest University of Finance and Economics); Jinyu Guo (University of Electronic Science and Technology of China); Yang Yang (University of

Electronic Science and Technology of China)
Large Language Models (LLMs) have demonstrated strong potential in recommendations due to their powerful reasoning capabilities. However, existing methods typically rely on a single reasoning path to drive the entire Top-K recommendations. This paradigm is prone to reasoning path collapse, where limiting exploration of potentially superior and diverse reasoning paths within the LLMs space. As a result, both the accuracy and diversity of the recommendation outcomes are constrained. To address this issue, we propose a novel model, Divergent Reasoning for LLM-based Recommendation, named DivReason. Inspired by the structure of intellect theory, which emphasizes a two-stage cognitive process of divergent thinking followed by convergent thinking, DivReason is designed with two core components: the Divergent Reasoning Path Generation Module and the Reasoning Path Aggregation Module. In the first module, DivReason introduces a training-free form of controlled uncertainty to promote diverse reasoning, leveraging Monte Carlo Dropout and Directional Perturbation to expand exploration in the latent reasoning space. In the Reasoning Path Aggregation Module, we adaptively select a subset of high-quality reasoning paths from the entire path pool and aggregate them into a unified reasoning representation. Meanwhile, we further adopt an alternating reinforcement learning strategy to optimize the model, explicitly balancing accuracy and diversity during training. Extensive experimental results show that DivReason effectively mitigates the issue of reasoning path collapse, while improving both the accuracy and diversity of LLM-based recommendations.

LWGR: Lagrangian-Constrained Personalized World Knowledge for Generative Recommendation [Full]

Lingyu Mu (Institute of Information Engineering, Chinese Academy of Sciences); Hao Deng (Alibaba International Digital Commerce Group); Haibo Xing (Alibaba International Digital Commerce Group); Kaican Lin (Alibaba International Digital Commerce Group); Zhitong Zhu (Institute of Information Engineering, Chinese Academy of Sciences); Zhengxiao Liu (Institute of Information Engineering, Chinese Academy of Sciences); Zheng Lin (Institute of Information Engineering, Chinese Academy of Sciences); Xiaoyi Zeng (Alibaba International Digital Commerce Group); Yu Zhang (Alibaba International Digital Commerce Group); Jinxin Hu (Alibaba International Digital Commerce Group)
Recent progress in large language model (LLM) based generative recommendation (GR) shows that leveraging LLM world knowledge can substantially improve performance. However, existing methods rely on fixed, manually designed instructions to generate semantic knowledge and directly incorporate it into GR, which has two limitations: (1) fixed instructions cannot capture the multidimensional heterogeneity of user interests; (2) uncontrollable knowledge fusion may conflict with behavioral signals and harm recommendations.To address these limitations, we propose \textbf{LWGR}, a framework that leverages \textbf{L}agrangian constraints to transfer users' personalized \textbf{W}orld knowledge from LLMs into \textbf{G}enerative \textbf{R}ecommendation. LWGR enhances GR along two axes: knowledge extraction and fusion. It builds user personalized soft instructions to extract behavior-relevant LLM world knowledge. Then, it formulates knowledge fusion as an optimization problem with explicitly bounded performance degradation, solved via a Lagrangian primal–dual method that selectively incorporates beneficial knowledge.We further design two training strategies for different LLM scales and a deployment scheme that combines nearline precomputation with lightweight online serving. Experiments on multiple public datasets and one industrial dataset show that LWGR outperforms eight state-of-the-art baselines by up to 11.23\% and brings a 1.35\% revenue lift on a large-scale advertising platform, demonstrating its effectiveness and practicality.

Beyond Static Best-of-N: Bayesian List-wise Alignment for LLM-based Recommendation [Full]

Ruijun Chen (University of Science and Technology of China); Chongming Gao (University of Science and Technology of China); Jiawei Chen (Zhejiang University); Weiqin Yang (Zhejiang University); Xiangnan He (University of Science and Technology of China)
Large Language Models have revolutionized recommender systems (LLM4Rec) by leveraging their generative capabilities to model complex user preferences. However, existing LLM4Rec methods primarily rely on token-level objectives, making it difficult to optimize list-level and non-differentiable metrics (e.g., NDCG, fairness) that define actual recommendation quality. While Best-of-N (BoN) directly optimizes these metrics during inference, its high computational cost hinders real-world deployment. To address this, BoN Alignment aims to distill the search capability into the model itself, yet current approaches suffer from two critical limitations: (1) Indiscriminate Supervision, where the static reference fails to distinguish the relative quality of candidates exceeding its empirical range, leading to a loss of ranking guidance; and (2) Gradient Decay, where the effective supervision signal rapidly diminishes as the evolving policy improves, resulting in inefficient optimization. To overcome these challenges, we propose BLADE (Bayesian List-wise Alignment via Dynamic Estimation). Unlike static approaches, BLADE introduces a Bayesian framework that continuously updates the target distribution by fusing historical priors with dynamic evidence from the model's current rollouts. This mechanism constructs a self-evolving target that adapts to the model's growing capabilities, ensuring the training signal remains informative throughout the learning process. Extensive experiments on three real-world datasets demonstrate that BLADE significantly outperforms state-of-the-art baselines. Crucially, it breaks the static performance upper bound, achieving sustained gains in both ranking accuracy (Recall, NDCG) and complex list-wise metrics (Fairness, Diversity). The code is available via https://github.com/RegionCh/BLADE.

GNN-based Item Indexing for LLM-enhanced Recommendation [Full]

Mao Senlin (Nanjing University of Aeronautics and Astronautics); Ji Zhang (Nanjing University of Aeronautics and Astronautics); Peng Zhang (Guangzhou University); Ze Wang (Tiangong University); Xiaoyao Zheng (Anhui Normal University); Jia Wang (Xi'an Jiaotong-Liverpool University)
Large language models (LLMs) have transformed recommender systems through strong semantic understanding and generalization. However, the design of item identifiers remains a critical bottleneck that directly affects recommendation quality. Traditional metadata-based identifiers introduce length variability and semantic ambiguity, whereas existing collaborative indexing (CID) approaches often neglect item attributes, show limited cross-dataset generalizability, and incur high computational cost at scale. To address these limitations, we propose a Graph Neural Network (GNN)–based item indexing framework with three coordinated innovations. First, we construct attribute-enriched co-occurrence graphs and use a GNN encoder to fuse item features with collaborative signals, yielding semantically informed representations that work well for attribute-rich catalogs. Second, we replace recursive spectral clustering with hierarchical agglomerative clustering on GNN embeddings, enabling direct control of index length via tree depth and reducing hyperparameter tuning across datasets. Third, we exploit localized message passing rather than global eigendecomposition, which provides considerably better runtime efficiency and is amenable to mini-batch training, supporting online index updates as interactions evolve. Across five benchmarks, GID achieves strong average ranking performance, showing larger improvements on sparse and attribute-rich datasets while remaining competitive in dense settings. The framework is robust under both seen and unseen prompt templates, which supports practical LLM-based recommendation. On sequential recommendation, GID improves HR@10 by 7.9% on average over the strongest baseline in each dataset.

Privacy, Security, and Robustness in IR · *Eureka 1* ·

Tuesday 13:30–15:00

Spectral-Adaptive Adversarial Hashing for Robust Image Retrieval [Full]

Gang Zhou (Beijing University of Posts and Telecommunications); Shibiao Xu (Beijing University of Posts and Telecommunications); Xiaolong Zheng (Institute of Automation, Chinese Academy of Sciences); Daniel Dajun Zeng (Institute of Automation, Chinese Academy of Sciences)
Deep hashing is widely used in large-scale image retrieval systems due to its efficient retrieval performance. However, its susceptibility to adversarial attacks limits its security in practical applications. Adversarial training is the most effective method for improving robustness, but it often leads to a significant trade-off between robustness and retrieval accuracy. In this paper, we conduct spectral analysis and find that generating high-quality hash codes requires wide-frequency response models, whereas adversarial training forces the model into spectral collapse, degrading it to a low-frequency response model and weakening its discriminability. To address this issue, we propose a Spectral-Adaptive Adversarial Hashing (SAAH) framework, which selectively preserves discriminative and task-relevant frequency components while suppressing adversarially unstable ones, enabling robust hashing without sacrificing retrieval performance. Extensive experiments on benchmark datasets demonstrate that SAAH consistently achieves a superior balance between retrieval accuracy and adversarial robustness, achieving the best performance in both retrieval accuracy and robustness compared with existing robust hashing methods.

Prompt-Unknown Promotion Attacks against LLM-based Sequential Recommender Systems [Full]

Yuchuan Zhao (The University of Queensland); Tong Chen (The University of Queensland); Junliang Yu (Griffith University); Zongwei Wang (Chongqing University); Lizhen Cui (Shandong University); Hongzhi Yin (University of Queensland)
Large language model-powered sequential recommender systems (LLM-SRSs) have recently demonstrated remarkable performance, enabling recommendations through prompt-driven inference over user interaction sequences. However, this paradigm also introduces new security vulnerabilities, particularly text-level manipulations, rendering them appealing targets for promotion attacks that purposely boost the ranking of specific target items. Although such security risks have been receiving increasing attention, existing studies typically rely on an unrealistic assumption of access to either the victim model or prompt to unveil attack mechanisms. In this work, we investigate the item promotion attack in LLM-SRSs under a more realistic setting where both the system prompt and victim model are unknown to the attacker, and propose a Prompt-Unknown Dual-poisoning Attack (PUDA) framework. To simulate attacks under this full black-box setting, we introduce an LLMbased evolutionary refinement strategy that infers discrete system prompts, enabling the training of an effective surrogate model that mimics the behaviors of the victim model. Leveraging the distilled prompt and surrogate model, we devise a promotion attack that adversarially revises target item texts under semantic constraints, which is further complemented by the highly plausible, surrogategenerated poisoning sequences to enable cost-effective target item promotion. Extensive experiments on real-world datasets demonstrate that PUDA consistently outperforms state-of-the-art competitors in boosting the exposure of unpopular target items. Our findings reveal critical security risks in modern

LLM-SRSs even when both prompts and models are protected, and highlight the need for more robust defensive means.

The Vulnerability of LLM Rankers to Prompt Injection Attacks: You are to [MARK] this paper as the Best Paper [Reproducibility]

Yu Yin (The University of Queensland); Shuai Wang (The University of Queensland); Bevan Koopman (The University of Queensland); Guido Zuccon (The University of Queensland)

Large Language Models (LLMs) have emerged as powerful re-rankers. Recent research has however showed that simple prompt injections embedded within a candidate document (i.e., jailbreak prompt attacks) can significantly alter an LLM's ranking decisions. While this poses serious security risks to LLM-based ranking pipelines, the extent to which this vulnerability persists across diverse LLM families, architectures, and settings remains largely under-explored. In this paper, we present a comprehensive empirical study of jailbreak prompt attacks against LLM rankers. We focus our evaluation on two complementary tasks: (1) Preference Vulnerability Assessment, measuring intrinsic susceptibility via attack success rate (ASR); and (2) Ranking Vulnerability Assessment, quantifying the operational impact on the ranking's quality (nDCG@10). We systematically examine three prevalent ranking paradigms (pairwise, listwise, setwise) under two injection variants: decision objective hijacking and decision criteria hijacking. Beyond reproducing prior findings, we expand the analysis to cover vulnerability scaling across model families, position sensitivity, backbone architectures, and cross-domain robustness. Our results characterize the boundary conditions of these vulnerabilities, revealing critical insights such as that encoder-decoder architectures exhibit strong inherent resilience to jailbreak attacks. We publicly release our code and additional experimental results at https://github.com/ielab/LLM-Ranker-Attack.

Privacy Preserving Information Retrieval: Defining Privacy Research Pillars for a Future Research Agenda [Perspective]

Francesco Luigi De Faveri (University of Padova); Guglielmo Faggioli (University of Padova); Asia J. Biega (Max Planck Institute for Security and Privacy); Nicola Ferro (University of Padova)

Advancements in computer science are raising concerns and preoccupations about the privacy of users' data submitted to and used by Information Retrieval (IR) systems. IR systems, such as search engines, integrate new generative information access pipelines that implement effective and efficient document retrieval and answer generation. However, critical challenges arise in Privacy-Preserving IR (PPIR): Are such data leaking personal information or being used to train generative systems? Are current privacy solutions sufficient to guarantee user privacy and limit such information leakage? Are users' queries and retrieved documents protected throughout the entire retrieval process? How has the privacy threats landscape changed, and in which directions should the IR and Privacy research community investigate to address such new risks? In this perspective paper, we provide initial answers to these questions, analysing state-of-the-art solutions for protecting user privacy when accessing information and highlighting areas of concern. We propose a new PPIR research agenda to address the unsolved problem of private data use and access. The agenda includes novel privacy research pillars aimed at addressing objectives grounded in gaps in the literature, user survey findings, and structured interviews with experts from research, industry, and regulatory bodies. By defining these privacy research pillars, we advise the IR community to pursue research toward a more resilient privacy direction that can address future challenges stemming from rapidly advancing technology eager for user data.

SynDiSC: High-Quality Tabular Data Synthesis with Distributional and Semantic Consistency [Full]

Fan Wu (Central South University); Haoye Pan (Central South University); Hao Wu (Nanjing University); Kai Qian (Central South University); Shucheng Li (Central South University); Feng Lyu (Central South University)

Synthesizing high-quality tabular data is essential for privacy-preserving data analysis. However, this task remains challenging due to two key factors: (1) distribution complexity : imbalanced and skewed data make it challenging to learn the data distribution accurately; and (2) semantic coherence : implicit relationships and logical dependencies among fields must be preserved to ensure valid and meaningful synthetic samples. To address these issues, we propose SynDiSC, a high-quality tabular data synthesis approach that enforces both distributional and semantic consistency. It comprises three core designs: (1) a distribution-aware encoding that effectively handles heterogeneous data types and complex distributions; (2) a multi-dimensional semantic conditioning that leverages multi-dimensional conditional dependencies to enforce semantic validity during generation; and (3) a conditional consistency controller that guides the generator to produce diverse samples satisfying multiple conditional constraints while mitigating mode collapse. Extensive experiments on datasets from various application domains demonstrate that SynDiSC significantly improves data quality, conditional controllability, and downstream task performance compared to state-of-the-art methods. Our code and data samples are open-sourced in the GitHub repository. https://github.com/Knightz9/SynDiSC.

How Far Are We from Automatically Identifying Violations of the Data Minimization Principle in Privacy Policies? [Full]

Ziyan Zhou (Institute of Information Engineering, Chinese Academy of Sciences); Yanru He (Institute of Information Engineering, Chinese Academy of Sciences); Yunchuan Guo (Institute of Information Engineering, Chinese Academy of Sciences); Haoyang Yu (Institute of Information Engineering, Chinese Academy of Sciences); Liang Fang (Institute of Information Engineering, Chinese Academy of Sciences); Fenghua Li (Institute of Information Engineering, Chinese Academy of Sciences)

Data protection laws and regulations require service providers to disclose data practices in privacy policies, specifying what personal information is processed and for what purposes. For compliance, these data practices must adhere to the data minimization principle, limiting the processing of personal information to what is directly relevant and necessary for the service purposes. However, data minimization is context-dependent, making violations difficult to define and quantify in privacy policies. Meanwhile, privacy policies are semantically complex and unstructured, hindering accurate extraction of fine-grained data practices and large-scale automated evaluation. To address these issues, we propose DataMini, a human--LLM collaborative evaluation framework for identifying violations of the data minimization principle in privacy policies. First, DataMini categorizes data minimization violations into two dimensions: inherent violations and contextual violations, establishing fine-grained evaluation criteria. Second, we construct a compliance baseline by mining high-frequency patterns from large-scale privacy policies and integrating expert knowledge to derive compliance mappings for human--LLM collaborative evaluation. Finally, the compliance baseline can automatically verify data practices that satisfy the data minimization principle, enabling the framework to focus exclusively on identifying suspected violations to improve efficiency and accuracy. Extensive evaluations demonstrate that DataMini exhibits superior data practice extraction accuracy of 83.46% and achieves an F1-score of 0.8180 for identifying data minimization violations in privacy policies, reducing manual evaluation effort by approximately 80%.

Video Retrieval · *Eureka 2 · Tuesday 13:30–15:00*

Cross-modal Representation Shift Refinement for Point-supervised Video Moment Retrieval [TOIS]

Kun Wang (School of Software, Shandong University, Jinan, China); Yupeng Hu (School of Software, Shandong University, Jinan, China); Hao Liu (School of Software, Shandong University, Jinan, China); Jiang Shao (School of Software, Shandong University, Jinan, China); Liqiang Nie (School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China)

Video Moment Retrieval (VMR) aims to retrieve temporal moments in videos that align with natural language queries, a task requiring cross-modal reasoning between video and text. Among various supervision paradigms, point-supervised VMR has emerged as a practical solution, leveraging single-frame annotations to significantly reduce annotation costs while maintaining competitive retrieval performance. However, this sparse supervision approach induces cross-modal representation shift. This shift complicates the model's ability to accurately capture action sequences and associate text with visual content. To tackle this, we propose a novel framework called pseuDo fRame-based tempOral and semaNtic rEfinement (DRONE) with two key modules: (1) Pseudo-Frame Temporal Alignment (PTA), which embeds textual queries as pseudo-frames to enhance temporal coherence, and (2) Curriculum-Guided Semantic Refinement (CSR), which uses a progressive contrastive learning strategy to refine semantic representations from easy to hard cases. Extensive experiments show that DRONE achieves effective retrieval performance while keeping annotation costs low.

From Interference to Stability: Adversarial Reliability Correction for Video Moment Retrieval with Relevance Feedback [Full]

Hao Liu (Shandong University); Yupeng Hu (Shandong University); Kun Wang (Shandong University); Junchao Wang (Shandong University); Ruping Cao (Shandong University); Yutao Yao (Shandong University); Zilu Cai (Shandong University)

Video Moment Retrieval (VMR) aims to retrieve target video moments that correspond to natural language queries. Most existing methods rely on a positive-only assumption that the queried moment always exists within the video, which limits their reliability in practical scenarios. Departing from this restrictive setting, we study Video Moment Retrieval with Relevance Feedback (VMR-RF), which requires models to both retrieve relevant moments and reject irrelevant queries. This task remains challenging due to the following issues: 1) Intrinsic Semantic Interference caused by visually similar but irrelevant moments, and 2) Propagative Decision Irreversibility induced by unidirectional relevance prediction. In light of these, we introduce AdversaRial Reliability cOrrection netWork (ARROW) for VMR-RF. ARROW adopts an active discriminative strategy through two synergetic components: a Gradient-induced Semantic Adversary (GSA) that probes model vulnerabilities by actively amplifying semantic interference, and an Adversarial Reliability Predictor (ARP) that quantifies prediction stability under such interference to effectively suppress unreliable decisions. Extensive experiments validate the effectiveness of ARROW.

Generation-Augmented Video Corpus Moment Retrieval [Full]

Mingjin Kuai (Nanjing University); Qianyin Xiao (Nanjing University); Juncheng Li (Zhejiang University); Jin Peng (Nanjing University); Lizi Liao (Singapore Management University); Wei Ji (Nanjing University)

Video Corpus Moment Retrieval (VCMR) requires models to efficiently retrieve and precisely locate specific moments relevant to natural language queries within a massive, untrimmed video corpus. However, existing discriminative approaches typically rely on shallow visual-textual feature matching mechanisms, which often struggle to capture fine-grained semantic differences. To address this limitation, we propose Video-GAR, a novel framework that reframes the conventional retrieval task from superficial matching to generative understanding, positing that the capability for query reconstruction evidences deep semantic comprehension. Specifically,

Video-GAR orchestrates three synergistic components: To overcome the computational efficiency bottleneck, we construct a Bi-Mamba backbone that leverages the linear complexity of state-space models for efficient global context modeling. Building on these representations, we introduce a generation-augmented fusion module, in which a training-only decoder acts as a semantic regularizer to implicitly calibrate cross-modal attention without increasing inference overhead. Finally, to ensure fine-grained precision, we propose a boundary-aware localization strategy that integrates boundary modeling with categorical supervision. Experiments on two benchmark datasets demonstrate that Video-GAR significantly improves retrieval and localization accuracy while maintaining outstanding inference speed.

Think with Grounding: Curriculum Reinforced Reasoning with Video Grounding for Long Video Understanding [Full]

Houlun Chen (DCST, Tsinghua University); Xin Wang (DCST, BNRist, Tsinghua University); Guangyao Li (DCST, Tsinghua University); Yuwei Zhou (DCST, Tsinghua University); Yihan Chen (DCST, Tsinghua University); Jia Jia (DCST, BNRist, Tsinghua University); Wenwu Zhu (DCST, BNRist, Tsinghua University)

Long video understanding (LVU) is challenging due to rich and complicated multimodal clues in long temporal range. Current methods adopt reasoning to improve the model's ability to analyze complex video clues in long videos via text-form reasoning. However, the existing literature suffers from the fact that the text-only reasoning under fixed video context may exacerbate hallucinations since detailed crucial clues are often ignored under limited video context length due to the temporal redundancy of long videos. To address this gap, we propose Video-TwG, a curriculum reinforced framework that employs a novel Think-with-Grounding paradigm, enabling video LLMs to actively decide when to perform on-demand grounding during interleaved text–video reasoning, selectively zooming into question-relevant clips only when necessary. Video-TwG can be trained end-to-end in a straightforward manner, without relying on complex auxiliary modules or heavily annotated reasoning traces. In detail, we design a Two-stage Reinforced Curriculum Strategy, where the model first learns think-with-grounding behavior on a small short-video GQA dataset with grounding labels, and then scales to diverse general QA data with videos of diverse domains to encourage generalization. Further, to handle complex think-with-grounding reasoning for various kinds of data, we propose the TwG-GRPO algorithm, which features the fine-grained grounding reward, self-confirmed pseudo reward, and accuracy-gated mechanism. Finally, we propose to construct a new TwG-51K dataset that facilitates training. Experiments on Video-MME, LongVideoBench, and MLVU show that Video-TwG consistently outperforms strong LVU baselines. Further ablation validates the necessity of our Two-stage Reinforced Curriculum Strategy and shows our TwG-GRPO better leverages diverse unlabeled data to improve grounding quality and reduce redundant groundings without sacrificing QA performance. https://github.com/hlchen23/Video-TwG

Agentic Spatio-Temporal Grounding via Collaborative Reasoning [Full]

Heng Zhao (IHPC, Agency for Science, Technology and Research(A*STAR)); Yew-Soon Ong (IHPC, Agency for Science, Technology and Research(A*STAR)); Joey Tianyi Zhou (IHPC, Agency for Science, Technology and Research(A*STAR))

Spatio-Temporal Video Grounding (STVG) aims to retrieve the spatio-temporal tube of a target object or person in a video given a text query. Most existing approaches perform frame-wise spatial localization within a predicted temporal span, resulting in redundant computation, heavy supervision requirements, and limited generalization. Weakly-supervised variants mitigate annotation costs but remain constrained by the dataset-level train-and-fit paradigm with an inferior performance. To address these challenges, we propose the Agentic Spatio-Temporal Grounder (ASTG) framework for the task of STVG in an open-world and zero-shot setting. Specifically, two specialized agents SRA (Spatial Reasoning Agent) and TRA (Temporal Reasoning Agent) constructed leveraging modern Multi-modal Large Language Models (MLLMs) work collaboratively to retrieve the target tube in an autonomous and self-guided manner. Following a propose-and-evaluation paradigm, ASTG duly decouples spatio-temporal reasoning and automates the tube extraction, verification and temporal localization processes. With a dedicated visual memory and dialogue context, ASTG achieves architectural efficiency by minimizing the number of reasoning calls compared to exhaustive per-frame reasoning, eliminating the logical redundancy inherent in joint-reasoning systems. Experiments on popular benchmarks demonstrate the superiority of the proposed approach where it outperforms existing weakly-supervised and zero-shot approaches by a margin and is comparable to some of the fully-supervised methods.

Denoise and Align: Diffusion-Driven Foreground Knowledge Prompting for Open-Vocabulary Temporal Action Detection [Full]

Sa Zhu (Institute of Information Engineering, Chinese Academy of Sciences); Wanqian Zhang (Institute of Information Engineering, Chinese Academy of Sciences); Lin Wang (Hangzhou Dianzi University); Jinchao Zhang (Institute of Information Engineering, Chinese Academy of Sciences); Cong Wang (Zhejiang University); Bo Li (Institute of Information Engineering, Chinese Academy of Sciences)

Open-Vocabulary Temporal Action Detection (OV-TAD) aims to localize and classify action segments of unseen categories in untrimmed videos, where effective alignment between action semantics and video representations is critical for accurate detection. However, existing methods struggle to mitigate the semantic imbalance between concise, abstract action labels and rich, complex video contents, inevitably introducing semantic noise and misleading cross-modal alignment. To address this challenge, we propose DFAlign, the first framework that leverages diffusion-based denoising to generate foreground knowledge for the guidance of action–video alignment. Following the 'conditioning, denoising and aligning' manner, we first introduce the Semantic-Unify Conditioning (SUC) module, which unifies action-shared and action-specific semantics as conditions for diffusion denoising. Then, the Background-Suppress Denoising (BSD) module generates foreground knowledge by progressively removing background redundancy from videos through denoising process. This foreground knowledge serves as effective intermediate semantic anchor between video and text representations, mitigating the semantic gap and enhancing the discriminability of action-relevant segments. Furthermore, we introduce the Foreground-Prompt Alignment (FPA) module to inject extracted foreground knowledge as prompt tokens into text representations, guiding model's attention towards action-relevant segments and enabling precise cross-modal alignment. Extensive experiments demonstrate that our method achieves state-of-the-art performance on two OV-TAD benchmarks. The code repository is provided as follows: https://github.com/Sasa77777779/DFAlign_SIGIR26.git.

Search Behavior · *Eureka 3 · Tuesday 13:30–15:00*

A Comparative Study of Users' Information-Seeking Practices Across Search Engines and Generative AI Chatbots [Full]

Elsa Lichtenegger (University of Zurich); Aleksandra Urman (University of Zurich); Aniko Hannak (University of Zurich)

Web search engines have long been the primary gateway to online information, but generative AI chatbots are emerging as alternative tools. Yet we lack empirical understanding of how users choose between these tools and how their information-seeking behaviors differ across these contexts. Understanding how users select, use, and assess information retrieval (IR) systems is essential for evaluating them in ways that reflect user expectations across contexts. To shed light on user perceptions and usage of IR systems, we surveyed 84 participants about their recent search engine and GenAI chatbot use, analyzing task types, contextual triggers, outcomes, and tool preferences. Our findings reveal that users approach search engines and GenAI chatbots with distinct philosophies. With search engines, they follow a 'Browse and Verify' approach, prioritizing reliability and source verifiability. With GenAI chatbots, they adopt a 'Delegate and Consume' approach, valuing quick summarization and personalized content. Despite these stated preferences, actual usage shows that users frequently turn to GenAI chatbots for actionable guidance in high-stakes domains such as health and relationship matters, hinting at a divergence between stated philosophies and real-world behavior. These findings underscore the need for IR systems that support reliable information assessment, transparent sources, and user-guided evaluation.

Understanding and Modeling Heterogeneous Search Behavior [Full]

Nuha Abu Onq (RMIT University); Chenglong Ma (RMIT University); Mark Sanderson (RMIT University); Falk Scholer (RMIT University)

We investigate between-user drivers of query variability through a controlled between-subject, full-factorial user study that manipulates age, gender, and language proficiency across six backstory-driven search tasks. From initial queries, session logs, and post-task interviews, we quantify how demographic and task factors shape query-, task-, and session-level behaviors. We further derive a small set of interpretable latent search dimensions from user evidence to analyze and simulate heterogeneous query behavior. Our results show that age is the most consistent predictor of query formulation and search interaction patterns. Gender and language differences are more selective, and task context is further associated with these patterns. The latent dimensions help explain the differences as variation in search strategy rather than uniform differences in engagement or ability. The paper provides a trait-informed view of heterogeneous search behavior that supports more user-aware analysis and robustness-oriented evaluation in IR.

An Eye Tracking Study: Are AI Overviews Changing Search Behavior? [Full]

Sara Allawati (RMIT University); Dana McKay (RMIT University); Mark Sanderson (RMIT University); Paul Thomas (Microsoft); Johanne Trippas (RMIT University)

We conduct a lab eye-tracking study to examine how users interact with search engines that place Generative Artificial Intelligence (GenAI) results above traditionally ranked search results, also known as "ten blue links", and we use an engagement scale to evaluate their experience. Our aim is to study how users interact with search engine interfaces that incorporate GenAI content, assess users' willingness to scroll past GenAI content to view the traditional search results, and explore how these interactions differ from existing literature on scanning search engine result pages. % We show that GenAI content is changing how people search, but the "golden triangle" remains valid, where the top-left section of the search page attracts the most attention. Searchers are still engaging with the blue links in patterns consistent with the literature; however, they engage significantly more with GenAI content. Finally, we outline future directions to deepen our understanding of search behavior in the era of GenAI.

Cognitive Style Shapes Search Behaviours: An fNIRS Study of Exploratory Search [Full]

Huimin Tang (The University of Nottingham Ningbo China); Boon Giin Lee (The University of Nottingham Ningbo China); Dave Towey (The University of Nottingham Ningbo China); Max L. Wilson (University of Nottingham); Matthew Pike (The University of Nottingham Ningbo China)
Cognitive style, a user's habitual approach to information processing, offers a promising approach to personalising information searching (IS) systems, yet the underlying style-related search behaviours remain poorly understood. This study investigates how the Wholist–Analytic and Verbal–Imagery dimensions of cognitive style influence search behaviour and prefrontal cortex (PFC) activation during the Post-focus stage of exploratory search. Forty participants completed comparison search tasks while we recorded behavioural metrics, subjective workload ratings, and functional near-infrared spectroscopy (fNIRS) data. Our results demonstrate that cognitive style significantly predicted search engine results page (SERP) interaction patterns: analytics and imagers adopted structured navigation with detailed reading, whereas wholists and verbalisers preferred sporadic navigation with rapid scanning. Critically, fNIRS revealed distinct PFC activation patterns, specifically in the Ventrolateral PFC (VLPFC) and Dorsolateral PFC (DLPFC), corresponding to these behavioural differences. By mapping neuro-cognitive profiles to IR behaviours, this work provides empirical grounding for designing "style-aware" adaptive IR interfaces that tailor information density and navigational support to individual cognitive profiles.

Following the Eye-Tracking Evidence: Established Web-Search Assumptions Fail in Carousel Interfaces [Full]

Jingwei Kang (University of Amsterdam); Maarten de Rijke (University of Amsterdam); Harrie Oosterhuis (University of Amsterdam)
Carousel interfaces have been the de-facto standard for streaming media services for over a decade. Yet, there has been very little research into user behavior with such interfaces, which thus remains poorly understood. Due to this lack of empirical research, previous work has assumed that behaviors established in single-list web-search interfaces, such as the F-pattern and the examination hypothesis, also apply to carousel interfaces, for instance when designing click models or evaluation metrics. We analyze a recently-released interaction and examination dataset resulting from an eye-tracking study performed on carousel interfaces to verify whether these assumptions actually hold. We find that (i)~the F-pattern holds only for vertical examination and not for horizontal swiping; additionally, we discover that, when conditioned on a click, user examination follows an L-pattern unique to carousel interfaces; (ii)~click-through-rates conditioned on examination indicate that the well-known examination hypothesis does not hold in carousel interfaces; and (iii)~contrary to the assumptions of previous work, users generally ignore carousel headings and focus directly on the content items. Our findings show that many user behavior assumptions, especially concerning examination patterns, do not transfer from web search interfaces to carousel recommendation settings. Our work shows that the field lacks a reliable foundation on which to build models of user behavior with these interfaces. Consequently, a re-evaluation of existing metrics and click models for carousel interfaces may be warranted.

Agentic Search in the Wild: Intents and Trajectory Dynamics from 14M+ Real Search Requests [Full]

Jingjie Ning (Carnegie Mellon University); João Coelho (INESC-ID, Carnegie Mellon University); Yibo Kong (Carnegie Mellon University); Yunfan Long (Carnegie Mellon University); Bruno Martins (INESC-ID, Instituto Superior Técnico, University of Lisbon); João Magalhães (NOVA LINCS NOVA University Lisbon); Jamie Callan (Carnegie Mellon University); Chenyan Xiong (Carnegie Mellon University)
LLM-powered search agents are increasingly being used for multi-step information seeking tasks, yet the IR community lacks empirical understanding of how agentic search sessions unfold and how retrieved evidence is reflected in later queries. This paper presents a large-scale log analysis of agentic search based on 14.44M search requests (3.97M sessions) collected from DeepResearchGym, i.e., an open-source search API accessed by external agentic clients. We sessionize the logs, assign session-level intents and step-wise query-reformulation labels using LLM-based annotation, and propose Context-driven Term Adoption Rate (CTAR) to quantify whether newly introduced query terms are lexically traceable to previously retrieved evidence. Our analyses reveal distinctive behavioral patterns. First, over 90% of multi-turn sessions contain at most ten steps, and 89% of inter-step intervals fall under one minute. Second, behavior varies by intent. Fact-seeking sessions exhibit high repetition that increases over time, while sessions requiring reasoning sustain broader exploration. Third, query reformulations are often traceable to retrieved evidence across steps. On average, 54% of newly introduced query terms appear in the accumulated evidence context, with additional traceability to earlier steps beyond the most recent retrieval. These findings provide candidate signals for repetition-aware stopping, intent-adaptive retrieval budgeting, and explicit cross-step context tracking. We released the anonymized logs, making them available at a public HuggingFace repository.

Multi-Vector Retrieval · *Courtyard 1+2 · Tuesday*

13:30–15:00

Multi-Vector Index Compression in Any Modality [Full]

Hanxiang Qin (Johns Hopkins University); Alexander Martin (Johns Hopkins University); Rohan Jha (Johns Hopkins University); Chunsheng Zuo (Johns Hopkins University); Reno Kriz (Johns Hopkins University); Benjamin Van Durme (Microsoft)
We study efficient multi-vector retrieval for late interaction in any modality. Late interaction has emerged as a dominant paradigm for information retrieval across modalities, but its computation and storage costs grow linearly with document length, making it costly for multimodal corpora. To address this limitation, we explore methods for compressing multi-vector document representations under a constant vector budget. We introduce four approaches for index compression, including a novel attention-guided clustering (øurs). øurs uses an attention-guided mechanism to identify the most semantically salient parts of a document as cluster centroids and to weight token aggregation. Evaluating these methods on retrieval tasks spanning 4 modality settings (\beir, \vidore, \msrvtt, µltivent), we show that øurs consistently outperforms other parameterized compression methods (sequence resizing and memory tokens), provides greater flexibility in index size than non-parametric hierarchical clustering, and achieves competitive or improved performance compared to a full, uncompressed index.

Reproduction Beyond Benchmarks: ConstBERT and ColBERT-v2 Across Backends and Query Distributions [Reproducibility]

Utshab Kumar Ghosh (Missouri University of Science and Technology); Ashish David (Missouri University of Science and Technology); Shubham Chatterjee (Missouri University of Science and Technology)
Reproducibility must validate architectural robustness, not just numerical accuracy. We evaluate ColBERT-v2 and ConstBERT across five dimensions, finding that while ConstBERT reproduces within 0.05% MRR@10 on MS-MARCO, both models show a drop of 86–97% on long, narrative queries (TREC ToT 2025). Ablations prove this failure is architectural: performance plateaus at 20 words because the MaxSim operator's uniform token weighting cannot distinguish signal from filler noise. Furthermore, undocumented backend parameters create an 8-point gap due to ConstBERT's sparse centroid coverage, and fine-tuning with 3× more data actually degrades performance by up to 29%. We conclude that architectural constraints in multi-vector retrieval cannot be overcome by adaptation alone. Code: https://github.com/utshabkg/multi-vector-reproducibility.

Comparing Token Pruning Approaches for Multi-Vector Retrieval [Reproducibility]

Ferdinand Schlatt (Friedrich-Schiller-Universität Jena); Hanno Barschel (Friedrich-Schiller-Universität Jena); Matthias Hagen (Friedrich-Schiller-Universität Jena)
The computational costs of the transformer-based multi-vector retrieval model ColBERT depend on the number of vectors used to represent queries and documents. Common strategies to lower the costs thus prune the vectors to a fixed number or relative to the sequence length. We compare standard pruning approaches like weighted token pruning or IDF-based pruning and analyze the impact on the downstream effectiveness of respective ColBERT models. Our experiments indicate that weighted pruning can yield a better effectiveness--efficiency trade-off than other pruning techniques, but we also find that very simplistic pruning techniques can yield very effective ColBERT models when trained properly.

A Voronoi Cell Formulation for Principled Token Pruning in Late-Interaction Retrieval Models [Full]

Yash Kankanampati (Université Sorbonne Paris Nord); Yuxuan Zong (Sorbonne Université); Nadi Tomeh (Université Sorbonne Paris Nord); Benjamin Piwowarski (Sorbonne Université); Joseph Le Roux (Université Sorbonne Paris Nord)
Late-interaction models such as ColBERT offer competitive performance across various retrieval tasks but require storing a dense embedding for each document token, leading to a substantial index storage overhead. Past works address this by attempting to prune low-importance token embeddings based on statistical and empirical measures, but they often either lack formal grounding or are ineffective. To address these shortcomings, we introduce a framework grounded in hyperspace geometry and cast token pruning as a Voronoi cell estimation problem in the embedding space. By interpreting each token's influence as a measure of its Voronoi region, our approach enables principled pruning that retains retrieval quality while reducing index size. Through our experiments, we demonstrate that this approach serves not only as a competitive pruning strategy but also as a valuable tool for improving and interpreting token-level behavior within dense retrieval systems.

NumColBERT: Non-Intrusive Numeracy Injection for Late-Interaction Retrieval Models [Full]

Haruki Fujimaki (University of Tsukuba); Makoto P. Kato (University of Tsukuba)
This study addresses the challenge of improving dense retrieval performance for queries containing numerical conditions, such as "companies with more than one billion dollars in R&D expenditure." Although recent research has underscored the limitations of standard models in handling numeric information across domains such as finance, e-commerce, and medicine, existing solutions typically decompose queries into textual and numerical components and score them separately using dedicated methods. These approaches intrude upon late-interaction retrieval models such as ColBERT and incur considerable challenges in deployment, latency, and maintainability. To overcome these limitations, we propose NumColBERT, an inference-time non-intrusive method that enhances numerically conditioned retrieval while preserving the original late-interaction mechanism and providing unified scoring across textual and numerical content. Because NumColBERT retains the standard ColBERT indexing and MaxSim scoring pipeline, existing optimizations and ecosystem components developed for

ColBERT can be directly reused, facilitating practical deployment. NumColBERT introduces a Numerical Gating Mechanism and a Numerical Contrastive Learning objective to enable numerical conditions to contribute more effectively to retrieval within the standard ColBERT scoring mechanism. The gating mechanism dynamically amplifies the influence of tokens carrying critical numerical constraints while suppressing context-neutral mentions such as model numbers or dates. The contrastive objective explicitly shapes the embedding space to reflect numerical magnitudes and conditions, enabling numerical values to be distinguished within the shared representation space. Experimental results show that NumColBERT substantially outperforms standard fine-tuning baselines and achieves accuracy that matches or exceeds that of prior approaches that rely on separate textual and numerical scoring. These findings demonstrate the feasibility of numerically conditioned retrieval with a non-intrusive inference pipeline and present a maintainable solution for real-world deployment.

PLAID-PRF: Pseudo-Relevance Feedback with Centroid-like Tokens in PLAID [Full]

Xiao Wang (University of International Business and Economics); Sean MacAvaney (University of Glasgow); Craig Macdonald (University of Glasgow)

Multi-vector dense retrieval models, such as ColBERT, achieve strong retrieval effectiveness by modelling fine-grained token-level interactions between queries and documents. Methods such as PLAID use centroid-based quantisation of each token's vector to reduce the index size and speed up retrieval while maintaining strong effectiveness. In this work, we introduce PLAID-PRF, a method that performs Pseudo-Relevance Feedback (PRF) over PLAID to reformulate ColBERT's query vectors based on the top-retrieved results. In contrast with prior methods that perform PRF on multi-vector retrieval models, PLAID-PRF keeps computational costs low by leveraging the internal PLAID centroid vectors, treating them similarly to tokens in traditional PRF methods. The method selects a small and diverse set of high-utility expansion vectors and appends them to the original query, rerunning PLAID to refine both candidate generation and final scoring. Extensive experiments on the standard in-domain MSMARCO and four out-of-domain BEIR benchmarks show that PLAID-PRF consistently improves retrieval effectiveness over various baselines. In particular, PLAID-PRF improves over PLAID by up to 4.3% nDCG@10 and 7.3% MRR@10, while introducing substantially less computation overhead than prior PRF methods. The results demonstrate that our proposed centroid-aware PRF method offers an effective and lightweight mechanism to improve the quality of top-ranked retrieved results. Overall, this work enables effective and efficient feedback-aware late-interaction retrieval without expensive query-time document-token clustering.

LLMs for Recommendation · *Goldfields* · *Tuesday*

16:30–17:45

Bridging LLM Embeddings and VAE Parameters for Disentangled Recommendation [Full]

Nhu-Thuat Tran (Singapore Management University); Hady W. Lauw (Singapore Management University)

Disentangled recommendation within the Variational Autoencoder (VAE) framework aims to capture multiple user interests. While effective, these VAEs are fundamentally constrained by their reliance on interaction data alone, lacking the rich external semantic knowledge needed to properly structure and separate latent interests. Meanwhile, Large Language Models (LLMs) excel at deriving profound user preference signals from textual data. Prevailing methods for integrating LLMs into recommendation, however, either focus on single-interest modeling or perform a shallow fusion by aligning LLM and VAE representation spaces. Thus, they fail to fundamentally shape the VAE's latent space for multi-interest learning, hindering recommendation performance. To bridge this gap, we propose to fundamentally shift the integration point: instead of aligning representation spaces, we bridge the LLM-generated embedding space directly to the VAE's parameter space. Our framework designs hypernetworks to generate parameters governing interest discovery of a disentangled VAE conditioned on LLM-derived user embeddings. This direct bridge injects rich semantic knowledge into the model's learning foundation, preserving the VAE's power for disentangled representation while dramatically enhancing its modeling capability with informative, LLM-structured priors. Extensive experiments on benchmark datasets show that our method yields significant performance gains, unveiling the untapped potential of using LLMs for disentangled recommendation.

Modular Representation Compression: Adapting LLM Representations for Efficient and Effective Recommendation [Full]

Yunjia Xi (Tencent Youtu Lab); Menghui Zhu (Huawei Noah's Ark Lab); Jianghao Lin (Shanghai Jiao Tong University); Bo Chen (Huawei Noah's Ark Lab); Ruiming Tang (Huawei Noah's Ark Lab); Yong Yu (Shanghai Jiao Tong University); Weinan Zhang (Shanghai Jiao Tong University)

Recently, large language models (LLMs) have advanced recommendation systems (RSs), and recent works have begun to explore how to integrate LLMs into industrial RSs. While most approaches deploy LLMs offline to generate and pre-cache augmented representations for RSs, high-dimensional representations from LLMs introduce substantial storage and computational costs. Thus, it is crucial to compress LLM representations effectively. However, we identify a counterintuitive phenomenon during representation compression: Mid-layer Representation Advantage (MRA), where representations from middle layers of LLMs outperform those from final layers in recommendation tasks. This degraded final layer renders existing compression methods, which typically compress on the final layer, suboptimal. We interpret this based on modularity theory that LLMs develop spontaneous internal functional modularity and force the final layer to specialize in the proxy training task. Thus, we propose Modular Representation Compression (MARC) to explicitly control the modularity of LLMs. First, Modular Adjustment explicitly introduces compression and task adaptation modules, enabling the LLM to operate strictly as a representation-learning module. Next, to ground each module to its specific task, Modular Task Decoupling uses information constraints and different network structures to decouple tasks. Extensive experiments validate that MARC addresses MRA and produces efficient representations. Notably, MARC achieved a 2.82% eCPM lift in an online A/B test within a large-scale commercial search advertising scenario.

Rethinking Semantic–Collaborative Integration: Why Alignment Is Not Enough [Perspective]

Maolin Wang (City University of Hong Kong); Dongze Wu (Beihang University); Jianing Zhou (Beihang University); Hongyu Chen (City University of Hong Kong); Beining Bao (City University of Hong Kong); Yu Jiang (Chinese University of Hong Kong); Chenbin Zhang (Unaffiliated); Chang Wang (Jingdong); Jian Liu (Beihang University); Lei Sha (Beihang University)

Large language models (LLMs) have become an important semantic infrastructure for modern recommender systems. A prevailing paradigm integrates LLM-derived semantic embeddings with collaborative representations via representation alignment, implicitly assuming that the two views encode a shared latent entity and that stronger alignment yields better results. We formalize this assumption as the global low-complexity alignment hypothesis and argue that it is stronger than necessary and often structurally mismatched with real-world recommendation settings. We propose a complementary perspective in which semantic and collaborative representations are treated as partially shared yet fundamentally heterogeneous views, each containing both shared and view-specific factors. Under this shared-plus-private latent structure, enforcing global geometric alignment may distort local structure, suppress view-specific signals, and reduce informational diversity. To support this perspective, we develop complementarity-aware diagnostics that quantify overlap, unique-hit contribution, and theoretical fusion upper bounds. Empirical analyses on sparse recommendation benchmarks reveal low item-level agreement between semantic and collaborative views and substantial oracle fusion gains, indicating strong complementarity. Furthermore, controlled alignment probes show that low-capacity mappings capture only shared components and fail to recover full collaborative geometry, especially under distribution shift. These findings suggest that alignment should not be treated as the default integration principle. We advocate a shift from alignment-centric modeling to complementarity fusion-centric, complementarity-aware design, where shared factors are selectively integrated while private signals are preserved. This reframing provides a principled foundation for the next generation of LLM-enhanced recommender systems.

Filling the Gaps: Selective Knowledge Augmentation for LLM Recommenders [Full]

Jaehyun Lee (Pohang University of Science and Technology); Sanghwan Jang (Pohang University of Science and Technology); SeongKu Kang (Korea University); Hwanjo Yu (Pohang University of Science and Technology)

Large language models (LLMs) have recently emerged as powerful training-free recommenders. However, their knowledge of individual items is inevitably uneven due to imbalanced information exposure during pretraining, a phenomenon we refer to as knowledge gap problem. To address this, most prior methods have employed a naive uniform augmentation that appends external information for every item in the input prompt. However, this approach not only wastes limited context budget on redundant augmentation for well-known items but can also hinder the model's effective reasoning. To this end, we propose KnowSACKP(Knowledge-aware Selective Augmentation with Comparative Knowledge Probing) to mitigate the knowledge gap problem. KnowSACKP estimates the LLM's internal knowledge by evaluating its capability to capture collaborative relationships and selectively injects additional information only where it is most needed. By avoiding unnecessary augmentation for well-known items, KnowSACKP focuses on items that benefit most from knowledge supplementation, thereby making more effective use of the context budget. KnowSACKP requires no fine-tuning step, and consistently improves both recommendation accuracy and context efficiency across four real-world datasets. Our code is available at https://github.com/nowhyun/KnowSA_CKP.

RecPFN: Prior-Fitted Networks for In-Context-Based Recommendations [Full]

En Zhi Tan (SAP SE); Jia Xiang Lim (SAP SE); Bryan Lijie Chew (SAP SE); Tze Minh Ng (SAP SE); Benjamin Yan Han Yap (SAP SE)

We introduce RecPFN, a prior-fitted network that brings in-context learning to sequential recommendation. RecPFN is pretrained entirely on synthetic clickstream environments sampled from a broad structural causal prior, enabling it to amortize Bayesian-style inference from a small support set. At inference, a lightweight decoder-only transformer conditions on a handful of domain sequences and produces next-item predictions for queries in a single forward pass, without any weight updates. Across eight public benchmarks, RecPFN achives state-of-the-art zero-shot performance while remaining strongly competitive with supervised methods in low-compute and low-data regimes. It is deployment-efficient and robust to domain shift, outperforming strong zero-shot baselines that rely on large real-interaction corpora.

RecPFN provides a practical path toward generalizable, data-efficient recommenders and opens avenues for richer priors, longer-context ICL, and multimodal extensions. Code for training and evaluation will be made publicly available by the conference date.

Graph Representation Learning · *Room 110 · Tuesday 16:30–17:45*

Graph2text or Graph2token: A Perspective of Large Language Models for Graph Learning [TOIS]

Shuo Yu (School of Computer Science and Technology, Dalian University of Technology, Dalian, China); Yingbo Wang (Dalian University of Technology, Dalian, China); Ruolin Li (School of Software, Dalian University of Technology, Dalian, China); Guchun Liu (Dalian University of Technology, Dalian, China); Yanming Shen (School of Computer Science and Technology, Dalian University of Technology, Dalian, China); Shaoxiong Ji (Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany); Bowen Li (School of Computing Technologies, RMIT University, Melbourne, Australia); Fengling Han (School of Computing Technologies, RMIT University, Melbourne, Australia); Xiuzhen Zhang (School of Computing Technologies, RMIT University, Melbourne, Australia); Feng Xia (School of Computing Technologies, RMIT University, Melbourne, Australia)

Graphs are prevalent in numerous real-world applications. Previous methods directly model graph structures and achieve significant success. However, these methods encounter bottlenecks due to the inherent irregularity of graphs. An innovative solution is converting graphs into textual representations, thereby harnessing the powerful capabilities of Large Language Models (LLMs) to process and comprehend graphs. In this paper, we present a comprehensive review of methodologies for applying LLMs to graphs, termed LLM4graph. The core of LLM4graph lies in transforming graphs into texts for LLMs to understand and analyze. Thus, we propose a novel taxonomy of LLM4graph methods from the view of the transformation. Specifically, existing methods can be divided into two paradigms: Graph2text and Graph2token, which transform graphs into texts or tokens as the input of LLMs, respectively. We point out four challenges during the transformation to systematically present existing methods from a problem-oriented perspective. For practical concerns, we provide a guideline for researchers on selecting appropriate models and LLMs for different graphs and hardware constraints. To empirically evaluate our taxonomy and different technical choices, we conduct experiments with representative methods in Graph2text and Graph2token. We also identify five future research directions for LLM4graph.

Learning Noise-Resilient and Transferable Graph-Text Alignment via Dynamic Quality Assessment [Full]

Yuhang Liu (Tianjin University); Minglai Shao (Tianjin University); Zengyi Wo (Baidu); Yunlong Chu (Tianjin University); Bing Hao (Tianjin University); Shengzhong Liu (Shanghai Jiao Tong University); Ruijie Wang (Beihang University); Jianxin Li (Beihang University)

Pre-training graph foundation models (GFMs) on text-attributed graphs (TAGs) is important for web-scale retrieval and recommendation, where graph entities are matched with textual descriptions. Existing CLIP-style graph-text aligners typically assume one-to-one correspondence: each node is pulled close only to its paired text, and all other pairs are treated as negatives. This overlooks the many-to-many relations common in real TAGs, where a node and its local neighborhood can be semantically related to multiple texts, and vice versa. Meanwhile, TAG supervision is often imperfect: noisy or weak node-text links introduce false-positive pairs, causing contrastive learning to align mismatched semantics. These limitations reveal a fundamental trade-off: leveraging expressive many-to-many signals increases semantic coverage but may propagate errors under noise, whereas strict one-to-one training is more conservative yet still suffers when mismatched pairs remain in the training set. Therefore, we propose ADAligner, a quality-aware graph–text alignment framework that adapts between expressive many-to-many and conservative one-to-one objectives based on estimated alignment reliability. ADAligner tracks batch-level reliability online and adjusts optimization accordingly—promoting soft, subgraph-level alignment when supervision is clean while emphasizing reliable one-to-one alignment by filtering low-confidence pairs under noise. We provide theoretical analysis showing that this closed-loop adaptation is stable and convergent. Experiments on nine TAG benchmarks show that, under 30% mismatched node-text supervision, ADAligner consistently improves cross-modal retrieval by 144.70% on average, zero-/few-shot node classification by 26.13%, and link prediction by 4.70% over the strongest multimodal baseline, demonstrating strong robustness to alignment noise across both unsupervised and transfer settings. Our code is available at https://github.com/karmaisacat-13/ADAligner.

Mitigating Structural Overfitting: A Distribution-Aware Rectification Framework for Missing Feature Imputation [Full]

Yifan Song (The Hong Kong University of Science and Technology (Guangzhou)); Fenglin Yu (Carnegie Mellon University); Yihong Luo (Hong Kong University of Science and Technology); Xingjian Tao (The Hong Kong University of Science and Technology (Guangzhou)); Siya Qiu (Hong Kong University of Science and Technology); Kai Han (Shanghai University of Finance and Economics); Jing Tang (The Hong Kong University of Science and Technology (Guangzhou))

Incomplete node features are ubiquitous in real-world scenarios such as user profiling and cold-start recommendation, which severely hinders the practical deployment of graph learning systems (e.g., GNNs). Existing solutions typically rely on diffusion-based structural smoothing (e.g., feature propagation) to impute missing values. However, we find that these approaches suffer from structural overfitting, leading to three progressive challenges: 1) performance degradation on disjoint graphs, 2) loss of semantic diversity due to over-smoothing, and 3) feature distribution shift when generalizing to unseen graph structures (inductive tasks). To address these challenges, we introduce the DART framework. It begins by employing Global Structural Augmentation (GSA), which establishes global correlations to bridge disjoint components and extend diffusion coverage. Building upon this, we design a semantic rectifier based on masked autoencoding. This module learns the latent feature manifold to recover natural semantic details. Crucially, we introduce a test-time distribution rectification mechanism that projects structurally biased features back onto the learned manifold during inference, effectively bridging the inductive distribution gap. Furthermore, considering that synthetic masking fails to reflect realworld sparsity, we present a new dataset Sailing collected from voyage records with naturally missing attributes. Extensive experiments on six public datasets and Sailing demonstrate that DART significantly outperforms state-of-the-art methods in both transductive and inductive settings. Our code and dataset are available at https://github.com/yfsong00/DART.

Hierarchical Cluster-based Open-World Graph Active Learning [Full]

Yayong Li (The University of Queensland); Zhengyi Du (Lanzhou University of Technology); Hong Zhang (Lanzhou University of Technology); Jonathan Wilton (The University of Queensland); Jinran Wu (The University of Queensland); Zongli Liu (Lanzhou University of Technology); Nan Ye (The University of Queensland)

Existing works in graph active learning (GAL) mostly assume a closed-world setting where all classes are known in advance, while in practice, we need to deal with the open-world setting where novel classes are encountered during learning. Apart from selecting informative nodes for refining the current classifier, the open-world GAL algorithms are also expected to discover novel classes. Motivated by the observation that identifying an informative region can be easier than finding the most informative example, we propose a novel hierarchical cluster-based algorithm for open-world GAL. Our algorithm performs clustering in the feature space to identify an informative region which either contains informative examples for known classes or novel classes, then performs a novel semi-supervised sub-clustering on the selected cluster to identify the most informative example. To facilitate the discovery of informative regions, we introduce a cluster-based self-distillation loss between ego and final embeddings to learn well-clustered node embeddings that are more likely to align with the classes. We also introduce a novel cluster informativeness score for the clusters, which measures not only how uncertain the cluster is (similar to standard active learning algorithms), but whether the cluster is likely to contain novel classes. Our semi-supervised sub-clustering algorithm partitions the selected cluster into subregions for known classes and an additional uncertain subregion, then we query the label for a representative node for the uncertain subregion. Extensive experiments on five benchmark datasets demonstrate the effectiveness of the proposed method with a more balanced improvement over novel classes. Ablation study shows that our cluster-based self-distillation loss, informativeness score, and the semi-supervised sub-clustering strategy are all beneficial. Empirical analysis also reveals how our informativeness score is effective for novel class discovery and identifying informative examples for refining the decision boundary.

Inductive Subgraphs as Shortcuts: Causal Disentanglement for Heterophilic Graph Learning [Full]

Xiangmeng Wang (The Hong Kong Polytechnic University); Qian Li (Curtin University); Haiyang Xia (University of Macau); Hao Miao (The Hong Kong Polytechnic University); Qing Li (The Hong Kong Polytechnic University); Guandong Xu (The Education University of Hong Kong)

Heterophily is a prevalent property of real-world graphs and is well known to impair the performance of homophilic Graph Neural Networks (GNNs). Prior work has attempted to adapt GNNs to heterophilic graphs through non-local neighbor extension or architecture refinement. However, the fundamental reasons behind misclassifications remain poorly understood. In this work, we take a novel perspective by examining recurring inductive subgraphs, empirically and theoretically showing that they act as spurious shortcuts that mislead GNNs and reinforce non-causal correlations in heterophilic graphs. To address this, we adopt a causal inference perspective to analyze and correct the biased learning behavior induced by shortcut inductive subgraphs. We propose a debiased causal graph that explicitly blocks confounding and spillover paths responsible for these shortcuts. Guided by this causal graph, we introduce Causal Disentangled GNN (CD-GNN), a principled framework that disentangles spurious inductive subgraphs from true causal subgraphs by explicitly blocking non-causal paths. By focusing on genuine causal signals, CD-GNN substantially improves the robustness and accuracy of node classification in heterophilic graphs. Extensive experiments on real-world datasets not only validate our theoretical findings but also demonstrate that our proposed CD-GNN outperforms state-of-the-art heterophily-aware baselines.

Dense Retrieval · *Eureka 1 · Tuesday 16:30–17:45*

Internalizing Explicit Reasoning into Latent Space for Dense Retrieval [Full]

Jiajie Jin (Gaoling School of Artificial Intelligence, Renmin University of China); Yanzhao Zhang (Alibaba Group); Mingxin Li (Alibaba Group);

Dingkun Long (Alibaba Group); Pengjun Xie (Alibaba Group); Yutao Zhu (Gaoling School of Artificial Intelligence, Renmin University of China); Zhicheng Dou (Gaoling School of Artificial Intelligence, Renmin University of China)

Large Language Models (LLMs) have fundamentally transformed dense retrieval, upgrading backbones from discriminative encoders to generative architectures. However, a critical disconnect remains: while LLMs possess strong reasoning capabilities, current retrievers predominantly utilize them as static encoders, leaving their potential for complex reasoning unexplored. To address this, existing approaches typically adopt "rewrite-then-retrieve" pipelines to generate explicit Chain-of-Thought (CoT) rationales before retrieval. However, this incurs prohibitive latency. Conversely, implicit reasoning methods utilizing latent tokens offer efficiency but often suffer from semantic degeneration due to the lack of explicit supervision. In this paper, we propose LaSER, a novel self-distillation framework that internalizes explicit reasoning into the latent space of dense retrievers. Operating on a shared LLM backbone, LaSER introduces a dual-view training mechanism: an Explicit view that explicitly encodes ground-truth reasoning paths, and a Latent view that performs implicit latent thinking. To bridge the gap between these views, we design a multi-grained alignment strategy. Beyond standard output alignment, we introduce a trajectory alignment mechanism that synchronizes the intermediate latent states of the latent path with the semantic progression of the explicit reasoning segments. This allows the retriever to "think" silently and effectively without autoregressive text generation. Extensive experiments on both in-domain and out-of-domain reasoning-intensive benchmarks demonstrate that LaSER significantly outperforms state-of-the-art baselines. Furthermore, analyses across diverse backbones and model scales validate the robustness of our approach, confirming that our unified learning framework is essential for eliciting effective latent thinking. Our method successfully combines the reasoning depth of explicit CoT pipelines with the inference efficiency of standard dense retrievers. The code, model, and training data are available at https://github.com/RUC-NLPIR/LaSER.

Recasting Web-Scale Query Suggestion as dense retrieval: Efficient, Up-to-Date, and Context-Aware Suggestions [Full]

Sosuke Nishikawa (The University of Tokyo); Naoki Yoshinaga (Institute of Industrial Science, The University of Tokyo); Nobuhiro Kaji (LY Corporation)

Query suggestion (QS) in web search must provide efficient and up-to-date suggestions that are relevant to the session context. Recent studies treat QS as generation, which captures session context well but suffers from high latency and costly retraining to maintain information freshness. In this study, we recast QS as dense retrieval: given a search session, the next query is retrieved from a large index of historical queries using efficient approximate nearest neighbor search. We propose Context-aware Asymmetric Dual Encoder for QS (CADE-QS), which uses an asymmetric dual encoder to model query -to- suggestion dependency and incorporates session history into the query encoder via context-aware contrastive learning. We evaluate our method on two real-world search-log datasets: a recent Japanese web search log and the public AOL log. CADE-QS rivals strong generative baselines in quality while reducing end-to-end latency to around 30 ms on CPU, representing an order-of-magnitude improvement. Detailed analyses confirm CADE-QS's robust context awareness, unidirectional modeling, and practicality for reflecting evolving information via index refreshes, as well as the effectiveness of an adaptive hybrid strategy for low-coverage scenarios. Our code is publicly available at https://github.com/lycorp-jp/cadeqs.

Scaling Laws for Embedding Dimension in Information Retrieval [Full]

Julian Killingback (University of Massachusetts Amherst); Mahta Rafiee (University of Massachusetts Amherst); Manas Madine (University of Massachusetts Amherst); Hamed Zamani (University of Massachusetts Amherst)

Dense retrieval, which encodes queries and documents into a single dense vector, has become the dominant neural retrieval approach due to its simplicity and compatibility with fast approximate nearest neighbor algorithms. As the tasks dense retrieval performs grow in complexity, the fundamental limitations of the underlying data structure and similarity metric---namely vectors and inner-products---become more apparent. Recent work has shown theoretical limitations inherent to single vectors and inner-products, which are generally tied to the embedding dimension. Given the importance of embedding dimension for retrieval capacity, understanding how dense retrieval performance changes as embedding dimension is scaled is fundamental to building next-generation retrieval models that balance effectiveness and efficiency. In this work, we conduct a comprehensive analysis of the relationship between embedding dimension and retrieval performance. Our experiments include a range of model sizes from two model families to construct a detailed picture of embedding scaling behavior. We find that the scaling behavior fits a power law, allowing us to derive scaling laws for performance given only the embedding dimension, as well as a joint law accounting for embedding dimension and model size. Our analysis shows that for evaluation tasks aligned with the training task, performance continues to improve as embedding size increases, though with diminishing returns. For evaluation data that are less aligned with the training task, we find that performance is less predictable, with performance degrading with larger embedding dimensions for certain tasks. We hope our work provides additional insight into the limitations of embeddings and their behavior as well as offers a practical guide for selecting model and embedding dimension to achieve optimal performance with reduced storage and compute costs.

A Replicability Study of Joint Product Quantisation for Effective Space-Efficient Dense Retrieval [Reproducibility]

Craig Macdonald (University of Glasgow); Zhlli Shen (University of Glasgow); Nicola Tonellotto (University of Pisa)

Dense retrieval uses embedded dense representations for documents to be obtained from language models ingested in vector stores. To reduce large space requirements and/or efficiency concerns, approximate nearest neighbours algorithms can be deployed, typically at the cost of effectiveness. Among existing approaches, Product Quantisation (PQ) is one well-known algorithm, where embeddings are split along dimensions, and then assigned to similar cluster centroids on a per-split basis. Joint Product Quantisation (JPQ) improves over PQ by further training the obtained centroid embeddings, as well as the query encoder, for enhanced effectiveness. In this replicability study, we reimplement JPQ, comparing its performance on the STAR biencoder model with that reported in the original paper, as well as considering its generalisability in several dimensions: to other biencoders such as TCT and TAS-B; to larger models such as RepLLama; and beyond document retrieval, to a retrieval augmented generation task. Among many interesting results, we find that JPQ does work on more modern biencoder models such as TCT but not on larger models such as RepLLama; however, it does scale to large corpora, allowing dense retrieval for RAG that is as effective from an index that is 32x larger.

Lighting the Way for BRIGHT: Reproducible Baselines with Anserini, Pyserini, and RankLLM [Reproducibility]

Sahel Sharifymoghaddam (University of Waterloo); Yijun Ge (University of Waterloo); Raghav Vasudeva (University of Waterloo); Jimmy Lin (University of Waterloo)

Retrieval benchmarks for large language models (LLMs) should reflect the long, reasoning-intensive queries typical of retrieval-augmented generation (RAG). We present a systematic study of BRIGHT, a reasoning-focused retrieval benchmark, along with strong, reproducible reference methods integrated into Anserini, Pyserini, and RankLLM. We evaluate lexical, sparse, dense, and fusion-based retrievers, as well as LLM rerankers, under long-query settings. In reproducing BRIGHT's lexical baseline, we identify a key under-documented detail: query-side BM25 (BM25Q), which applies BM25 weighting to the query itself. On long, multi-sentence queries, BM25Q consistently outperforms standard BM25, making it the strongest lexical baseline for reasoning-oriented retrieval. We further audit the BRIGHT corpus, uncovering data quality issues that impact evaluation, and offer mitigation. Finally, we study the generalizability of BM25Q across five additional benchmarks, finding its gains largely specific to BRIGHT, while fusion with standard BM25 provides the most consistent improvements across datasets.

Social Media Analysis · *Eureka 2 · Tuesday 16:30–17:45*

SNBot: Modeling Self–Neighborhood Representation Discrepancy for Social Bot Detection [Full]

Qilong Lin (Soochow University); Jingya Zhou (Soochow University)

The proliferation of social bots poses a persistent threat to online social platforms, making accurate and efficient detection increasingly critical. Although recent graph-based methods have achieved notable progress by modeling user interactions, many of them implicitly assume neighborhood consistency and degrade in sparse, directed, and heterophilic social graphs, where a user's neighborhood may poorly reflect its own attributes. In this work, we propose SNBot, a novel social bot detection framework that explicitly models the discrepancy between node self-representations and their neighborhood embeddings. SNBot first learns unified self-representations from multi-modal user attributes through a lightweight type-aware encoding scheme. It then performs message passing on an augmented directed graph using a bidirectional aggregation mechanism, which separately captures incoming and outgoing interactions to better characterize asymmetric behaviors. By preserving and exploiting self–neighborhood representation discrepancies rather than over-smoothing them, SNBot produces more discriminative node representations. Extensive experiments on multiple real-world benchmark datasets demonstrate that SNBot consistently outperforms state-of-the-art methods.

Depression Detection from Social Media: A Mutual Guidance Multi-modal Network with Complementary Graph Learning [Full]

Guocheng Hu (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)); Chaoqun Zheng (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)); Ruifan Zuo (Shandong University); Fengling Li (University of Technology Sydney); Dan Shi (Nanjing University of Science and Technology); Xiaofeng Qu (University of Jinan); Wenpeng Lu (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences))

Depression has become a critical global public health challenge, creating an urgent need for automated and scalable screening solutions. Social media platforms, which capture rich and spontaneous multi-modal behavioral data, offer a promising avenue for detecting early signs of mental distress. However, existing depression detection methods predominantly rely on static multi-modal fusion strategies and frequently fail to effectively tackle cross-modal semantic gaps. To address these limitations, we propose a Mutual Guidance Multi-modal Network with Complementary Graph Learning

(MGMN) for depression detection by observing individuals' behavioral performance on social media. Specifically, a cross-modal mutual guidance mechanism is designed to dynamically construct a complementary graph by using mutual similarities within and across visual and acoustic modalities common in social media. More specifically, based on this complementary graph, a modality-specific adaptive residual learning module is applied to each modality to stabilize deep feature learning and preserve modality-specific and complementary information via graph-conditioned adaptive residual fusion. Furthermore, the refined uni-modal features are subsequently fed into a joint-modal fusion and prediction module to output the final disease prediction probability. Extensive experiments on the MUD3, LMVD, and D-vlog datasets demonstrate our proposed method's superiority over state-of-the-art methods, confirming that the proposed framework provides a robust and effective solution for mental health monitoring. Codes are available at https://github.com/Petofi-romance/MGMN

Deja Vu in Plots: Leveraging Cross-Session Evidence with Retrieval-Augmented LLMs for Live Streaming Risk Assessment [Full]

Yiran Qiao (Institute of Computing Technology, Chinese Academy of Sciences); Xiang Ao (Institute of Computing Technology, Chinese Academy of Sciences); Jing Chen (Bytedance China); Yang Liu (Institute of Computing Technology, Chinese Academy of Sciences); Qiwei Zhong (ByteDance China); Qing He (Institute of Computing Technology, Chinese Academy of Sciences)

The rise of live streaming has transformed online interaction, enabling massive real-time engagement but also exposing platforms to complex risks such as scams and coordinated malicious behaviors. Detecting these risks is challenging because harmful actions often accumulate gradually and recur across seemingly unrelated streams. To address this, we propose CS-VAR (Cross-Session Evidence-Aware Retrieval-Augmented Detector) for live streaming risk assessment. In CS-VAR, a lightweight, domain-specific model performs fast session-level risk inference, guided during training by a Large Language Model (LLM) that reasons over retrieved cross-session behavioral evidence and transfers its local-to-global insights to the small model. This design enables the small model to recognize recurring patterns across streams, perform structured risk assessment, and maintain efficiency for real-time deployment. Extensive offline experiments on large-scale industrial datasets, combined with online validation, demonstrate the state-of-the-art performance of CS-VAR. Furthermore, CS-VAR provides interpretable, localized signals that effectively empower real-world moderation for live streaming.

HCPRA: A Hierarchical Cognition–Perception–Reasoning Agent Framework for Emotion-Cause Pair Extraction in Conversations [Full]

Botao Wang (Northwestern Polytechnical University); Lianwei Wu (Northwestern Polytechnical University); Shuhan Guo (Northwestern Polytechnical University); Kang Wang (Northwestern Polytechnical University); Qingyan Wang (Northwestern Polytechnical University); Tingran Zhang (Northwestern Polytechnical University); Jiapeng Liu (Xi'an Jiaotong University); Hikmat Ullah Khan (University of Sargodha)

Emotion-Cause Pair Extraction in Conversations (ECPEC) aims to identify speakers' emotions and the corresponding causes within conversation contexts. This task has gained increasing attention in recent years. Existing methods typically operate from the perspective of text object, inputting the entire conversation text and relying on neural networks to highlight utterance features as a means of emotion perception, while employing pair concatenation as a single path cause reasoning strategy. However, in conversation scenarios, humans are the true origin of emotions, while language text merely serves as the medium of expression. Consequently, these methods lack the modeling of the cognition and reasoning processes behind the human's behavior from the perspective of the speaker subject. To address this issue, we propose a novel Hierarchical Cognition–Perception–Reasoning Agent (HCPRA) framework. The framework is oriented around the speaker, simulating their personality cognition and conversation scenario cognition. Moreover, it utilizes hybrid memory and implicit emotion enhancement to simulate the human emotion perception. Additionally, it employs a hierarchical emotion cause reasoning mechanism to extract interpretable relationships between the speaker's emotions and their causes through emotion awareness and multi-path cause reasoning. Experimental results demonstrate that our approach achieves state-of-the-art performance on three datasets, with F1 improvement up to 15.61%.

Decoding Multimodal Cues: Unveiling the Implicit Meaning Behind Hateful Videos [Full]

Junyu Lu (Dalian University of Technology); Deyi Ji (Tencent); Liqun Liu (Tencent); Xiaokun Zhang (City University of Hong Kong); Youlin Wu (Dalian University of Technology); Roy Ka-Wei Lee (Singapore University of Technology and Design); Peng Shu (Tencent); Huan Yu (Tencent); Jie Jiang (Tencent); Bo Xu (Dalian University of Technology); Liang Yang (Dalian University of Technology); Hongfei Lin (Dalian University of Technology)

Hateful videos have become prevalent on online platforms, highlighting an urgent need for effective detection. However, existing studies primarily focus on binary classification and fail to provide contextual rationales that reveal the implicit meanings behind these judgments, significantly undermining model explainability. To fill this gap, we aim to achieve explainable hateful video detection, enabling models to provide contextual rationales that integrate relevant evidence and logical reasoning alongside decisions. This approach can comprehensively enhance the understanding of video content and the explainability of the decision-making process. We first introduce two datasets, Ex-HateMM and Ex-ImpliHateVid, for explainable hateful video detection. Each dataset provides fine-grained annotations of multimodal harmful elements, along with contextual rationales. We then propose an Information Augmentation and Reasoning Enhancement (IARE) framework designed for explainable detection. The framework employs an information augmentation phase that leverages the multimodal chain-of-thought to integrate harmful elements, thereby enriching rationale evidence. Additionally, IARE incorporates a reasoning enhancement phase, in which Direct Preference Optimization guides the model toward correct reasoning paths and away from incorrect ones, thereby improving the logical coherence of its justifications. We conduct extensive experiments on the two datasets, comparing multiple baselines with our proposed IARE framework. The results demonstrate that IARE achieves state-of-the-art performance while also generating accurate rationales.

Fairness and Responsible IR · *Eureka 3 · Tuesday 16:30–17:45*

Post-hoc Provider Fairness Adaptation via Hierarchical Exposure Alignment [Full]

Jingzhi Li (Hefei University of Technology); Zhiyong Cheng (Hefei University of Technology); Richang Hong (Hefei University of Technology); Meng Wang (Hefei University of Technology)

Provider exposure fairness is crucial for sustaining a healthy content ecosystem and preventing monopolization in recommender systems. Yet, most existing methods either incorporate fairness constraints during model training, requiring expensive retraining when fairness objectives change, or rely on post-hoc reranking with fixed criteria, which lacks adaptability to diverse fairness requirements. To overcome these limitations, we propose Post-hoc Fairness Adaptation (PFA), a lightweight framework that equips a frozen recommender with a fairness adapter, enabling flexible fairness control without retraining the backbone model. Specifically, the fairness adapter learns personalized additive score adjustments from user–item embeddings, which are injected into the original ranking scores to steer provider exposure toward fairness. To train the adapter, we minimize the Kullback-Leibler (KL) divergence between the actual and the target fair exposure distributions. However, this global objective implicitly treats all providers equally, ignoring structural disparities such as imbalanced provider group sizes and heterogeneous exposure within groups. Consequently, fairness may appear satisfied at an aggregate level while severe inter-group and intra-group exposure imbalances persist, undermining practical fairness. To address this, we design Hierarchical Exposure Fairness Alignment (HEFA), which explicitly balances inter- and intra-group provider exposure disparities, enabling flexible adaptation to diverse fairness requirements. To mitigate potential accuracy degradation, PFA jointly optimizes HEFA with a differentiable NDCG loss, enabling end-to-end fairness optimization while preserving ranking quality. Extensive experiments on three public datasets demonstrate that PFA achieves substantial fairness gains with negligible accuracy loss, consistently outperforming strong baselines. Code is available at https://github.com/Tam-JQK/Post-train.

Equity vs. Equality: Optimizing Ranking Fairness for Tailored Provider Needs [Full]

Yiteng Tu (DCST, Tsinghua University); Weihang Su (DCST, Tsinghua University); Shuguang Han (Alibaba Group); Yiqun Liu (Tsinghua University); Qingyao Ai (Quancheng Laboratory)

Ranking plays a central role in connecting users and providers in Information Retrieval (IR) systems, making provider-side fairness an important challenge. While recent research has begun to address fairness in ranking, most existing approaches adopt an equality-based perspective, aiming to ensure that providers with similar content receive similar exposure. However, it overlooks the diverse needs of real-world providers, whose utility from ranking may depend not only on exposure but also on outcomes like sales or engagement. Consequently, exposure-based fairness may not accurately capture the true utility perceived by different providers with varying priorities. To this end, we introduce an equity-oriented fairness framework that explicitly models each provider's preferences over key outcomes such as exposure and sales, thus evaluating whether a ranking algorithm can fulfill these individualized goals while maintaining overall fairness across providers. Based on this framework, we develop EquityRank, a gradient-based algorithm that jointly optimizes user-side effectiveness and provider-side equity. Extensive offline and online simulations demonstrate that EquityRank offers improved trade-offs between effectiveness and fairness and adapts to heterogeneous provider needs.

FairSpec: Expert Specialization for Fair LLM-based Recommendation [Full]

Yuchen Zheng (Nankai University); Xuan Pan (Tiangong University); Jing Wang (Nankai University); Chuanchang Zhang (Nankai University); Xi Lin (Shanghai Jiao Tong University); Chunyao Song (Nankai University); Xiangrui Cai (Nankai University); Xiaojie Yuan (Nankai University)

The integration of Large Language Models (LLMs) into recommender systems has introduced a powerful paradigm for personalized recommendation. However, due to biases stemming from sensitive user attributes (e.g., gender, age), LLMs may generate discriminatory and unfair recommendations, even for users with similar preferences. Existing methods to improve LLM-based recommendation fairness, like prompt engineering or data-processing, are often unstable and fail to address unfairness rooted in pre-training data and model parameters. While fine-tuning approaches show

promise, improving fairness typically comes at the cost of degraded recommendation performance, as the model's parameters conflates user preferences with sensitive attributes. To address this, we propose \textbf{FairSpec}, a novel lightweight fine-tuning framework designed to enhance fairness while preserving recommendation quality. FairSpec has a Parameter-Efficient Mixture-of-Experts architecture (PE-MoE) with a fairness-aware router. The experts are specialized into utility experts and fairness experts, and the latter are adversarially trained to suppress sensitive information. Furthermore, FairSpec introduces Fairness-Aware Direct Preference Optimization (FA-DPO), which employs a composite reward function to jointly optimize recommendation performance and fairness. Results on three public benchmarks with four LLM backbones demonstrate that FairSpec achieves the best overall performance-fairness balance, attaining the top average rank across all metrics. Furthermore, visualizations of routing weights illustrate the strategic behavior and effectiveness of the proposed fairness-aware routing mechanism. Our code can be found at \url{https://github.com/qin2mu4/FairSpec}.

The Attention Market: Interpreting Online Fair Re-ranking as Manifold Optimization under Walrasian Equilibrium [Full]

Chen Xu (Renmin University of China); Wei Chu (Renmin University of China); Wenyu Hu (Renmin University of China); Fengran Mo (Université de Montréal); Jun Xu (Renmin University of China); Maarten de Rijke (University of Amsterdam)

Fair re-ranking aims to promote long-tail items and enhance diversity within groups in information retrieval. While previous research on online fairness-aware re-ranking has shown promising outcomes, our comprehensive evaluation of online fair re-ranking methods over 20 settings reveals significant performance disparities among existing methods. To uncover the root causes of these inconsistencies, we reformulate fair re-ranking within an attentional market framework governed by a Walrasian Equilibrium, where the fairness is treated as a taxation cost. This market-based formulation is then coupled with manifold optimization, demonstrating that seeking this equilibrium is equivalent to performing gradient descent on a specific ranking manifold constructed by the market. Different re-ranking settings induce distinct manifold geometries, and these intrinsic geometric differences dictate the gradient landscapes and optimization trajectories. We propose ManifoldRank, an efficient online fair re-ranking algorithm. ManifoldRank adjusts gradients to align with the ranking manifold, considering various contextual settings. On the supply side, it incorporates a gradient adjustment based on different fairness requirements, accounting for associated costs. On the demand side, it empirically predicts an additional gradient adjustment term derived from the ranking scores. By integrating these two gradient adjustments, ManifoldRank effectively balances fairness and accuracy. Experimental results across multiple datasets confirm ManifoldRank's effectiveness.

Towards Inclusive Retrieval-Augmented Generation: Challenges and Opportunities for Cognitively Impaired Users [Perspective]

Claire Rogers (University of Strathclyde); Asmaa Z. A. M. Alqadri (University of Strathclyde); Fiona A. Beaton (NeuraSearch Laboratory); Yashar Moshfeghi (University of Strathclyde)

Current Information Retrieval (IR) systems, including conversational IR paradigms enhanced by Retrieval-Augmented Generation (RAG), assume that users can formulate coherent queries and reliably interpret retrieved information. In dementia care contexts, these assumptions fail: queries are unstable, feedback signals are unreliable, relevance fluctuates with cognitive state, and multi-user interaction is essential. Dementia impacts over 50 million people globally, yet current RAG research has yet to consider use cases where users have dementia or Mild Cognitive Impairment (MCI), and so RAG systems continue to be designed without considering people with dementia, MCI, or their caregivers as users, creating a fundamental mismatch between system capabilities and this population's needs. Our perspective outlines functional requirements, not optional features, for RAG systems serving cognitively diverse populations, including temporal user modelling that tracks cognitive trajectories over time, caregiver-in-the-loop retrieval that enables verification and oversight, consent-aware evidence access that respects fluctuating capacity, and cognitive-load-aware presentation that adapts complexity to comprehension level. These requirements would collectively transform RAG from a cognitively-stable user paradigm into a cognitively-adaptive one. Drawing on participatory workshops with 41 members of the public and dementia-care community, our perspective was shaped through empirical evidence from those with lived experience of dementia or MCI. We advocate for inclusive IR through our cognitively-adaptive RAG that supports cognitive independence in dementia care, and call on the IR community to advance accessibility and equity through inclusive retrieval system design.

Faithfulness and Robustness in RAG · *Courtyard 1+2* · *Tuesday 16:30–17:45*

Tokens to Types: Context Editing with Selective Entity Abstraction for Grounded Generation [Full]

Rounak Sharma (Adobe Research); Debabrata Mahapatra (Adobe Research); Shiv Kumar Saini (Adobe Research)

Large language models (LLMs) frequently prioritize parametric world knowledge over provided context -- a failure mode that is particularly catastrophic in enterprise or counterfactual settings where local facts contradict web-scale training data. While modern reasoning models improve general response quality, they fail to resolve these underlying prior knowledge biases even when generating a high volume of costly thinking tokens. We propose a context-editing framework that addresses this by performing selective abstraction over entities that appear in both the context and the question. Our approach replaces these overlapping entities with typed, indexed identifiers (e.g., Paris with City ■angle1\rangle), suppressing spurious lexical associations while preserving the relational structure required for grounded reasoning. This mechanism is decoupled into an offline preprocessing stage and a lightweight inference-time substitution, requiring no model retraining. Experiments across counterfactual and multi-hop knowledge-conflict benchmarks demonstrate consistent accuracy gains across many model families, open- and closed-sourced, especially for small to medium sized LLMs (0.5B to 18B). Notably, our framework yields up to a 25% improvement over state-of-the-art instruct models and 13% over reasoning models, establishing symbolic abstraction as a highly cost-efficient solution for ensuring context fidelity in LLMs.

Exploring Knowledge Conflicts for Faithful LLM Reasoning: Benchmark and Method [Full]

Tianzhe Zhao (Xi'an Jiaotong University); Jiaoyan Chen (The University of Manchester); Shuxiu Zhang (Hunan University); Haiping Zhu (Xi'an Jiaotong University); Qika Lin (National University of Singapore); Jun Liu (Xi'an Jiaotong University)

Large language models (LLMs) have achieved remarkable success across a wide range of applications especially when augmented by external knowledge through retrieval-augmented generation (RAG). Despite their widespread adoption, recent studies have shown that LLMs often struggle to perform faithful reasoning when conflicting knowledge is retrieved. However, existing work primarily focuses on conflicts between external knowledge and the parametric knowledge of LLMs, leaving conflicts across external knowledge largely unexplored. Meanwhile, modern RAG systems increasingly emphasize the integration of unstructured text and (semi-)structured data like knowledge graphs (KGs) to improve knowledge completeness and reasoning faithfulness. To address this gap, we introduce ConflictQA, a novel benchmark that systematically instantiates conflicts between textual evidence and KG evidence. Extensive evaluations across representative LLMs reveal that, facing such cross-source conflicts, LLMs often fail to identify reliable evidence for correct reasoning. Instead, LLMs become more sensitive to prompting choices and tend to rely exclusively on either KG or textual evidence, resulting in incorrect responses. Based on these findings, we further propose XoT, a two-stage explanation-based thinking framework tailored for reasoning over heterogeneous conflicting evidence, and verify its effectiveness with extensive experiments.

Beyond Semantic Relevance: Counterfactual Risk Minimization for Robust Retrieval-Augmented Generation [Full]

Peiyang Liu (Peking University); Qiang Yan (PX Securities); Ziqiang Cui (City University of Hong Kong); Di Liang (Tencent Technology); Xi Wang (Peking University); Wei Ye (Peking University)

Standard Retrieval-Augmented Generation (RAG) systems predominantly rely on semantic relevance as a proxy for utility. However, this assumption collapses in realistic decision-making scenarios where user queries are laden with cognitive biases, such as false premises or confirmation bias. In such cases, maximizing relevance paradoxically promotes the retrieval of sycophantic evidence that reinforces hallucinations, a critical failure we term the ``Relevance-Robustness Gap''. To bridge this gap, we propose CoRM-RAG (Counterfactual Risk Minimization for RAG), a framework that aligns retrieval with decision safety rather than mere similarity. Grounded in causal intervention, we introduce a Cognitive Perturbation Protocol to simulate user biases during training, which is then distilled into a lightweight Evidence Critic. This scoring module learns to identify documents that possess sufficient evidential strength to steer the model toward correctness despite adversarial query perturbations. Extensive experiments on decision-making benchmarks demonstrate that CoRM-RAG significantly outperforms strong dense retrievers and LLM-based rerankers in adversarial settings, while enabling effective risk-aware abstention through reliable robustness scoring. Our code is available at https://github.com/PeiYangLiu/CoRM-RAG.git.

LLM-Oriented Information Retrieval: A Denoising-First Perspective [Perspective]

Lu Dai (Hong Kong University of Science and Technology); Liang Sun (Hong Kong University of Science and Technology (Guangzhou)); Fanpu Cao (Hong Kong University of Science and Technology (Guangzhou)); Ziyang Rao (Hong Kong University of Science and Technology (Guangzhou)); Cehao Yang (Hong Kong University of Science and Technology (Guangzhou)); Hao Liu (Hong Kong University of Science and Technology (Guangzhou)); Hui Xiong (Hong Kong University of Science and Technology)

Modern information retrieval (IR) is no longer consumed primarily by humans but increasingly by large language models (LLMs) via retrieval-augmented generation (RAG) and agentic search. Unlike human users, LLMs are constrained by limited attention budgets and are uniquely vulnerable to noise; misleading or irrelevant information is no longer just a nuisance, but a direct cause of hallucinations and reasoning failures. In this perspective paper, we argue that denoising-maximizing usable evidence density and verifiability within a context window-is becoming the primary bottleneck across the full information access pipeline. We conceptualize this paradigm shift through a four-stage framework of IR challenges: from inaccessible to undiscoverable, to misaligned, and finally to unverifiable. Furthermore, we provide a pipeline-organized taxonomy of signal-to-noise optimization techniques, spanning indexing, retrieval, context engineering, verification,

and agentic workflow. We also present research works on information denoising in domains that rely heavily on retrieval such as lifelong assistant, coding agent, deep research, and multimodal understanding.

Lost in the Evidence? Reproducing Document Position and Context Size Effects in RAG [Reproducibility]

Jorge Gabín (Universidade da Coruña); Anxo Pérez (Universidade da Coruña); Javier Parapar (Universidade da Coruña)

Retrieval-Augmented Generation (RAG) systems rely on retrieved documents being concatenated into a model's input context, making both document ordering and context size critical yet controversial design choices. Prior work reports position-based effects such as lost in the middle and related long-context phenomena. However, empirical findings remain inconsistent and hard to reproduce across models, datasets, and evaluation protocols. In this paper, we present a systematic reproducibility study that revisits these claims and examines how they evolve with contemporary LLMs under a controlled evaluation framework. We first show that topic sampling is a major source of variance: small topic sets can mask or exaggerate ordering effects. Based on repeated subset sampling across multiple topic budgets, we provide a practical calibration procedure that identifies topic counts yielding stable trends at feasible cost. Using these fixed topic sets, we then reproduce and extend results on position sensitivity, re-evaluating lost in the middle and positional biases in modern LLMs. Then, we also study a more realistic RAG scenario in which relevance is mediated by a retriever rather than oracle access to ground-truth documents. In this setting, we re-examine a recent industry study and identify discrepancies to evaluation choices such as limited topic coverage and reliance on LLM-based judges. Finally, we conduct an analysis of how retrieval order and context size affect downstream LLM performance under imperfect retrieval. Our results demonstrate that both factors interact strongly with retrieval quality and model choice, and that conclusions drawn from idealised setups do not always transfer to real-world RAG pipelines. We release all code and configurations to support reproducibility and future work on robust RAG evaluation.

Context-Aware Recommendation · *Goldfields* ·

Wednesday 10:30–12:00

Inferring Targets from Calibrated Hesitations via Mutual Information Maximization in Multi-Behavior Recommendation [Full]

Cheng Li (Anhui Normal University); Yong Xu (Anhui Normal University); Suhua Tang (The University of Electro-Communications); Xin He (Anhui Normal University); Jianfeng Sun (Anhui Normal University); Jinde Cao (Southeast University)

Multi-behavior recommendation enriches user preference modeling by incorporating diverse auxiliary interactions. However, most existing methods simply treat interactions without target behaviors as absolute negative feedback. This strategy ignores an important intermediate state known as user hesitation, where users exhibit strong intent but fail to complete the final conversion due to various reasons. Consequently, models cannot distinguish true disinterest from intended but hesitant behavior, which introduces substantial noise into preference modeling. To address this issue, we propose a novel framework named Calibrated Hesitation Analysis for Multi-Behavior Recommendation via Mutual Information Maximization (CHARM). Specifically, we aggregate auxiliary behaviors that lead to successful conversions into latent intent representations and train an inference network by maximizing the mutual information between these intents and observed target behaviors. We then apply this network to auxiliary behaviors without conversion, under the assumption that the conversion had occurred, in order to infer latent conversion probabilities and identify high-intent hesitation candidates. Furthermore, to distinguish genuine hesitation from interaction termination caused by competing item choices, we design a competitor substitution penalty strategy to refine hesitation confidence scores. Finally, the calibrated hesitation set is incorporated into the recommendation process to improve ranking quality. Extensive experiments on three real-world datasets demonstrate that CHARM consistently outperforms existing state-of-the-art methods. The source code is available at https://github.com/city59/CHARM.

Incremental Multi-Behavior Recommendation [Full]

Jiahao Gong (Shenzhen University); Weike Pan (Shenzhen University)

Multi-behavior recommendation, unlike traditional single-behavior recommendation focusing solely on the target behavior (e.g., purchase), exploits multiple types of user-item interactions (e.g., view, favorite, add-to-cart, purchase) to mitigate data sparsity and learn more comprehensive user preferences. However, real-world data is inherently transmitted in a streaming manner, and most existing multi-behavior recommendation models are trained on static offline datasets, which cannot efficiently handle continuously arriving data streams. Full retraining incurs high computational cost and long training time. In contrast, fine-tuning is efficient but prone to "catastrophic forgetting", resulting in performance degradation. Incremental recommendation techniques offer a potential solution, but the existing methods are mostly designed for cases with one single behavior, ignoring valuable information contained in heterogeneous behaviors. To fill this gap, we introduce a new and important problem, i.e., incremental multi-behavior recommendation (IMBR), requiring a model to handle streaming data while utilizing valuable heterogeneous behaviors. To this end, we propose a novel solution called Behavior-aware Incremental Graph Convolution Network with Multi-Task Learning (BIGCN-MTL) for IMBR. Specifically, our BIGCN-MTL contains two key modules: behavior-aware incremental graph convolution network (BIGCN), which employs incremental cascading graph convolution to preserve both historical neighborhood information and behavior cascading dependency, and historical interest distillation (HID), which maintains users' preference stability and prevents behavioral semantic forgetting via non-interest consistency constraint and preference contrastive distillation. Extensive experiments on four real-world datasets demonstrate that our BIGCN-MTL achieves superior recommendation performance and improves training efficiency compared to full retraining. The source code of our BIGCN-MTL is available at https://github.com/codingPeter1/BIGCN-MTL.

Distribution-aware Re-representations for Multi-Scenario Recommendations [Full]

Qi Sun (The University of Melbourne); Yulin Xu (University of Southern California); Zelin Wang (Independent); Xiaoyu Kang (Institute of Information Engineering, Chinese Academy of Sciences); Keyan Jin (Macao Polytechnic University); Jiechao Gao (Stanford University)

Modern applications provided personalized recommendations across diverse scenarios, including the homepage, local pages, and live streams on platforms like TikTok. These scenarios exhibit varying user behavior patterns, resulting in heterogeneous yet interrelated distributions. Existing Multi-Scenario Recommendation (MSR) methods usually use parameter-sharing networks for shared features and scenario-specific networks for unique features. However, these methods fail to handle different distribution across scenarios, resulting in representation entanglement and localization, which hinder effective knowledge transfer and compromise performance. In this paper, we propose a Distribution-aware Re-representations (DAR) method for MSR. Its core idea is to construct distribution-aware prototype spaces and learn disentangled re-representations around global prototypes. Specifically, DAR employs a Multi-gate Mixture of Experts (MMoE) to obtain scenario-shared representations, and uses independent networks to learn scenario-specific representations. These representations are then projected into scenario-shared and scenario-specific prototype spaces, producing scenario-shared re-representations (capturing global information) and scenario-specific re-representations (focusing on distributional differences). During this process, DAR utilizes Unbalanced Optimal Transport (UOT) to compute the transport relationships between representations and global prototypes, taking these as pseudo-labels for re-representation learning. Moreover, to prevent prototype entanglement, a matrix orthogonalization constraint ensures independence among global prototypes. The effectiveness of DAR is demonstrated through extensive offline experiments conducted on four datasets, as well as online A/B tests on a video platform.

Mining Informative Interests via Latent Cross Reasoning for Search Enhanced Recommendation [Full]

Teng Shi (Renmin University of China); Weicong Qin (Renmin University of China); Weijie Yu (University of International Business and Economics); Xiao Zhang (Renmin University of China); Ming He (AI Lab at Lenovo Research); Jianping Fan (AI Lab at Lenovo Research); Jun Xu (Renmin University of China)

Search and recommendation (S&R) are fundamental components of modern commercial platforms, enabling users to access and explore information efficiently. User behaviors in these scenarios reflect different aspects of user intent, providing an opportunity for joint modeling of S&R. However, effectively leveraging search logs to enhance recommendation remains a challenging task. Existing methods often encode S&R histories either jointly or separately; however, they tend to regard all search signals as equally informative, thereby neglecting that many search behaviors can be irrelevant or even detrimental to recommendation performance. In practice, however, search histories frequently contain noisy or outdated behaviors that may introduce spurious correlations and degrade recommendation performance. Motivated by the human decision-making process, where one first identifies recommendation intent and then selectively reasons about relevant search signals, we propose LCR-SER, a latent cross reasoning method for search-enhanced recommendation. LCR-SER first encodes the user's S&R history into a unified latent representation that captures users' global interests. It then performs iterative reasoning in the latent space to dynamically identify informative search signals that are most relevant to the recommendation. To further guide this reasoning process, we introduce contrastive learning to align the reasoning states with the target items. In addition, we employ reinforcement learning to directly optimize ranking-oriented metrics, enabling LCR-SER to refine its reasoning strategy toward improved recommendation performance. Experiments on public datasets demonstrate that LCR-SER consistently outperforms strong baselines, validating the effectiveness of latent reasoning in enhancing search-aware recommendation.

Price-Aware Recommender Systems: A Cross-Paradigm Reproducibility Study [Reproducibility]

Fernando Medina-Quispe (Universidad Arturo Prat); David Contreras Aguilar (La Salle-Universitat Ramon Llull); Ludovico Boratto (University of Cagliari); Maria Salamó (Universitat de Barcelona)

Price is a key determinant of user purchase behavior in e-commerce. As traditional recommendation techniques typically fail to account for price information, Price-Aware Recommender Systems have emerged in response, aiming to explicitly incorporate price as a core feature to enhance the overall quality and relevance of recommendations. Recent advances include session-based approaches such as CoHHN (Heterogeneous Hypergraphs), PASBR (Graph Neural Network), and the

collaborative-filtering approach PUP (Graph Convolutional Networks). All reporting improved performance on their respective evaluation datasets. However, these methods have never been compared under controlled experimental conditions, limiting our understanding of their relative performance. This work addresses this gap through a reproducibility study over CoHHN, PASBR, and PUP. We successfully conducted the first cross-evaluation across six public datasets using standardized metrics (HR@20, MRR@20, NDCG@20). Our findings indicate that CoHHN provides the strongest and most consistent performance across evaluation metrics and datasets, while PASBR shows domain-limited competitiveness and PUP exhibits the weakest generalization, highlighting CoHHN's heterogeneous hypergraph architecture as the most robust framework for Price-Aware Recommendation. Source code available at https://anonymous.4open.science/r/reproducibility-sigir2026-E2CE.

Multi-Perspective Driven Expected Location Preferences for Next POI Recommendations [Full]

Pengxiang Lan (Northeastern University); Enneng Yang (Northeastern University); Yuliang Liang (Northeastern University); Jianzhe Zhao (Northeastern University); Guibing Guo (Northeastern University); Hai Zhao (Northeastern University)

Next point-of-interest (POI) recommendation aims to predict users' next destinations from their mobility trajectories. Nevertheless, accurately capturing user preferences from check-in data remains challenging, as user behaviors are shaped by complex contextual factors and varying visit intentions. Despite substantial progress, existing methods still face two critical limitations. (i) They typically assume that all check-in behaviors faithfully reflect users' intrinsic preferences, overlooking behaviors driven by external contextual factors (e.g., work requirements) that misalign with users' authentic preference space, limiting the ability to understand dynamic user behavior patterns. (ii) Some of the POIs that users check in at may just be the closest available options to their expected visit intentions, rather than places they genuinely prefer. Existing methods neglect users' expected intentions, leading to inaccurate identification of users' authentic preferences. In this paper, we propose a novel next POI recommendation method (MPDC) that leverages Multi-Perspective modeling and Diffusion-based Contrastive learning. From the behavior-aware perspective, a behavior-aware graph prompt module mitigates external contextual influences via bidirectional sequence modeling and a preference filter, while capturing visit preferences. From the spatial-aware perspective, a spatial-aware hyperbolic graph module leverages hyperbolic space to model higher-order, nonlinear geographic patterns, effectively capturing the influence of spatial factors. In addition, a diffusion-based preference contrast module leverages the generative ability of diffusion models to learn latent expected location preferences that reflect user intent under different contextual conditions. Comprehensive experiments across five real-world datasets demonstrate that MPDC significantly outperforms state-of-the-art algorithms. Our code is available at https://github.com/LanPangxiang/LaMDA2026.

CTR Prediction · *Room 110 · Wednesday 10:30–12:00*

Hierarchical Denoising Entire Space Multi-Task Model for Post-Click Conversion Rate Prediction with Noisy Labels [Full]

Yan Lyu (Peking University); Haoxuan Li (Peking University); Tianyu Xia (Peking University); Xiang Li (Peking University); Xiangnan Feng (Fudan University); Chunyuan Zheng (Peking University); Xiao-Hua Zhou (Peking University)

The post-click Conversion Rate (CVR) prediction plays a pivotal role in industrial recommender systems, directly serving as a decisive factor for ranking strategies and revenue optimization. While Entire Space Multi-Task Models (e.g., ESMM) have become the dominant framework by effectively addressing sample selection bias and data sparsity, they rely heavily on the assumption that the observed click and conversion labels perfectly reflect user preferences. In real-world scenarios, however, observed data is inevitably contaminated by label noise. Crucially, within the entire space modeling framework, this noise exhibits a complex hierarchical structure, where the superimposition of noise from antecedent click labels and subsequent conversion labels results in conventional single-task denoising methods being ineffective. To bridge this gap, we propose HiDe-ESMM (Hierarchical Denoising Entire Space Multi-Task Model), a unified framework designed to systematically mitigate hierarchical label noise. First, adopting the loss correction paradigm, we derive a statistically unbiased CTCVR loss via backward correction and identify the requisite noise rates based on the weak separability assumption. Second, to alleviate the numerical instability inherent in the unbiased loss approach, we introduce a robust forward correction strategy as an efficient alternative and theoretically prove the consistency of its optimal solution with the optimal solution on clean data. Extensive experiments on one semi-synthetic dataset and two real-world industrial datasets demonstrate that HiDe-ESMM significantly outperforms state-of-the-art baselines, validating its robustness in handling hierarchical noisy labels.

Hypergraph Diffusion-Based Sequential Ensemble for CTR Prediction [Full]

Zeheng Zhong (Peking University); Hongzhi Liu (Peking University); Gong Chen (Tencent Inc); Boyuan Ren (Peking University); Guomin Qin (Tencent Inc); Zhonghai Wu (Peking University)

Click-through Rate (CTR) prediction is a crucial task in online advertising and recommender systems. Theoretically, proper ensemble of multiple different CTR prediction models can improve the prediction effectiveness. Unfortunately, most of the existing ensemble learning methods for CTR prediction are designed for cross-sectional data and neglect user historical behavior sequence which is important for predicting user's future click behavior. To address the above issues, we propose a Hypergraph Diffusion-based Sequential Ensemble framework for CTR prediction (HDSE). Specifically, considering the inherent capability of diffusion models in space exploration, we design a generalized conditional diffusion model, which adaptively identifies critical information from diverse sequential models, to capture user's dynamic interest evolution across diverse contexts. To avoid corrupting the item dependencies caused by isotropic Gaussian noise used in traditional diffusion models, we construct a behavior hypergraph and design an anisotropic hypergraph-based smoothing operator to inject structure-aware noise for better exploring the user's interest space. For large-scale application scenarios, we propose a computationally efficient approximation method for estimating hypergraph propagation matrix in the smoothing operator. Extensive experiments on three real-world datasets demonstrate the effectiveness of the proposed model.

FedMM: Federated Collaborative Signal Quantization for Multi-Market CTR Prediction [Full]

Jun Zhang (Shenzhen University); Dugang Liu (Shenzhen University); Xing Tang (Shenzhen Technology University); Xiuqiang He (Shenzhen Technology University); Zhong Ming (Shenzhen University)

Online platforms such as Amazon and Netflix serve users across multiple countries and regions, underscoring the importance of multi-market recommendation (MMR). Most MMR methods adopt a pre-training and fine-tuning paradigm, in which a unified model is first trained on centralized, global data and subsequently adapted to specific markets. However, this approach ignores the privacy of market data. While traditional federated learning preserves privacy, it typically aims to obtain a global model by aggregating model parameters and does not account for significant market heterogeneity. Additionally, because ID spaces are disjoint across markets, embedding-based aggregation strategies become ineffective. To overcome these challenges, we propose a federated collaborative signal quantization (FedMM) method for multi-market click-through rate (CTR) prediction. Our core idea leverages a discrete codebook mechanism to achieve privacy-preserving transmission and align disjoint ID spaces. We further employ a hierarchical codebook structure to capture cross-market shared patterns and market-specific characteristics. Specifically, we deploy a residual quantized variational autoencoder (RQ-VAE) with a dual-layer codebook mechanism for each market to quantize collaborative embeddings. The first layer utilizes a global federated codebook, updated via aggregation to capture universally shared collaborative patterns, while the second layer maintains a local codebook to learn market-specific semantics. Finally, the learned discrete codes, which integrate both general and specific collaborative signals, are incorporated into downstream CTR models to enhance prediction accuracy across all markets. Extensive experiments on benchmark datasets demonstrate that FedMM significantly improves recommendation performance with privacy guarantees.

Exploring Test-time Scaling via Prediction Merging on Large-Scale Recommendation [Full]

Fuyuan Lyu (McGill University); Zhentai Chen (Shenzhen Technology University); Jingyan Jiang (Shenzhen Technology University); Lingjie Li (Shenzhen Technology University); Xing Tang (Shenzhen Technology University); Xiuqiang He (Shenzhen Technology University); Xue Liu (McGill University)

Inspired by the success of language models (LM), scaling up deep learning recommendation systems (DLRS) has become a recent trend in the community.All previous methods tend to scale up the model parameters during training time.However, how to efficiently utilize and scale up computational resources during test time remains underexplored, which can prove to be a scaling-efficient approach and bring orthogonal improvements in LM domains.The key point in applying test-time scaling to DLRS lies in effectively generating diverse yet meaningful outputs for the same instance.We propose two ways: One is to explore the heterogeneity of different model architectures.The other is to utilize the randomness of model initialization under a homogeneous architecture.The evaluation is conducted across eight models, including both classic and SOTA models, on three benchmarks. Sufficient evidence proves the effectiveness of both solutions.We further prove that under the same inference budget, test-time scaling can outperform parameter scaling.Our test-time scaling can also be seamlessly accelerated with the increase in parallel servers when deployed online, without affecting the inference time on the user side.Code is available here\footnote{https://github.com/aTitye/TTS4CTR}.

HyFormer: Revisiting the Roles of Sequence Modeling and Feature Interaction in CTR Prediction [Full]

Yunwen Huang (ByteDance AML); Shiyong Hong (ByteDance Search); Xijun Xiao (ByteDance AML); Jinqiu Jin (ByteDance Search); Xuanyuan Luo (ByteDance AML); Zhe Wang (ByteDance Search); Zheng Chai (ByteDance AML); Shikang Wu (ByteDance Search); Yuchao Zheng (ByteDance AML); Jingjian Lin (ByteDance Search)

Industrial large-scale recommendation models (LRMs) face the challenge of jointly modeling long-range user behavior sequences and heterogeneous non-sequential features under strict efficiency constraints. However, most existing architectures employ a decoupled pipeline: long sequences are first compressed with a query-token based sequence compressor like LONGER, followed by fusion with dense features through token-mixing modules like RankMixer, which thereby limits both the representation capacity and the interaction flexibility. This paper presents HyFormer, a unified hybrid transformer architecture that tightly integrates long-sequence modeling and

feature interaction into a single backbone. From the perspective of sequence modeling, we revisit and redesign query tokens in LRMs, and frame the LRM modeling task as an alternating optimization process that integrates two core components: Query Decoding which expands non-sequential features into Global Tokens and performs long sequence decoding over layer-wise key-value representations of long behavioral sequences; and Query Boosting which enhances cross-query and cross-sequence heterogeneous interactions via efficient token mixing. The two complementary mechanisms are performed iteratively to refine semantic representations across layers. Extensive experiments on billion-scale industrial datasets demonstrate that HyFormer consistently outperforms strong LONGER and RankMixer baselines under comparable parameter and FLOPs budgets, while exhibiting superior scaling behavior with increasing parameters and FLOPs. Large-scale online A/B tests in high-traffic production systems further validate its effectiveness, showing significant gains over deployed state-of-the-art models. These results highlight the practicality and scalability of HyFormer as a unified modeling framework for industrial LRMs.

Beyond Dense Connectivity: Explicit Sparsity for Scalable Recommendation [Full]

Yantao Yu (Alibaba International Digital Commercial Group); Sen Qiao (Alibaba International Digital Commercial Group); Lei Shen (Alibaba International Digital Commercial Group); Bing Wang (Alibaba International Digital Commercial Group); Xiaoyi Zeng (Alibaba International Digital Commercial Group)

Recent progress in scaling large models has motivated recommender systems to increase model depth and capacity to better leverage massive behavioral data. However, recommendation inputs are high-dimensional and extremely sparse, and simply scaling dense backbones (e.g., deep MLPs) often yields diminishing returns or even performance degradation. Our analysis of industrial CTR models reveals a phenomenon of implicit connection sparsity: most learned connection weights tend towards zero, while only a small fraction remain prominent. This indicates a structural mismatch between dense connectivity and sparse recommendation data; by compelling the model to process vast low-utility connections instead of valid signals, the dense architecture itself becomes the primary bottleneck to effective pattern modeling. We propose SSR (Explicit Sparsity for Scalable Recommendation), a framework that incorporates sparsity explicitly into the architecture. SSR employs a multi-view "filter-then-fuse" mechanism, decomposing inputs into parallel views for dimension-level sparse filtering followed by dense fusion. Specifically, we realize the sparsity via two strategies: a Static Random Filter that achieves efficient structural sparsity via fixed dimension subsets, and Iterative Competitive Sparse (ICS), a differentiable dynamic mechanism that employs bio-inspired competition to adaptively retain high-response dimensions. Experiments on three public datasets and a billion-scale industrial dataset from AliExpress (a global e-commerce platform) show that SSR outperforms state-of-the-art baselines under similar budgets. Crucially, SSR exhibits superior scalability, delivering continuous performance gains where dense models saturate. The code is available at https://github.com/Atticus666/SSRNet.

Item Tokenization · *Eureka 1 · Wednesday 10:30–12:00*

Learning Decomposed Contextual Token Representations from Pretrained and Collaborative Signals for Generative Recommendation [Full]

Yifan Liu (University of Illinois at Urbana-Champaign); Yaokun Liu (University of Illinois at Urbana-Champaign); Zelin Li (University of Illinois at Urbana-Champaign); Zhenrui Yue (University of Illinois at Urbana-Champaign); Ruichen Yao (University of Illinois at Urbana-Champaign); Yang Zhang (Miami University); Dong Wang (University of Illinois at Urbana-Champaign)

Recent advances in generative recommenders adopt a two-stage paradigm: items are first tokenized into semantic IDs using a pretrained tokenizer, and then large language models (LLMs) are trained to generate the next item via sequence-to-sequence modeling. However, these two stages are optimized for different objectives: semantic reconstruction during tokenizer pretraining versus user interaction modeling during recommender training. This objective misalignment leads to two key limitations: (i) suboptimal static tokenization, where fixed token assignments fail to reflect diverse usage contexts; and (ii) discarded pretrained semantics, where pretrained knowledge—typically from language model embeddings—is overwritten during recommender training on user interactions. To address these limitations, we propose to learn DEcomposed COntextual Token Representations (DECOR), a unified framework that preserves pretrained semantics while enhancing the adaptability of token embeddings. DECOR introduces contextualized token composition to refine token embeddings based on user interaction context, and decomposed embedding fusion that integrates pretrained codebook embeddings with newly learned collaborative embeddings. Experiments on three real-world datasets demonstrate that DECOR consistently outperforms state-of-the-art baselines in recommendation performance. Our code is available at https://github.com/yliuaa/DECOR.git.

Bi-Level Optimization for Generative Recommendation: Bridging Tokenization and Generation [Full]

Yimeng Bai (University of Science and Technology of China); Chang Liu (Beihang University); Yang Zhang (National University of Singapore); Dingxian Wang (Upwork); Frank Yang (Upwork); Andrew Rabinovich (Upwork); Wenge Rong (Beihang University); Fuli Feng (University of Science and Technology of China)

Generative recommendation is emerging as a transformative paradigm by directly generating recommended items, rather than relying on matching. Building such a system typically involves two key components: (1) optimizing the tokenizer to derive suitable item identifiers, and (2) training the recommender based on those identifiers. Existing approaches often treat these components separately—either sequentially or in alternation—overlooking their interdependence. This separation can lead to misalignment: the tokenizer is trained without direct guidance from the recommendation objective, potentially yielding suboptimal identifiers that degrade recommendation performance. To address this, we propose BLOGER, a Bi-Level Optimization for GEnerative Recommendation framework, which explicitly models the interdependence between the tokenizer and the recommender in a unified optimization process. The lower level trains the recommender using tokenized sequences, while the upper level optimizes the tokenizer based on both the tokenization loss and recommendation loss. We adopt a meta-learning approach to solve this bi-level optimization efficiently, and introduce gradient surgery to mitigate gradient conflicts in the upper-level updates, thereby ensuring that item identifiers are both informative and recommendation-aligned. Extensive experiments on multiple real-world datasets demonstrate that BLOGER consistently outperforms state-of-the-art generative recommendation methods while maintaining practical efficiency with no significant additional computational overhead, effectively bridging the gap between item tokenization and autoregressive generation. We release our code at https://github.com/Ten-Mao/BLOGER.

Differentiable Semantic ID for Generative Recommendation [Full]

Junchen Fu (University of Glasgow); Xuri Ge (Shandong University); Alexandros Karatzoglou (Amazon); Ioannis Arapakis (Telefónica Innovación Digital); Suzan Verberne (Leiden University); Joemon M. Jose (University of Glasgow); Zhaochun Ren (Leiden University)

Generative recommendation provides a novel paradigm in which each item is represented by a discrete semantic ID (SID) learned from rich content. Most methods treat SIDs as predefined and train recommenders under static indexing. In practice, SIDs are optimized only for content reconstruction rather than recommendation accuracy. This leads to an objective mismatch : the system optimizes an indexing loss to learn the SID, and a recommendation loss for interaction prediction, but because the tokenizer is trained independently, the recommendation loss cannot update it. A natural approach is to make semantic indexing differentiable so recommendation gradients can directly influence SID learning, but this often causes codebook collapse with only a few codes used. We attribute this to early deterministic assignments that limit codebook exploration, leading to imbalance and unstable optimization. In this paper, we therefore propose DIGER (Differentiable Semantic ID for GEnerative Recommendation). DIGER is a first step towards an effective differentiable semantic ID for generative recommendation. The Gumbel noise explicitly encourages early-stage exploration over codes, mitigating collapse and improving code utilization. To better balance exploration and convergence, we introduce two uncertainty decay strategies that reduce the Gumbel noise, enabling a gradual shift from early-stage exploration to the exploitation of learned SIDs. Extensive experiments across multiple public datasets demonstrate consistent improvements from differentiable semantic ID. These results confirm the effectiveness of aligning indexing and recommendation objectives through differentiable SIDs. This identifies differentiable SID as a promising area of study. Our code is released under https://github.com/junchen-fu/DIGER.

CARD: Non-Uniform Quantization of Visual Semantic Unit for Generative Recommendation [Full]

Yibiao Wei (University of Electronic Science and Technology of China); Jie Zou (University of Electronic Science and Technology of China); Pengfei Zhang (University of Electronic Science and Technology of China); Xiao Ao (University of Electronic Science and Technology of China); Weikang Guo (Southwest University of Finance and Economics); Zeyu Ma (University of Electronic Science and Technology of China); Yang Yang (University of Electronic Science and Technology of China)

Generative recommendation frameworks typically represent items as discrete Semantic IDs (SIDs). While existing studies have sought to enhance SID construction by incorporating multimodal content, collaborative signals, or more advanced quantization techniques, learning high-quality SIDs still faces two key challenges: (1) The two-stage generative recommendation paradigm (SID construction and autoregressive generation) provides insufficient supervision for heterogeneous fusion, which hinders learning high-quality SIDs, and (2) non-uniform embeddings lead to codeword imbalance and generation bias. To address these challenges, we propose a novel generative recommendation framework, called CARD. CARD introduces a visual semantic unit that unifies textual, visual, and collaborative signals into a structured visual representation prior to encoding, enabling holistic semantic modeling and effectively alleviating the semantic gap, thereby reducing the reliance on supervision signals during SID learning. Furthermore, to deal with the highly non-uniform distribution of item semantic embeddings in recommendation scenarios, we develop a non-uniform quantization framework (NU-RQ-VAE), which incorporates a learnable and invertible non-uniform transformation into the quantization process to map skewed semantic distributions into a more balanced latent space, thereby significantly improving codebook utilization and quantization accuracy. Experiments on multiple datasets show that CARD consistently outperforms baseline methods under various settings; meanwhile, the proposed non-uniform transformation module is plug-and-play and remains robust across different quantization schemes. Code is available at

https://github.com/HAI-UESTC/CARD.

Universal Item Tokenization for Transferable Generative Recommendation [Full]

Bowen Zheng (Renmin University of China); Hongyu Lu (Wechat, Tencent); Yu Chen (Wechat, Tencent); Xin Zhao (Renmin University of China); Ji-Rong Wen (Renmin University of China)

Recently, generative recommendation has emerged as a promising paradigm, attracting significant research attention. The basic framework involves an item tokenizer, which represents each item as a sequence of codes serving as its identifier, and a generative recommender that predicts the next item by autoregressively generating the target item identifier. However, in existing methods, both the tokenizer and the recommender are typically domain-specific, limiting their ability for effective transfer or adaptation to new domains. To this end, we propose UTGRec, a Universal item Tokenization approach for transferable Generative Recommendation. Specifically, we design a universal item tokenizer for encoding rich item semantics by adapting a multimodal large language model (MLLM). By devising tree-structured codebooks, we discretize content representations into corresponding codes for item tokenization. To effectively learn the universal item tokenizer on multiple domains, we introduce two key techniques in our approach. For raw content reconstruction, we employ dual lightweight decoders to reconstruct item text and images from discrete representations to capture general knowledge embedded in the content. For collaborative knowledge integration, we assume that co-occurring items are similar and integrate collaborative signals through co-occurrence alignment and reconstruction. Finally, we present a joint learning framework to pre-train and adapt the transferable generative recommender across multiple domains. Extensive experiments on four public datasets demonstrate the superiority of UTGRec compared to both traditional and generative recommendation baselines.

Drift-Aware Incremental Token Adaptation with Collaborative Semantics for Generative Recommendation [Full]

Yuebo Feng (Fudan University); Jiahao Liu (Fudan University); Mingzhe Han (Fudan University); Dongsheng Li (Microsoft Research Asia); Hansu Gu (Independent); Peng Zhang (Fudan University); Tun Lu (Fudan University); Ning Gu (Fudan University)

Generative recommendation commonly adopts a two-stage pipeline in which a learnable tokenizer maps items to discrete token sequences (i.e., identifiers) and an autoregressive generative recommender model (GRM) performs prediction based on these identifiers. Recent tokenizers further incorporate collaborative signals so that items with similar user-behavior patterns receive similar codes, substantially improving recommendation quality. However, real-world environments evolve continuously: new items cause identifier collision and shifts, while new interactions induce collaborative drift in existing items (e.g., changing co-occurrence patterns and popularity). Fully retraining both tokenizer and GRM is often prohibitively expensive, yet naively fine-tuning the tokenizer can alter token sequences for the majority of existing items, undermining the GRM's learned token–embedding alignment. To balance plasticity and stability for collaborative tokenizers, we propose DACT, a Drift-Aware Continual Tokenization framework with two stages: (i) tokenizer fine-tuning, augmented with a jointly trained Collaborative Drift Identification Module (CDIM) that outputs item-level drift confidence and enables differentiated optimization for drifting and stationary items; and (ii) hierarchical code reassignment using a relaxed-to-strict strategy to update token sequences while limiting unnecessary changes. Experiments on three real-world datasets with two representative GRMs show that DACT consistently achieves better performance than baselines, demonstrating effective adaptation to collaborative evolution with reduced disruption to prior knowledge. Our implementation is publicly available at https://github.com/HomesAmaranta/DACT for reproducibility.

Misinformation and Bias in Search · *Eureka 2* ·

Wednesday 10:30–12:00

Retrieval-Augmented Multimodal Model for Fake News Detection [Full]

Yiheng Li (University of International Business and Economics); Weihai Lu (Peking University); Hanyi Yu (University of Southern California); Yue Wang (Upstart Holdings, Inc.)

In recent years, multi-modal multi-domain fake news detection has garnered increasing attention. Nevertheless, this direction presents two significant challenges: (1) Failure to Capture Cross-Instance Narrative Consistency: existing models usually evaluate each news in isolation, fail to capture cross-instance narrative consistency, and thus struggle to address the spread of cluster-based fake news driven by social media; (2) Lack of Domain-Specific Knowledge for Reasoning: conventional models, which rely solely on knowledge encoded in their parameters during training, struggle to generalize to new or data-scarce domains (e.g., emerging events or niche topics). To tackle these challenges, we introduce Retrieval-Augmented Multimodal Model for Fake News Detection (RAMM). First, RAMM employs a Multimodal Large Language Model (MLLM) as its backbone to capture cross-modal semantic information from news samples. Second, RAMM incorporates an Abstract Narrative Alignment Module. This component adaptively extracts abstract narrative consistency from diverse instances across distinct domains, aggregates relevant knowledge, and thereby enables the modeling of high-level narrative information. Finally, RAMM introduces a Semantic Representation Alignment Module, which aligns the model's decision-making paradigm with that of humans—specifically, it shifts the model's reasoning process from direct inference on multimodal features to an instance-based analogical reasoning process. Extensive experimental results on three public datasets validate the efficacy of our proposed approach. Our code is available at the following link: \hrefhttps://github.com/li-yiheng/RAMM https://github.com/li-yiheng/RAMM

Mitigating Adversarial Attacks by Transferring LLM-generated Narrative Reasoning for Robust Fake News Detection [Full]

Mengyang Chen (Institute of Information Engineering, Chinese Academy of Sciences); Lingwei Wei (Institute of Information Engineering, Chinese Academy of Sciences); Wei Zhou (Institute of Information Engineering, Chinese Academy of Sciences); Songlin Hu (Institute of Information Engineering, Chinese Academy of Sciences)

Propagation-based fake news detectors primarily extract structural patterns from news propagation trees via graph neural networks (GNNs), which are crucial for trustworthy information access on social platforms. However, these systems remain vulnerable to adversarial message injection, increasingly enabled by large language models (LLMs). Such attacks pollute both semantic and structural signals, causing GNN-based aggregators to fuse logically conflicting content and yield unreliable representations. To address this, we propose LLM-TKT, a novel framework that distills LLM-based narrative reasoning into lightweight GNNs for robust fake news detection. The framework operates in two stages. First, we construct an offline LLM-driven narrative hub to synthesize global propagation narratives and diagnose local node-level coherence. Second, we design a dual-level narrative alignment to learn the semantic invariance of propagation with the guidance of propagation narratives. It filters unreliable neighbor nodes via local consistency and optimizes graph representations via global anchoring. Experiments on three real-world datasets demonstrate that LLM-TKT significantly outperforms existing methods, particularly in defending against sophisticated LLM-driven injection attacks without incurring runtime LLM inference costs.

ExDR: Explanation-driven Dynamic Retrieval Enhancement for Multimodal Fake News Detection [Full]

Guoxuan Ding (Institute of Information Engineering, Chinese Academy of Sciences); Yuqing Li (Institute of Information Engineering, Chinese Academy of Sciences); Ziyan Zhou (Institute of Information Engineering, Chinese Academy of Sciences); Zheng Lin (Institute of Information Engineering, Chinese Academy of Sciences); Daren Zha (Institute of Information Engineering, Chinese Academy of Sciences); Jiangnan Li (WeChat AI, Tencent Inc.)

The rapid spread of multimodal fake news poses a serious societal threat, as its evolving nature and reliance on timely factual details challenge existing detection methods. Dynamic Retrieval-Augmented Generation provides a promising solution by triggering keyword-based retrieval and incorporating external knowledge, thus enabling both efficient and accurate evidence selection. However, it still faces challenges in addressing issues such as redundant retrieval, coarse similarity, and irrelevant evidence when applied to deceptive content. In this paper, we propose ExDR—an Explanation-driven Dynamic Retrieval-Augmented Generation framework for Multimodal Fake News Detection. Our framework systematically leverages model-generated explanations in both the retrieval triggering and evidence retrieval modules. It assesses triggering confidence from three complementary dimensions, constructs entity-aware indices by fusing deceptive entities, and retrieves contrastive evidence based on deception-specific features to challenge the initial claim and enhance the final prediction. Experiments on two benchmark datasets, AMG and MR2, demonstrate that ExDR consistently outperforms previous methods in retrieval triggering accuracy, retrieval quality, and overall detection performance, highlighting its effectiveness and generalization capability.

FAIR-RAG: An End-to-End Framework for Mitigating Political Bias through Fair Retrieval-Augmented Generation [Full]

Jaebeom You (Graduate School of Data Science, Seoul National University of Science and Technology); Kisung Lee (Division of Computer Science, Louisiana State University); Hyuk-Yoon Kwon (Graduate School of Data Science, Seoul National University of Science and Technology)

Retrieval-Augmented Generation (RAG) systems can amplify political bias from underlying web corpora. To empirically demonstrate this amplification, we first analyze 16,254 documents from the C4 dataset and 24,300 LLM-generated responses, revealing significant left-leaning and supportive stance bias that can propagate strongly from retrieval to generation. To mitigate this amplification of political bias, we propose FAIR-RAG, an end-to-end framework integrating (1) multi-LLM persona-based annotation, (2) a vector database with political-stance metadata, and (3) a multi-stage fairness engine designed for each of the three stages in RAG systems. FAIR-RAG achieves Attention Weighted Rank Fairness of 97.51 (82.1% improvement) and Perspective Balance of 51.01/82.37 (average 5.6% improvement over state-of-the-art) while maintaining high output quality (Context Precision: 0.974/0.975, Faithfulness: 0.994/0.996). Ablation studies confirm that all three components must operate collaboratively for optimal bias mitigation. This work provides a foundational framework for developing trustworthy and equitable AI information systems. All source code and experimental scripts are publicly available at: https://github.com/bigbases/FAIR-RAG.

PurifAI: Detecting and Fixing Search-Induced Distortions in Web-Augmented LLMs [Full]

Guoqing Wang (Peking University); Zhao Zhang (Peking University); Zeyu Sun (Institute of Software Chinese Academy of Sciences); Xiaofei Xie

(Singapore Management University); Yizhou Chen (Peking University); Yanchao Tan (Fuzhou University); Dan Hao (Peking University)
As Large Language Models (LLMs) increasingly serve as interfaces for proprietary data (e.g., enterprise knowledge bases, legal statutes), ensuring their fidelity to trusted internal information is paramount. While integrating real-time web search can enhance model utility, it introduces a critical vulnerability: the ingestion of conflicting, misleading, or hallucinated content from the open web can override the model's adherence to its verified internal knowledge. We define this failure mode as search-induced distortion, a significant risk in high-stakes domains where the internal knowledge base serves as the absolute ground truth. To address this challenge, we present PurifAI, a proactive, model-agnostic, cache-level purification system designed for safety- and compliance-sensitive deployments. Rather than serving as a general fact-checking engine, PurifAI is explicitly designed to preserve knowledge alignment with a pre-defined trusted knowledge core. It automatically generates diagnostic probes from trusted documents to identify and neutralize searched web content that conflicts with the canonical internal source before such content distorts the LLM's responses. Extensive evaluations on news, encyclopedic, and legal domains show that PurifAI effectively improves alignment with the trusted core, identifying and blocking distortive content at the source and achieving repair success rates above 70% across mainstream LLMs. Our work offers a practical safeguard for enterprises and other high-stakes applications seeking to integrate web-augmented LLMs without compromising policy consistency and trusted knowledge integrity.

LLM-based Evaluation and Relevance Assessment ·

Eureka 3 · Wednesday 10:30–12:00

U-NIAH: Unified RAG and LLM Evaluation for Long Context Needle-in-a-Haystack [TOIS]

Yunfan Gao (Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China); Yun Xiong (Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China); Wenlong Wu (College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China); Bohan Li (College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China); Yijie Zhong (College of Design and Innovation, Tongji University, Shanghai, China); Haofen Wang (College of Design and Innovation, Tongji University, Shanghai, China)
Recent advancements in Large Language Models (LLMs) have significantly extended context windows, igniting discussions about the necessity of Retrieval-Augmented Generation (RAG). U-NIAH, a unified Needle-In-A-Haystack (NIAH) framework, systematically evaluates LLMs and RAG methods in controlled long-context settings. It extends beyond traditional NIAH by incorporating more practical and complex scenarios like multi-needle, long-needle, and needle-in-needle configurations and leveraging the synthetic dataset to mitigate LLM biases. The experiments aim to address three research questions in long-context scenarios: (1) performance trade-offs between LLMs and RAG, (2) error patterns in RAG, and (3) RAG's limitations in complex settings. Results show that smaller LLMs benefit more from RAG. In all settings, RAG achieves a win rate of 82.58% over direct answers. Additionally, it is found that retrieval noise and chunk ordering degrade RAG performance, and we further summarized typical error patterns, including omissions due to noise, hallucinations under high noise critical conditions, and self-doubt behaviors, as well as how these phenomena vary with context length. Finally, in some challenging scenarios, experiments show that deep reasoning models are more easily affected by distractors. These findings highlight the complementary roles of RAG and LLMs and offer actionable insights for optimizing deployment strategies.

Hybrid Pooling with LLMs via Relevance Context Learning [Full]

David Otero (IRLab, CITIC, Universidade da Coruña); Javier Parapar (IRLab, CITIC, Universidade da Coruña)
High-quality relevance judgements over large query sets are essential for evaluating Information Retrieval (IR) systems, yet manual annotation remains costly and time-consuming. Large Language Models (LLMs) have recently shown promise as automatic relevance assessors, but their reliability is still limited. Most existing approaches rely on zero-shot prompting or in-context learning (ICL) with a small number of labelled examples. However, standard ICL treats examples as independent instances and fails to explicitly capture the underlying relevance criteria of a topic, restricting its ability to generalise to unseen query-document pairs. To address this limitation, we introduce Relevance Context Learning (RCL), a novel framework that leverages human relevance judgements to explicitly model topic-specific relevance criteria. Rather than directly using labelled examples for in-context prediction, RCL first prompts an LLM (Instructor LLM) to analyse sets of judged query-document pairs and generate explicit narratives that describe what constitutes relevance for a given topic. These relevance narratives are then used as structured prompts to guide a second LLM (Assessor LLM) in producing relevance judgements. To evaluate RCL in a realistic data collection setting, we propose a hybrid pooling strategy in which a shallow depth-k pool from participating systems is judged by human assessors, while the remaining documents are labelled by LLMs. Experimental results demonstrate that RCL substantially outperforms zero-shot prompting and consistently improves over standard ICL. Overall, our findings indicate that transforming relevance examples into explicit, context-aware relevance narratives is a more effective way of exploiting human judgements for LLM-based IR dataset construction.

Learning to Rank with Multi-Criteria LLM-Judge Annotations [Full]

Naghmeh Farzi (University of New Hampshire); Laura Dietz (University of New Hampshire)
Large Language Models (LLMs) are increasingly used as automated judges (LLM judges) to evaluate Information Retrieval (IR) systems, offering a cost-effective complement to human assessments. However, most prior work treats evaluation mainly as a tool for comparison rather than as a signal for improving the retrieval system. We study whether criterion grades from Multi-Criteria LLM-Judge relevance labeling, which decomposes relevance into Exactness, Coverage, Topicality, and Contextual Fit, can serve as effective ranking features for learning-to-rank (L2R). We evaluate this approach using manual relevance labels from TREC TREC DL 2019, DL 2020, and DL 2023. We then analyze how the Multi-Criteria LLM-Judge feature importance varies relative to retrieval scores across IR system performance levels. The results show that although weaker IR systems benefit the most from this treatment, we see improvements ranging from 14% to 36% in NDCG@20 across diverse IR systems. Contextual Fit often becomes the primary discriminator, exceeding the importance of the retrieval scores. As system quality increases, retrieval scores become the dominant feature, with Multi-Criteria LLM-Judge grades providing complementary, dataset-dependent signals. These findings show that LLM-based relevance labeling can go beyond comparison by turning Multi-Criteria LLM-Judge grades into effective ranking features for reranking.

Topic-Specific Classifiers are Better Relevance Judges than Prompted LLMs [Full]

Lukas Gienapp (University of Kassel); Martin Potthast (University of Kassel); Andrew Yates (Johns Hopkins University); Harrisen Scells (University of Tübingen); Eugene Yang (Johns Hopkins University)
The unjudged document problem, where systems that did not contribute to the original judgement pool may retrieve documents without a relevance judgement, is a key obstacle to the reuseability of test collections in information retrieval. This is commonly dealt with by treating unjudged documents as non-relevant, and recently, the use of large language models (LLMs) as a relevance judge (LLM-as-a-judge) emerged. However, this has been criticized, among other things, as circular, since the same LLM can be used as the ranker and the judge. We propose to train topic-specific relevance classifiers instead: By finetuning monoT5 with independent LoRA weight adaptation on the judgments of a single assessor for a single topic's pool, we align it to that assessor's notion of relevance for that topic. The system rankings obtained through our classifier's relevance judgments achieve a Spearmans' ρ correlation of >0.94 with ground truth system rankings. As little as 128 initial human judgments per topic suffice to improve the comparability of models, while achieving more reliability than existing LLM-as-a-judge approaches, and maintaining human judgments as the gold standard for retrieval evaluation. Code, models, and data are available.

How Variability Influences Podcast Search: Queries, Transcriptions, and Judges [Full]

Watheq Mansour (The University of Queensland); J. Shane Culpepper (The University of Queensland); Andrew Yates (Johns Hopkins University); Joel Mackenzie (The University of Queensland)
Podcasts have continued to grow in popularity over the last two decades, with more than 4.52 million podcasts and 584 million listeners across the globe in 2025. Developing effective search systems for web-scale podcast corpora is of vital importance. Previous research has approached the search task primarily through text representation using a single transcription of the audio content via automatic speech recognition (ASR) models. However, there is currently limited understanding about how variation in podcast representations influences ranking, retrieval, and relevance assessment. In this paper, we provide a comprehensive analysis of the effectiveness of the podcast search problem from three different sources of variation. First, we create a set of 5,745 unique, automatically generated query variations derived from the TREC 2020 and 2021 podcast track topic descriptions. Second, to enable a comprehensive and reproducible analysis, we build a diverse pool composed of a set of 17 different ranked retrieval runs (system variations) using state-of-the-art (SOTA) information retrieval (IR) systems for all of the query variations over a set of corpus (transcript) variations. Third, we use five large language models (LLMs) to automatically assess the relevance of query-document pairs resulting from every query-system-transcription combination, yielding over 1.8 million judged pairs. Our analysis studies the impact of these variations on both search and judgment effectiveness, resulting in a series of practical recommendations for podcast and audio retrieval tasks.

Formalized Information Needs Improve Large-Language-Model Relevance Judgments [Full]

Jüri Keller (TH Köln); Maik Fröbe (Friedrich-Schiller-Universität Jena); Björn Engelmann (TH Köln); Fabian Haak (TH Köln); Timo Breuer (TH Köln); Birger Larsen (Aalborg University); Philipp Schaer (TH Köln)
Cranfield-style retrieval evaluations with too few or too many relevant documents or with low inter-assessor agreement on relevance can reduce the reliability of observations. In evaluations with human assessors, information needs are often formalized as retrieval topics to avoid an excessive number of relevant documents while maintaining good agreement. However, emerging evaluation setups that use Large Language Models (LLMs) as relevance assessors often use only queries, potentially decreasing the reliability. To study whether LLM relevance assessors benefit from formalized information needs, we synthetically formalize information needs with LLMs into topics that follow the established structure from previous

human relevance assessments (i.e., descriptions and narratives). We compare assessors using synthetically formalized topics against the LLM-default query-only assessor on the~2019/2020~editions of TREC Deep Learning and Robust04. We find that assessors without formalization judge many more documents relevant and have a lower agreement, leading to reduced reliability in retrieval evaluations. Furthermore, we show that the formalized topics improve agreement between human and LLM relevance judgments, even when the topics are not highly similar to their human counterparts. Our findings indicate that LLM relevance assessors should use formalized information needs, as is standard for human assessment, and synthetically formalize topics when no human formalization exists to improve evaluation reliability.

Efficient Retrieval, Training, and Indexing · *Courtyard*

1+2 · Wednesday 10:30–12:00

ARHN: Answer-Centric Relabeling of Hard Negatives with Open-Source LLMs for Dense Retrieval [Full]

Hyewon Choi (LG AI Research); Jooyoung Choi (LG AI Research); Hansol Jang (LG AI Research); Hyun Kim (LG AI Research); Chulmin Yun (LG AI Research); Changwook Jun (LG AI Research); Stanley Jungkyu Choi (LG AI Research)

Neural retrievers are often trained on large-scale triplet data comprising a query, a positive passage, and a set of hard negatives. In practice, hard-negative mining can introduce false negatives and other ambiguous negatives, including passages that are relevant or contain partial answers to the query. Such label noise yields inconsistent supervision and can degrade retrieval effectiveness. We propose ARHN (Answer-centric Relabeling of Hard Negatives), a two-stage framework that leverages open-source LLMs to refine hard negative samples using answer-centric relevance signals. In the first stage, for each query–passage pair, ARHN prompts the LLM to generate a passage-grounded answer snippet or to indicate that the passage does not support an answer. In the second stage, ARHN applies an LLM-based listwise ranking over the candidate set to order passages by direct answerability to the query. Passages ranked above the original positive are relabeled to additional positives. Among passages ranked below the positive, ARHN exclude any that contain an answer snippet from the negative set to avoid ambiguous supervision. We evaluated ARHN on the BEIR benchmark under three configurations: relabeling only, filtering only, and their combination. Across datasets, the combined strategy consistently improves over either step in isolation, indicating that jointly relabeling false negatives and filtering ambiguous negatives yields cleaner supervision for training neural retrieval models. By relying strictly on open-source models, ARHN establishes a cost-effective and scalable refinement pipeline suitable for large-scale training.

Constructing Hard-Positive Query–Document Pairs for Dense Retrieval via Phrase Representativeness [Full]

Zhanyu Wu (Beihang University); Richong Zhang (Beihang University); Zhijie Nie (Beihang University)

Dense retrieval usually fails in two ways: ranking non-relevant documents too high, or ranking relevant documents too low. We focus on the second case from the query side: documents that are genuinely relevant but receive low retrieval scores under certain query formulations, forming hard-positive query–document pairs. To study this failure mode systematically, we build on recent token-alignment work, which analyzes vocabulary logits obtained by projecting retriever representations through an architecture-specific token prediction head and shows that low overlap among top-ranked logit tokens can cause relevant documents to be scored low. This suggests a practical route to hard positives: generate relevant queries that rely on document phrases that receive low ranks under these logits. We therefore define model-specific Token and Phrase Representativeness Scores (TRS/PRS) to discover tokens and key phrases that appear in a document but are poorly expressed by its embedding. Using high-PRS phrases as anchors, we automatically construct challenging yet relevant queries, yielding hard-positive query–document pairs. Experiments across multiple retrievers and datasets show that PRS-anchored queries induce substantially larger drops for dense retrievers than other query construction strategies, revealing a dense-specific weakness. Moreover, when mixed with standard hard negatives during contrastive fine-tuning, these hard positives improve retrieval on our constructed hard-positive benchmark and can also improve standard in-domain retrieval benchmarks. Our code is publicly available at https://github.com/wzy2001wzy/HardPositive.

Reproducing Complex Set-Compositional Information Retrieval [Reproducibility]

Vincent Degenhart (Radboud University); Dewi Timman (Radboud University); Arjen de Vries (Radboud University); Faegheh Hasibi (Radboud University); Mohanna Hoveyda (Radboud University)

Complex information needs may involve set-compositional queries using conjunction, disjunction, and exclusion, yet it remains unclear whether current retrieval paradigms genuinely satisfy such constraints or exploit `semantic shortcuts'. We conduct a reproducibility study to benchmark major retrieval families and reasoning-targeted methods on QUEST and QUEST+Variants, and introduce LIMIT+, a controlled benchmark where relevance depends on arbitrary attribute predicates and constraint satisfaction, and less on pretrained knowledge. Our findings show that (i) on QUEST, the best neural retrievers achieve an effectiveness that is more than double what can be achieved with BM25 (Recall@100 ${>}$0.41 vs.\ 0.20), but reasoning-targeted methods like ReasonIR and Search-R1 do not outperform general-purpose retrievers uniformly; (ii) on LIMIT+, gains fail to transfer, where the strongest QUEST method collapses from Recall@100${\approx}$0.42 to below 0.02, while classic lexical retrieval gains to ${\sim}$0.96. Lastly, (iii) stratifying by compositional depth reveals a consistent degradation across all methods, where algebraic sparse and lexical methods show more stable performance while dense approaches collapse.We release code and LIMIT+ data generation scripts to support future reproducibility and controlled evaluation.

Hypencoder Revisited: Reproducibility and Analysis of Non-Linear Scoring for First-Stage Retrieval [Reproducibility]

Arne Eichholtz (University of Amsterdam); Yongkang Li (University of Amsterdam); Jutte Vijverberg (University of Amsterdam); Tobias Groot (University of Amsterdam); Mohammad Aliannejadi (University of Amsterdam)

The Hypencoder, proposed by Killingback et al., is a retrieval framework that replaces the fixed inner-product scoring function used in standard bi-encoders with a query-specific neural network (the q-net), whose weights are generated by a hypernetwork from the contextualized query embeddings. This design enables more expressive relevance estimation while preserving independent query and document encoding. In this work, we conduct a reproducibility study of the Hypencoder and extend the original analysis in three directions. Our reproduction confirms that the Hypencoder outperforms a similarly trained bi-encoder baseline on in-domain and out-of-domain benchmarks, and that the proposed efficient search algorithm substantially reduces query latency with minimal performance loss. On hard retrieval tasks, we find partial support: the Hypencoder outperforms the baseline on DL-Hard and FollowIR, but not on TREC TOT, where checkpoint incompatibility and fine-tuning sensitivity complicate full verification. Beyond reproduction, we investigate three extensions: (i)~integrating alternative pre-trained encoders into the Hypencoder framework, where we find that performance gains depend on the encoder and fine-tuning strategy; (ii)~comparing query latency against a Faiss-based bi-encoder pipeline, revealing that standard bi-encoder retrieval remains faster under both exhaustive and efficient search settings; and (iii)~evaluating adversarial robustness, where we find that the q-net's non-linear scoring does not provide a consistent robustness disadvantage over inner-product scoring. Our code is publicly available at \url{https://github.com/arneeichholtz/Hypencoder-reprod}.

Benchmarking Filtered Approximate Nearest Neighbor Search Algorithms on Transformer-based Embedding Vectors [Full]

Patrick Iff (ETH Zurich); Paul Brügger (ETH Zurich); Marcin Chrapek (ETH Zurich); David Kochergin (ETH Zurich); Maciej Besta (ETH Zurich); Torsten Hoefler (ETH Zurich)

Advances in embedding models for text, image, audio, and video drive progress across multiple domains, including retrieval-augmented generation, recommendation systems, and others. Many of these applications require an efficient method to retrieve items that are close to a given query in the embedding space while satisfying a filter condition based on the item's attributes, a problem known as filtered approximate nearest neighbor search (FANNS). By performing an in-depth literature analysis on FANNS, we identify a key gap in the research landscape: publicly available datasets with embedding vectors from state-of-the-art transformer-based text embedding models that contain abundant real-world attributes covering a broad spectrum of attribute types and value distributions. To fill this gap, we introduce the arxiv-for-fanns dataset of transformer-based embedding vectors for the abstracts of over 2.7 million arXiv papers, enriched with 11 real-world attributes such as authors and categories. We benchmark eleven different FANNS methods on our new dataset to evaluate their performance across different filter types, numbers of retrieved neighbors, dataset scales, and query selectivities. We distill our findings into eight key observations that guide users in selecting the most suitable FANNS method for their specific use cases.

GIGP+: A CPU-GPU Co-Processing Engine for Multi-Vector Retrieval [Full]

Zheng Bian (Hong Kong Polytechnic University); Man Lung Yiu (Hong Kong Polytechnic University); Bo Tang (Southern University of Science and Technology)

Multi-vector retrieval models (e.g., ColBERTv2) offer high retrieval accuracy but suffer from efficiency problems at scale. Recently, several methods have been developed to enhance the efficiency of multi-vector retrieval. On one hand, the state-of-the-art GPU-based method PLAID-GPU exploits the massive parallelism of the GPU to accelerate computation, but it needs to process a considerable amount (e.g., ten thousand) of document candidates. On the other hand, the state-of-the-art (SOTA) CPU-based method IGP employs a more effective strategy to reduce the number of candidates, but fails to utilize the massive parallelism of the GPU. To get the best of both worlds, we propose GIGP+, a GPU-based method designed to achieve high parallelism and low computational overhead. Our contributions are: (1) an efficient candidate generation kernel that enjoys parallelism while retaining the effectiveness of IGP, (2) a score reordering mechanism that reduces the synchronization overhead and (3) a scheduling strategy for efficient batch processing. Our experiments demonstrate that GIGP+ achieves a 11.0× improvement in query per second (QPS) and reduces latency by 7.6× compared to PLAID-GPU, while maintaining equivalent retrieval accuracy. As for cloud pricing, GIGP+ delivers a 2.3× improvement in queries per dollar over SOTA CPU-based solutions.

Graph Collaborative Filtering · *Goldfields · Wednesday*

13:00–14:30

Graph Diffusion Gated Embeddings for Recommender Systems [Full]

Seungcheol Lee (LINE Plus Corporation); Taeyoung Roh (LINE Plus Corporation); Jiho Seo (LINE Plus Corporation); Soohyun Lim (LINE Plus Corporation)

Modern large-scale recommender systems rely on dense user and item embeddings, but using a full embedding table requires a large number of parameters. To reduce this cost, hashing tricks are commonly used to compress embedding tables. However, collisions and hard bucket assignments can degrade both performance and personalization. We introduce Graph Diffusion Gated Embeddings (GDE), which precompute multi-scale, type-separated heat-kernel diffusions from a small set of user/item seeds (via HK-relax) and convert them into degree-corrected probabilistic gating distributions. These gates mix a compact set of learnable seed-cluster embeddings to form node representations. Ablations show that the cross-only variant (cGDE), which uses only cross-type gating, achieves the strongest performance. Across both Top-K retrieval and click through rate prediction tasks, cGDE outperforms hashing-based baseline methods. Our code is available at https://github.com/latent-x/Graph-Diffusion-Gated-Embeddings-for-Recommender-Systems.

Disentangling from Collaborative and Semantic Views: Graph Collaborative Filtering for Q&A Recommendation [Full]

Changshuo Zhang (Renmin University of China); Teng Shi (Renmin University of China); Xiao Zhang (Renmin University of China); Yanping Zheng (Renmin University of China); Ruobing Xie (Tencent); Qi Liu (Tencent); Jun Xu (Renmin University of China)

Question and answer (Q&A) platforms usually recommend question-answer pairs to meet users' knowledge acquisition needs, unlike traditional recommendations that recommend only one item. This makes user behaviors more complex, and presents two challenges for Q&A recommendation, including: the collaborative information entanglement, which means user feedback is influenced by either the question or the answer; and the semantic information entanglement, where questions are correlated with their corresponding answers, and correlations also exist among different question-answer pairs. Traditional recommendation methods treat the question-answer pair as a whole or only consider the answer as a single item, which overlooks the two challenges and cannot effectively model user interests. To address these challenges, we introduce a graph neural network model named Question & Answer Graph Collaborative Filtering (QAGCF). QAGCF creates graphs separately from collaborative and semantic views to disentangle the collaborative and semantic information of question-answer pairs. The collaborative view disentangles questions and answers to individually model collaborative information, while the semantic view captures the semantic information both within and between question-answer pairs. These views are further merged into a global graph to integrate the collaborative and semantic information. Polynomial-based graph filters are used to address the high heterophily issues of the global graph. Additionally, contrastive learning is utilized to obtain robust embeddings during training. Extensive experiments on industrial and public datasets demonstrate that QAGCF consistently outperforms baselines and achieves state-of-the-art results.

Mind the Metric: Reproducibility and Fair Benchmarking of Spectral Graph Models for Collaborative Filtering [Reproducibility]

Domenico de Gioia (Politecnico di Bari); Claudio Pomo (Politecnico di Bari); Ludovico Boratto (University of Cagliari); Tommaso Di Noia (Politecnico di Bari)

Graph models that manipulate the frequency spectrum of user-item interactions to separate preference signals from noise often report significant improvements, but concerns about evaluation rigor and reproducibility persist. We conduct a reproducibility and replicability study that examines three major families: (i) spectral denoising methods, (ii) graph signal processing (GSP) models, and (iii) spectral propagation approaches. Reproducing published pipelines reveals a polarized landscape: while several works are fully reproducible, others rely on flawed metric implementations and incomplete hyperparameter disclosures. In particular, we observe systematic inflation of Recall in the spectral denoising methods due to an implementation error, and theoretically invalid ranking metrics in GSP models due to unordered prediction lists; conversely, the graph filtering models are consistently reproducible. Beyond reproduction, we establish a unified evaluation protocol on four datasets with consistent splits and hyperparameter optimization for all baselines, showing that strong classical methods (e.g., SLIM, Item-kNN) remain highly competitive and that no single spectral model dominates across domains. We further analyze robustness under varying data sparsity and assess beyond-accuracy properties, finding that spectral filtering often improves catalog exploration even when accuracy gains are marginal. Our code is available at https://github.com/sisinflab/Mind_the_Metric_SIGIR-26.

Verbalizing LightGCN: Direct Learning of Textual Representations from User-Item Interaction Graph via LLMs [Full]

Manh-Khanh Ngo Huu (Singapore Management University); Hady W. Lauw (Singapore Management University)

In this work, we propose VerbaLightGCN, a novel LLM-based recommendation framework that integrates the semantic understanding of LLMs with user-item interaction modeling. Traditional collaborative filtering (CF) models typically embed user and item IDs into a latent space to capture interaction signals. However, pretrained LLMs cannot natively interpret these learned embeddings. To bridge this gap, VerbaLightGCN adopts a CF-as-text paradigm, in which collaborative signals are encoded in textual form and directly learned from the user–item interaction graph, and are then combined with semantic information to construct user and item profiles that function as latent embeddings. Inspired by LightGCN, our method retains its message-passing design but replaces numerical embedding computations with a Chain-of-Thought prompting mechanism. This enables LLMs to simulate the LightGCN aggregation process through natural language. The result is a recommendation framework that unifies semantic understanding with collaborative signals in a fully language-native form. Experiments show that VerbaLightGCN achieves superior performance to both zero-shot LLM-based and traditional CF-based baselines. Further analysis reveals that the user and item profiles generated by VerbaLightGCN effectively capture both semantic preferences and collaborative filtering signals.

Chunk-Wise Quantization for Graph Collaborative Filtering [Full]

Kaixi Hu (Wuhan Textile University); Peipei Wang (Qilu University of Technology); Kaize Shi (University of Southern Queensland); Jingling Yuan (Wuhan University of Technology); Yu Yang (The Education University of Hong Kong); Guandong Xu (The Education University of Hong Kong); Lin Li (Wuhan University of Technology)

Energy efficiency has become a critical requirement, driving recommendation systems for resource-constrained environments such as edge devices. Model quantization offers an effective way to build low-bitwidth models while preserving accuracy. However, user–item interaction graphs contain numerous nodes and complex topological structures, leading nodes to exhibit unique similarities and differences. Existing quantization methods uniformly process parameters in high-dimensional DNN layers (e.g., linear, convolutional, or attention layers), while inadequately capturing such similarities among node embeddings. This paper proposes GraphQ, a chunk-wise quantization framework for graph collaborative filtering that supports both the training and post-training phases in a unified perspective. Our core idea is to adaptively partition node embeddings into multiple chunks based on the distribution of embedding values, and then apply chunk-wise quantization. Specifically, for quantization-aware training (QAT), we introduce learnable low-precision quantization factors that partition node embeddings into multiple chunks and are dynamically updated following message passing. For post-training quantization (PTQ), we first cluster nodes and then partition their dimensions into chunks for weight clipping. Extensive experiments on four real-world datasets show that GraphQ outperforms state-of-the-art QAT methods by an average of 27.49% in Recall@10 under the 256-dimensional embedding and 2-bit settings, and surpasses PTQ methods by 78.64% on average under 4-bit settings.

Fourier Kolmogorov-Arnold Network and Hypergraph Enhanced Contrastive Learning for Recommendation [Full]

Yuwen Liu (China University of Petroleum (East China)); Lianyong Qi (China University of Petroleum (East China)); Xucheng Zhou (China University of Petroleum (East China)); Xingyuan Mao (China University of Petroleum (East China)); Weiming Liu (Zhejiang University); Shuang Wang (China University of Petroleum (East China)); Xiaolong Xu (Nanjing University of Information Science and Technology); Haolong Xiang (Nanjing University of Information Science and Technology); Xuyun Zhang (Macquarie University); Wanchun Dou (Nanjing University)

Recommendation plays a crucial role in the modern Web ecosystem, powering personalized services across e-commerce, social platforms, and online content networks. To model complex user–item interactions in such Web environments, Graph Neural Networks (GNNs) have become a popular and effective approach due to their ability to capture relational dependencies. However, existing GNN-based methods face some challenges, such as limited capacity for nonlinear representation, inability to capture global structural information, and susceptibility to noise in user interaction data. Although self-supervised learning methods have been introduced to address these issues, these methods often overlook the intricate dependencies between users and items and fail to effectively utilize high-order global information. To address these challenges, we propose Fourier Kolmogorov-Arnold Network and Hypergraph Enhanced Contrastive Learning (FHCL) for recommendation. Our method constructs two complementary views: a graph generative view and a denoising view. In the graph generative view, we use the Fourier Kolmogorov-Arnold Network (Fourier KAN) to enhance the nonlinear representation capabilities by decomposing complex user-item interactions. Subsequently, we employ Variational Graph Auto-Encoders (VGAE) to reconstruct the graph structure, extracting meaningful structural information while mitigating the impact of noise. Then we use hypergraph learning to capture high-order global dependencies. In the denoising view, we introduce a denoising matrix to filter noisy edges and further combine hypergraph learning to improve user preferences. Finally, we integrate these views through contrastive learning to generate robust and accurate recommendations. Extensive experiments on two public datasets demonstrate the superior performance of FHCL, while comprehensive ablation studies validate the necessity and effectiveness of each component.

Sequential Recommendation · *Room 110 · Wednesday 13:00–14:30*

Divergence Meets Consensus: A Multi-Source Negative Sampling Framework for Sequential Recommendation [Full]

Yuanzi Li (Renmin University of China); Lingjie Wang (Shandong University); Jingyu Zhao (Shandong University); Zihang Tian (Renmin University of China); Yuhan Wang (Peking University); Lei Wang (Renmin University of China); Xu Chen (Renmin University of China)

Negative sampling is significant for training sequential recommendation models under implicit feedback. The predominant strategy, self-guided hard negative sampling, selects negatives based on the model's current state but suffers from three limitations: (1) the coupling between sampling and model updates triggers a vicious cycle that drives the model into local optima; (2) relying on current model parameters narrows sampling to a small region of the item space, reducing diversity and harming generalization; (3) identifying a hard negative requires scoring the entire candidate pool, causing substantial computational overhead with minimal information gain. To address these challenges, we propose MDCNS (Multi-source Divergence-Consensus for Negative Sampling), a novel ``Teacher-Peer-Self'' framework inspired by Vygotsky's Zone of Proximal Development (ZPD) theory. The proposed method comprises three components, including multi-source scoring, divergence re-ranking, and consensus distillation. Firstly, multi-source scoring incorporates peer and ensemble teacher models to inject external negative signals and break the self-reinforcement loop. Then, divergence re-ranking exploits prediction discrepancy between self and peer models to enhance sampling diversity. Finally, consensus distillation aligns the self-model with the teacher via KL divergence, simultaneously improving computational cost utilization. Extensive experiments on six real-world datasets and five backbone models show that MDCNS consistently outperforms state-of-the-art negative sampling methods, demonstrating strong effectiveness and generalization.

CATS: Cluster-Aware Thompson Sampling for Negative Mining in Sequential Recommendation [Full]

Giulia Di Teodoro (University of Pisa); Federico Siciliano (National Institute of Geophysics and Volcanology); Nicola Tonellotto (University of Pisa); Fabrizio Silvestri (Sapienza University of Rome)

Modern sequential recommendation systems often rely on negative sampling to efficiently train models over vast item corpora. However, common strategies such as uniform, popularity-based sampling, or hard-negative mining, often yield uninformative negatives, introduce popularity bias, or suffer from false negatives that hinder model learning. While several studies focus on identifying true negatives, none explore latent item representations to mitigate false negatives issue. We propose ■CATS■ (Cluster-Aware Thompson Sampling),an adaptive negative sampling method that balances exploration and exploitation by leveraging unsupervised item clustering and adaptive multi-armed bandit principles. The key insight behind CATS is that false negatives tend to lie in the same cluster as the positive item, whereas true hard negatives are more likely found in nearest clusters. CATS leverages this by adaptively sampling negatives from the positive item's cluster, its neighboring clusters, and the rest of the item space, and decaying the sampling probability from the positive item's cluster over time, reducing the risk of accumulating false negatives during training. Evaluation using state-of-the-art sequential models (SASRec, BERT4Rec, BSARec) and three benchmark datasets (MovieLens-1M, Amazon Beauty and BeerAdvocate) demonstrates that CATS consistently outperforms standard and advanced sampling baselines. Notably, CATS achieves improvements of over 20% in NDCG@10 compared to vanilla sampling methods, and of over 16% compared to more advanced methods. Our findings suggest that incorporating the latent item structure through adaptive sampling strategies like CATS can significantly enhance the performance of sequential recommendation systems. The complete code needed to reproduce the results presented in this paper can be found in our GitHub repository https://github.com/gditeodoro/CATS.

Balanced Frequency Decoupling: Energy-Aware Multi-Scale Preference Modeling for Sequential Recommendation [Full]

Jiahao Hu (Chongqing University); Wei Zhou (Chongqing University); Jie Liao (Chongqing University); Junlin Zhu (University of Electronic Science and Technology of China); Junhao Wen (Chongqing University); Hongyu Zhang (Chongqing University)

Frequency-domain sequential recommendation models enhance sequence representation capacity through spectral transformations. However, existing methods typically adopt coarse-grained spectrum reweighting strategies that strengthen high-frequency components while amplifying random noise within frequency bands. Moreover, they generally rely on a coupled modeling mechanism that handles both long-term and short-term preferences within a single backbone network, lacking dedicated modeling paths tailored to their distinct temporal characteristics. To address these challenges, we propose a Balanced Frequency Decoupling Sequential Recommendation model (BFDRec). Specifically, we design an energy-aware spectrum denoising mechanism to adaptively suppress low-energy noises according to the energy distribution within each frequency band while preserving salient behavioral fluctuation signals. Additionally, we construct a multi-scale decoupled architecture to model users' multi-scale preferences and adaptively integrate them through a dynamic gating mechanism, aligning the sequence modeling process with the distinct temporal characteristics of different frequency bands. Extensive experiments across five real-world datasets demonstrate that BFDRec effectively achieves noise suppression and accurate multi-scale preference modeling, with average improvements of 6.67% and 5.05% in HR and NDCG, respectively, over advanced baseline models. Our code is available at https://anonymous.4open.science/r/BFDRec.

Pay Attention to Sequence Split: Uncovering the Impacts of Sub-Sequence Splitting on Sequential Recommendation Models [Reproducibility]

Yizhou Dang (Northeastern University); Yifan Wu (Northeastern University); Minhan Huang (Northeastern University); Chuang Zhao (Tianjin University); Lianbo Ma (Northeastern University); Guibing Guo (Northeastern University); Xingwei Wang (Northeastern University); Zhu Sun (Singapore University of Technology and Design)

Sub-sequence splitting (SSS) has been demonstrated as an effective approach to mitigate data sparsity in sequential recommendation (SR) by splitting a raw user interaction sequence into multiple sub-sequences. Previous studies have demonstrated its ability to enhance the performance of SR models significantly. However, in this work, we discover that (i). SSS may interfere with the evaluation of the model's actual performance. We observed that many recent state-of-the-art SR models employ SSS during the data reading stage (not mentioned in the papers). When we removed this operation, performance significantly declined, even falling below that of earlier classical SR models. The varying improvements achieved by SSS and different splitting methods across different models prompt us to analyze further when SSS proves effective. We find that (ii). SSS demonstrates strong capabilities only when specific splitting methods, target strategies, and loss functions are used together. Inappropriate combinations may even harm performance. Furthermore, we analyze why sub-sequence splitting yields such remarkable performance gains and find that (iii). it evens out the distribution of training data while increasing the likelihood that different items are targeted. Finally, we provide suggestions for overcoming SSS interference, along with a discussion on data augmentation methods and future directions. We hope this work will prompt the broader community to re-examine the impact of data splitting on SR and promote fairer, more rigorous model evaluation. All analysis code and data are available at https://github.com/KingGugu/SSS4SR.

Coupling Global Context with Kinematic Evolution for Sequential Recommendation [Full]

Yueting Yang (Sichuan University); Yao Li (Sichuan University); Rongmei Zhao (Chengdu University); Shenggen Ju (Sichuan University)

Accurately modeling the continuous evolution of user preferences remains a fundamental challenge in sequential recommendation. While existing models have made significant strides in capturing sequential dependencies, they predominantly encode historical behaviors via discrete aggregation, often failing to explicitly model the continuous evolution trends and intrinsic momentum underlying preference shifts. To address this limitation, we propose Kinematic Evolution for Sequential Recommendation (KERec), which models user preference evolution from a continuous kinematic perspective. KERec adopts a dual-branch encoding architecture: the global context branch captures stable long-term preference structures, while the kinematic evolution branch explicitly models the first-order velocity and second-order acceleration of preference evolution based on Taylor series expansion and multi-scale difference operators. To effectively integrate these heterogeneous signals, we introduce an adaptive feature fusion mechanism that balances steady-state consistency with dynamic sensitivity. Furthermore, we design a momentum-guided contrastive learning paradigm that constrains augmentation views via evolutionary momentum, suppressing random noise while reinforcing the perception of genuine evolutionary trajectories. Extensive experiments on multiple public benchmark datasets demonstrate that KERec significantly outperforms state-of-the-art methods in recommendation accuracy and exhibits superior robustness under varying sequence lengths and noisy levels. The code of our method is available at https://github.com/yyt-1219/KERec.

Time-Interval-Aware Disentangled Expert Modeling for Next-Basket Recommendation [Full]

Zhiying Deng (Central China Normal University); Yuan Fu (Huazhong University of Science and Technology); Usman Farooq (Huazhong University of Science and Technology); Ziwei Tian (Huazhong University of Science and Technology); Wei Liu (Huazhong University of Science and Technology); Jianjun Li (Huazhong University of Science and Technology)

Next-basket recommendation (NBR) is a type of recommendation that aims to predict a set of items a user will purchase based on their historical transaction basket sequences. It is governed by a dynamic interplay between two distinct user intents: habitual repurchase, which involves repeating past behaviors, and exploratory interest, which involves discovering new items. However, existing NBR methods generally suffer from two limitations: (1) they often entangle these conflicting motives within a single representation, causing habits to overshadow discovery, and (2) they rely on discrete sequential modeling that ignores continuous-time intervals and item-specific periodicities. In this paper, we propose a novel solution named Time-Interval Disentangled Experts (TIDE) to address these challenges. TIDE incorporates a Hawkes-enhanced Fourier Time Encoding to capture item-specific temporal periodicities and dynamic decay. To decouple user intentions, TIDE utilizes a dual-expert architecture that integrates a Habit Expert for recurring needs and a Pattern-Guided Exploration Expert for discovery. Combined with an item-aware gating mechanism, TIDE adaptively balances repurchase and exploration. Extensive experiments on four diverse real-world datasets demonstrate that TIDE consistently outperforms representative state-of-the-art NBR methods.

Complex Reasoning and QA · *Eureka 1 · Wednesday*

13:00–14:30

From Existence to Exhaustiveness: Unveiling the Compounding Failures of Large Language Models in Multi-answer Event Temporal Reasoning [Full]

Shaojuan Wu (Shanxi University)

Large Language Models (LLMs) have achieved remarkable success in temporal reasoning. However, existing benchmarks predominantly adopt a "single-answer" paradigm, focusing on verifying the existence of a specific fact while overlooking the challenge of exhaustiveness. In real-world scenarios, entities often simultaneously play multiple roles or exist in multiple states within the same timeframe. To bridge this gap, we introduce MulTR, a comprehensive benchmark designed for multi-answer temporal reasoning from long unstructured contexts. Specifically, MulTR integrates structured temporal facts from Wikidata and natural language text from Wikipedia MulTR integrates structured temporal facts from Wikidata and natural language text from Wikipedia through a logic-driven synthesis process. Notably, we formulate two distinct settings, question-dependent and document-dependent, based on the presence of cue words in the question. It is designed to systematically decouple temporal reasoning capabilities from the uncertainty of the number of answers. Experiment results demonstrate that state-of-the-art models suffer from retrieval laziness, terminating the search process prematurely after locating the first valid piece of evidence. Consequently, their performance drops sharply when evaluated on strict exact match metrics. MulTR, as a diagnostic testing platform, reveal these defects and establish the rigorous standard for future research in dynamic knowledge processing. The MulTR benchmark and evaluation prompt are publicly available at https://github.com/TemporalNLP/MulTR.

STRIDE: Strategic Iterative Decision-Making for Retrieval-Augmented Multi-Hop Question Answering [Full]

Wei Chen (University of Science and Technology of China); Lili Zhao (University of Science and Technology of China); Zhi Zheng (University of Science and Technology of China); Huijun Hou (NIO Inc.); Tong Xu (University of Science and Technology of China)

Multi-hop question answering (MHQA) enables accurate answers to complex queries by retrieving and reasoning over evidence dispersed across multiple documents. Existing MHQA approaches mainly rely on iterative retrieval-augmented generation, which suffer from the following two major issues. On one hand, existing methods prematurely commit to surface-level entities rather than underlying reasoning structures, making question decomposition highly vulnerable to lexical ambiguity. On the other hand, existing methods overlook the logical dependencies among reasoning steps, resulting in uncoordinated execution. To address these issues, we propose STRIDE, a framework that separates strategic planning, dynamic control, and grounded execution. At its core, a Meta-Planner first constructs an entity-agnostic reasoning skeleton to capture the abstract logic of the query, thereby deferring entity grounding until after the reasoning structure is established, which mitigates disambiguation errors caused by premature lexical commitment. A Supervisor then orchestrates sub-question execution in a dependency-aware manner, enabling efficient parallelization where possible and sequential coordination when necessary. By dynamically deciding whether to retrieve new evidence or infer from existing facts, it avoids redundant queries and error propagation, while fusing cross-branch information and reformulating failed queries to enhance robustness. Grounded fact extraction and logical inference are delegated to specialized execution modules, ensuring faithfulness through explicit separation of retrieval and reasoning. While STRIDE is compatible with any large language models (LLMs), off-the-shelf open-source LLMs underperform closed-source counterparts in its structured reasoning pipeline. To close this gap, we further propose STRIDE-FT, a modular fine-tuning framework that uses self-generated execution trajectories from STRIDE, requiring neither human annotations nor stronger teacher models. Experiments show that STRIDE achieves robust and accurate reasoning on MHQA benchmarks, while STRIDE-FT effectively enhances open-source LLMs.

Recursive Short-to-Long Generalization for Multi-hop Reasoning [Full]

Mayi Xu (Wuhan University); Ke Sun (Wuhan University of Science and Technology); Jianhao Chen (Wuhan University); Qiankun Pi (Wuhan University); Guixin Su (Wuhan University); Yunfeng Ning (Wuhan University); Yongqi Li (Wuhan University); Yuanyuan Zhu (Wuhan University); Ming Zhong (Wuhan University); Jiawei Jiang (Wuhan University); Tieyun Qian (Wuhan University)

Reasoning is fundamental to human intelligence and critical for problem-solving. In practice, answering complex questions often requires abundant reasoning steps involving long reasoning paths. Existing multi-hop reasoning methods mainly focus on short-hop questions with only short reasoning paths, whose distribution is different from that of long paths. While training on long-hop data can improve corresponding performance, collecting it at scale is difficult and expensive. Hence, we explore the short-to-long generalization scenario for the first time, which aims to enhance the model's capability to handle long-hop questions using easily collected short-hop reasoning data. In this scenario, the reasoning model can only learn the features of short-hop questions due to the lack of long-hop data, and thus two challenges emerge. (1) Struggling to determine where to go next : as the number of steps increases, the model needs to reason about the next step based on a long context composed of more evidences. This reasoning process will confuse the reasoning model, which can only learn the pattern from a short context to the next step. (2) Hard to determine when to stop reasoning : too many reasoning hops may introduce excessive noise, while too few may fail to gather sufficient supporting information. The path lengths of long-hop questions vary significantly. Thus, a fixed stopping threshold, commonly used for short-hop questions with similar path lengths, cannot handle them effectively. Inspired by the classic recursive algorithm, we propose a novel Recursion-based Short-to-Long Generalization (RSLG) reasoning framework, which recursively decomposes long-hop questions into multiple short-hop questions that can be handled by a short-hop reasoning model. In this way, the above two challenges are transformed into how to decompose and when to stop decomposition. For how to decompose, we introduce the parallel and sequential recursion patterns to guide the recursive decomposition processes. A local-to-global recursive demonstration construction strategy is proposed to obtain the corresponding recursive demonstrations from the perspective of certainty, complexity, and diversity, respectively. For when to stop decomposition, we propose three recursive termination criteria in view of decomposability, redundancy, and relevance. Extensive experiments on six short-to-long settings across diverse domains demonstrate RSLG's superior performance, robustness, and efficiency in this challenging scenario. The code and data link: https://github.com/NLPGM/RSLG

Analytical Search [Perspective]

Yiteng Tu (DCST, Tsinghua University); Shuo Miao (DCST, Tsinghua University); Weihang Su (DCST, Tsinghua University); Yiqun Liu (DCST, Tsinghua University); Qingyao Ai (Quancheng Laboratory)

Analytical information needs, such as trend analysis and causal impact assessment, are prevalent across various domains including law, finance, science, and much more. However, existing information retrieval paradigms, whether based on relevance-oriented document ranking or retrieval-augmented generation (RAG) with large language models (LLMs), often struggle to meet the end-to-end requirements of such tasks at the corpus scale. They either emphasize information finding rather than end-to-end problem solving, or simply treat everything as question answering, offering limited control over reasoning, evidence usage, and verifiability. As a result, they struggle to support analytical queries that have diverse utility concepts and high accountability requirements. In this paper, we propose analytical search as a distinct and emerging search paradigm designed to fulfill these analytical information needs. Analytical search extends search as an evidence-governed, process-oriented analytical workflow that explicitly models analytical intent, retrieves evidence for fusion, and produces verifiable conclusions through structured, multi-step inference. We analyze the limitations of existing search systems in supporting analytical information needs and present a unified example framework that integrates query understanding, recall-oriented retrieval, reasoning-aware fusion, and adaptive verification. We also discuss potential research directions for the construction of analytical search engines. In this way, we highlight the conceptual significance and practical importance of analytical search and call on efforts toward the next generation of search engines that support analytical information needs.

TimelineReasoner: Advancing Timeline Summarization with Large Reasoning Models [Full]

Liancheng Zhang (Renmin University of China); Xiaoxi Li (Renmin University of China); Zhicheng Dou (Renmin University of China)

The proliferation of online news poses a challenge to extracting structured timelines from unstructured content. While recent studies have shown that Large Language Models (LLMs) can assist Timeline Summarization (TLS), these approaches primarily treat models as passive generators. The emergence of Large Reasoning Models (LRMs) presents an opportunity to reason over events actively, enabling iterative evidence acquisition, the detection of missing events, and the validation of temporal consistency. To leverage the reasoning capabilities of LRMs, we propose TimelineReasoner, a novel framework that shifts TLS from static generation to an active, reasoning-driven process. TimelineReasoner adopts a two-stage framework: Global Cognition, which tracks events at a macroscopic level and continuously updates a global event memory, and Detail Exploration, which identifies informational gaps and refines the timeline via targeted document retrieval. To support this, TimelineReasoner incorporates several specialized mechanisms, including an Event Scraper for retrieving temporal event descriptions, a Timeline Updater for refining the timeline, and a Supervisor for detecting gaps in the timeline and guiding retrieval. Experimental results on open-domain TLS datasets demonstrate that TimelineReasoner significantly outperforms existing LLM-based TLS methods in terms of timeline accuracy, coverage, and coherence. On closed-domain TLS datasets, our method performs on par with or exceeds state-of-the-art approaches. This work not only pushes the boundaries of TLS but also highlights the broader potential of LRM-based reasoning frameworks for timeline summarization.

Mitigating Bias in Large Language Model Based Question Answering through Causal Front Door Prompting [Full]

Yaqi Yang (RMIT University); Ziqi Xu (RMIT University); Jie Li (Sportsbet); Chenglong Ma (RMIT University); Jeffrey Chan (RMIT University); Mark Sanderson (RMIT University); Xin Zheng (RMIT University); Yongli Ren (RMIT University)

Large language models (LLMs) are widely used for question answering (QA) but can generate biased or stereotype-driven answers due to demographic associations learned during pre-training. Existing mitigation strategies often rely on model access or fine-tuning, which limits their applicability to closed-source LLMs. We propose a Causal Front Door Prompting framework

(CFDP) that reduces demographic influence by intervening on the chain of thought reasoning, which is treated as an observable mediator. CFDP samples and clusters multiple reasoning traces and estimates answer probabilities through weighted aggregation. Experiments on two widely used bias-sensitive QA benchmarks, BBQ and Stereotype, across major LLMs show that CFDP consistently improves fairness metrics without sacrificing QA accuracy. Ablation and sensitivity analyses confirm the value of each component, indicating that causal intervention on reasoning provides an effective and practical approach for bias mitigation in LLM-based QA.

Conversational and Personalized Information Seeking · *Eureka 2 · Wednesday 13:00–14:30*

NeurPIU: Neurobiologically Inspired Personalized Intent Understanding in Large Language Models [Full]

Zenghua Liao (National University of Defense Technology); Jinzhi Liao (National University of Defense Technology); Xiang Zhao (National University of Defense Technology)

Large language models (LLMs) excel when user goals are clearly specified, yet real-world queries are often vague and evolving, forcing LLMs to guess and leading to misaligned responses. Existing approaches attempt to clarify user intents through iterative questioning. While effective in alleviating disambiguation, this paradigm tends to merely provide standard and normal responses, which fails to meet the growing demand for diverse personalized expression of users. Based on the observation, we first identify its problem as the ignore of users' mental states, in which the complicated mental elements, unclear functional rules, and evolving mental states pose obstacles to approach the problem. Therefore, in this paper, we introduce the theory of the mentalizing network in the human brain and propose a neurobiologically inspired framework, i.e., NeurPIU, that endows LLMs with human-like mentalizing capabilities for personalized intent understanding. NeurPIU constructs an intent neural network that organizes users' long-term mental states into a three-layer graph; retrieves query-relevant states via spreading activation mechanism with temporal decay; and injects an encoded cognitive prefix into a frozen LLM through a lightweight LoRA-based cognitive model to guide response generation. The network is incrementally updated after each interaction to track evolving user cognition. Extensive experiments on four benchmarks show that NeurPIU consistently improves long-term dialogue quality, especially its plug-and-play feature, and generalizes to conversational recommendation and mental health counseling. A user study with physiological measurements further indicates that NeurPIU reduces interaction time by 53.95% while improving user experience ratings by 22.38%. All data and code are released.

IPQA: A Benchmark for Core Intent Identification in Personalized Question Answering [Full]

Jieyong Kim (Yonsei University); Maryam Amirizaniani (University of Washington); Soojin Yoon (Yonsei University); Dongha Lee (Yonsei University)

Intent identification serves as the foundation for generating appropriate responses in personalized question answering (PQA). However, existing benchmarks evaluate only response quality or retrieval performance without directly measuring intent identification capabilities. This gap is critical because without understanding which intents users prioritize, systems cannot generate responses satisfying individual information needs. To address this, we introduce the concept of core intents: intents users prioritize when selecting answers to satisfy their information needs. To evaluate these core intents, we propose IPQA, a benchmark for core Intent identification in Personalized Question Answering. Since users do not explicitly state their prioritized intents, we derive core intents from observable behavior patterns in answer selection, grounded in satisficing theory where users choose answers meeting their acceptance thresholds. We construct a dataset with various domains through systematic filtering, LLM-based annotation, and rigorous quality control combining automated verification with human validation. Experimental evaluations across state-of-the-art language models reveal that current systems struggle with core intent identification in personalized contexts. Models fail to identify core intents from user histories, with performance degrading as question complexity increases.

Do Simulated Users Need to Remember? Analyzing the Impact of Memory Models in Conversational Search Evaluation [Full]

Nailia Mirzakhmedova (Bauhaus Universität Weimar); Marcel Gohsen (Bauhaus Universität Weimar); Johannes Kiesel (GESIS - Leibniz Institute for the Social Sciences); Matthias Hagen (Friedrich-Schiller Universität Jena); Benno Stein (Bauhaus Universität Weimar)

Conversational search systems are typically evaluated using a fixed reference collection of conversations or through user studies with a live system. However, fixed-reference conversations can cover only a few plausible conversations, and user studies are costly, time-consuming, and often hard to reproduce. A promising alternative that avoids coverage and cost issues is user simulation, in which a computer program takes on the role of a user and interacts with the system under evaluation. But the complexity of human search behavior raises the question of how "realistic" the simulations actually need to be for reliable evaluations of conversational search systems. In this paper, we ask: Do simulated users need to remember? While real users may learn and forget information during conversational search sessions, which inspired previous research to also model memory capabilities in simulations, it remains unclear whether this actually influences the results of system evaluations. To investigate the impact of memory modeling, we analyze conversations of simulated users and of humans with four conversational search systems. Our results suggest that incorporating long-term memory into simulators can help reproduce system effectiveness rankings obtained from human conversations, whereas incorporating short-term memory can diminish the reproduction. We also find that simulators are generally valid and reproducible---and memory modeling even increases run-to-run reproducibility of system rankings---but overall, simulations approximate human evaluation scores better when "helpful" assistants are evaluated than when assistants with deteriorated response quality are assessed. Our code and data are available at https://github.com/webis-de/SIGIR-26.

Dynamic Memory Forest: Constructing and Tracing Conversational Trajectories for Long-Term Conversation [Full]

Cai Ke (Harbin Institute of Technology); Bin Liang (The Chinese University of Hong Kong); Xin Liu (Pengcheng Laboratory); Yue Yu (Pengcheng Laboratory); Hui Wang (Pengcheng Laboratory); Ruifeng Xu (Harbin Institute of Technology)

While large language models (LLMs) have made significant progress in expanding their context windows, they still face great challenges in effectively organizing and utilizing long-term memory to maintain conversation consistency and coherence. Summarizing historical conversations has achieved remarkable performance, which, however, loses conversational trajectory and association, making it difficult to precisely combine memories from different sessions in response to current queries. To address this, we propose the Dynamic Memory Forest (DMF), a novel Consolidation-then-Growth framework for long-term open-domain conversation, which simulates the consolidation and growth processes of human memory by dynamically organizing long-term conversation histories into a memory forest of memory trees. To be specific, inspired by the principles of synaptic consolidation and plasticity from Cognitive Science, we first consolidate each session into memory units that preserve thematic coherence ("Consolidation"). Then, we first structure these units into memory trees and then grow the forest by dynamically connecting them through an evolutionary grafting mechanism, called Group Relative Voting Optimization, which mimics synaptic connection to decide whether a new memory tree should be grafted onto the existing forest or grow independently ("Growth"). For retrieval, we design an Entropy-Driven Memory Walk, constructing a logically coherent memory path via a navigation policy that prioritizes exploring high-entropy nodes. Experiments on three long-term conversation datasets show that our DMF significantly outperforms baselines in enhancing response generation for LLMs.

Towards a General Intelligent Information Agent: Framework, Models, and Evaluation [Perspective]

Chengxiang Zhai (University of Illinois at Urbana-Champaign)

Despite their common goal of assisting users in information access, current Information Retrieval (IR) systems, such as search engines and recommender systems, are studied and deployed separately across application contexts, resulting in scattered user information and fragmented support for a user's task. Can we develop a single general intelligent system to unify those specific systems for assisting users with information access using multiple modes (e.g., search, recommendation, and conversation)? In this perspective paper, we present a vision for developing a general intelligent information agent (GIANT) that not only unifies the current specific systems for supporting information access but also goes beyond information access to provide personalized task completion for users. We propose to formalize GIANT generally as an agent interacting with its users to minimize their effort on finishing a task as well as the agent's resource overhead. We propose a probabilistic modeling framework for optimizing GIANT's interactions with its users and discuss how to estimate its four component models, including Situation Model, Content Model, Task Model, and User Model. We discuss how the framework can be refined to derive specific interaction strategies and implemented with a general Markov Decision Process (MDP) architecture. We further discuss how to evaluate GIANT using user simulation and conclude with an outline of some promising directions for future research.

Information Seeking Behavior in LLM-Based RAG: Mental Models and Missing Information [Perspective]

Bruno Nadalic Sotic (University of Amsterdam); Jaap Kamps (University of Amsterdam)

Retrieval-augmented generation (RAG) systems are rapidly becoming a primary entry point to information, replacing ranked lists with synthesized, citation-backed answers. While recent work has focused on improving RAG architectures and measuring answer correctness, much less is known about how users conceptualize these systems and how such mental models reshape information seeking. In this Perspectives Paper, we argue that RAG interfaces compress the visible stages of the Information Search Process: exploration, comparison, and synthesis migrate inside the system, while the user sees only a conversational request-response loop. We ground this argument in a qualitative mixed-methods study of nine participants using an LLM-based "chat with documents" interface for exploratory, known-item, list-based, and absent-item tasks. Most participants did not describe the system in terms of selective retrieval and ranking. Instead, they imported an analyst-centric mental model, assuming exhaustive reading, persistent memory, and near-perfect recall. These assumptions systematically reconfigured behavior: participants engaged in minimal query reformulation, delegated sense-making to the system, rarely inspected sources, and treated "not found" responses as strong evidence of nonexistence. A minority of participants articulated IR-consistent models and exhibited more iterative, verification-oriented behavior. We use these findings to question implicit

behavioral assumptions in interactive IR and RAG evaluation and to outline design implications for handling missing information, retrieval visibility, and verification scaffolding. Our perspective is that understanding users' mental models of both system operation and their own role is central to assessing the epistemic consequences of generative search.

Cross-Modal Retrieval · *Eureka 3 · Wednesday 13:00–14:30*

Unsupervised 2D Image-Based 3D Model Retrieval via Decision Boundary Alignment and Graph Semantic Propagation [Full]

Nian Hu (Tianjin University of Technology); Yibo Zhao (Tianjin University of Technology); Xinhui Li (Tianjin University of Technology); Chen Li (Tianjin University of Technology); Cong Liu (Nova University of Lisbon); Zan Gao (Tianjin University of Technology)

Unsupervised 2D image-based 3D model retrieval (IBMR) aims to retrieve semantically relevant 3D shapes for a given 2D image query when 3D annotations are unavailable. This setting is challenging due to severe modality gaps, category-imbalanced mini-batches, inconsistent cross-domain decision boundaries, and mismatched semantic neighborhood structures. In this paper, we propose a unified framework that integrates Category-Aligned Sampling (CAS), Decision Boundary Alignment (DBA), and Graph Semantic Propagation (GSP) into a single optimization paradigm. CAS constructs category-consistent mini-batches to stabilize crossmodal learning. Built upon CAS, DBA leverages a masked Margin Disparity Discrepancy to regularize cross-domain class decision boundaries via an adversarial min-max objective, encouraging discriminative separation beyond marginal feature matching. To complement boundary-level regularization, GSP builds a crossdomain affinity graph over 2D and 3D samples and propagates supervision-induced relational structure through semantic message passing, explicitly preserving instance-level neighborhood consistency that is critical for retrieval. Extensive experiments on MI3DOR and MI3DOR-2 demonstrate consistent improvements over representative unsupervised IBMR baselines.

Selective Constraint Learning for Unsupervised Cross-Domain Image Retrieval [Full]

Wensi Fang (Jilin University); Xiaodan Zhang (Jilin University); Xiaoyu Lian (Jilin University); Qiang Li (Jilin University); Shuai Lü (Jilin University)

Unsupervised cross-domain image retrieval aims to retrieve semantically consistent images across domains with significant domain gaps, which poses substantial challenges under the absence of annotations in both domains. Existing approaches primarily rely on internally derived supervision signals for representation learning and cross-domain alignment. However, such internally induced supervision tends to impose an upper bound on achievable retrieval performance, as it lacks stable semantic references to support reliable category-level correspondence across domains. To address these limitations, we propose Selective Constraint Learning (SCL), a framework that introduces external semantic guidance as a stable prior for unsupervised cross-domain image retrieval. Leveraging a pre-trained foundation model, SCL constructs a dual-scope constraint bank to capture high-confidence positive and negative semantic relations within and across domains. Based on this, we design a generic constraint loss to jointly facilitate intra-domain compactness and inter-domain alignment. In addition, prototypical geometry regularization is designed to enhance in-domain structural stability through prototype-centered pull-and-push forces. Extensive experiments on multiple benchmarks demonstrate that SCL consistently outperforms state-of-the-art methods.

Struct-Align: Zero-Shot Text-to-3D Scene Retrieval via Locality-Aware Structural Alignment [Full]

Xiong Li (Zhejiang University of Technology); Yikang Yan (Zhejiang University of Technology); Zhenyu Wen (Zhejiang University of Technology); Jie Su (Zhejiang University of Technology); Qi Chen (Hangzhou Normal University); Zhen Hong (Zhejiang University of Technology)

Text-to-3D Scene Retrieval (T3SR) aims to retrieve 3D scenes that match users' linguistic queries, enabling intuitive access to 3D scene repositories. Existing approaches rely on joint embedding learning with large amounts of paired text–scene data, which is expensive to collect and often fails to generalize under open-vocabulary queries and diverse scene distributions. In this paper, we propose Struct-Align, a foundation-model-driven framework for zero-shot T3SR that eliminates the need for paired training data. Our key insight is to reformulate T3SR as a single-modality structural alignment problem by converting both 3D scenes and textual queries into a shared, schema-aligned textual representation compatible with pretrained text embedding models. To reliably derive such representations from complex 3D environments, we introduce a role-decomposed scene structuring pipeline that mitigates generative instability and produces semantically consistent scene depictions. To address the inherent semantic asymmetry between query and scene representations, we further propose a locality-aware structural matching strategy that explicitly localizes query intent and performs instance- and relation-level alignment within query-relevant substructures. Extensive experiments on multiple benchmarks demonstrate that Struct-Align outperforms both training-based and zero-shot baselines while exhibiting strong robustness to domain shift.

Unifying Granularity and Reliability: A Robust and Efficient Framework for Text-based Person Retrieval [Full]

Jingchen Hao (Xi'an Jiaotong University); Jiang Liu (Xi'an Jiaotong University); Zhen Peng (Xi'an Jiaotong University); Yuting Zhang (China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd); Zhongjiang He (China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd); Weizhan Zhang (Xi'an Jiaotong University); Hao Sun (China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd)

Text-based person retrieval (TPR) has become a crucial task in cross-modal retrieval due to its broad application in fields such as public safety and criminal investigation. Existing TPR methods typically rely on fully fine-tuning large-scale pretrained vision-language models like CLIP, which incurs high computational costs and tends to exhibit poor generalization in unseen domains due to overfitting. Fortunately, Parameter-Efficient Transfer Learning (PETL) has emerged as a lightweight alternative. However, applying PETL to TPR remains challenging, as its limited adaptation capacity struggles to capture intricate identity cues and becomes highly susceptible to gradient interference from unreliable image-text pairs. To address these challenges, we present a PETL-based framework named UniGR that unifies granularity and reliability for robust and efficient TPR. Specifically, we design a multi-granularity relational adapter (MRA) to capture both coarse-grained global and fine-grained local relational features among tokens, equipping the generic backbone with the task-specific, precise understanding needed for TPR. To combat the noise sensitivity of PETL, a reliability-aware reweighting strategy (RRS) is introduced to adaptively down-weight unreliable samples during training. Furthermore, we propose a parameter-free cross-modal cyclic verification (CMCV) module to mitigate ambiguities in cross-modal matching computations and refine retrieval ranking further. Experiments on benchmarks corroborate the superiority of UniGR among parameter-efficient methods. Remarkably, with only 4.5% of trainable parameters, UniGR outperforms most fully fine-tuned methods while maintaining strong generalization.

Pretrain-then-Adapt: Uncertainty-Aware Test-Time Adaptation for Text-based Person Search [Full]

Jiahao Zhang (University of Macau); Shaofei Huang (University of Macau); Yaxiong Wang (Hefei University of Technology); Zhedong Zheng (University of Macau)

Text-based person search faces inherent limitations due to data scarcity, driven by stringent privacy constraints and the high cost of manual annotation. To mitigate this, existing methods usually rely on a Pretrain-then-Finetune paradigm, where models are first pretrained on synthetic person-caption data to establish cross-modal alignment, followed by fine-tuning on labeled real-world datasets. However, this paradigm lacks practicality in real-world deployment scenarios, where large-scale annotated target-domain data is typically inaccessible. In this work, we propose a new Pretrain-then-Adapt paradigm that eliminates reliance on extensive target-domain supervision through an offline test-time adaptation manner, enabling dynamic model adaptation using only unlabeled test data with minimal post-train time cost. To mitigate overconfidence with false positives of previous entropy-based test-time adaptation, we propose an Uncertainty-Aware Test-Time Adaptation (UATTA) framework, which introduces a bidirectional retrieval disagreement mechanism to estimate uncertainty, i.e., low uncertainty is assigned when an image-text pair ranks highly in both image-to-text and text-to-image retrieval, indicating high alignment; otherwise, high uncertainty is detected. This indicator drives offline test-time model recalibration without labels, effectively mitigating domain shift. We validate UATTA on four benchmarks, i.e., CUHK-PEDES, ICFG-PEDES, RSTPReid, and PAB, showing consistent improvements across both CLIP-based (one-stage) and XVLM-based (two-stage) frameworks. Ablation studies confirm that UATTA outperforms existing offline test-time adaptation strategies, establishing a new benchmark for label-efficient, deployable person search systems. Our code is available at https://github.com/nkuzjh/UATTA.

StructAlign: Structured Cross-Modal Alignment for Continual Text-to-Video Retrieval [Full]

Shaokun Wang (Harbin Institute of Technology, Shenzhen); Weili Guan (Harbin Institute of Technology, Shenzhen); Jizhou Han (Xi'an Jiaotong University); Jianlong Wu (Harbin Institute of Technology, Shenzhen); Yupeng Hu (Shandong University); Liqiang Nie (Harbin Institute of Technology, Shenzhen)

Continual Text-to-Video Retrieval (CTVR) is a challenging multimodal continual learning setting, where models must incrementally learn new semantic categories while maintaining accurate text-video alignment for previously learned ones, thus making it particularly prone to catastrophic forgetting. A key challenge in CTVR is feature drift, which manifests in two forms: intra-modal feature drift caused by continual learning within each modality, and non-cooperative feature drift across modalities that leads to modality misalignment. To mitigate these issues, we propose StructAlign, a structured cross-modal alignment method for CTVR. First, StructAlign introduces a simplex Equiangular Tight Frame (ETF) geometry as a unified geometric prior to mitigate modality misalignment. Building upon this geometric prior, we design a cross-modal ETF alignment loss that aligns text and video features with category-level ETF prototypes, encouraging the learned representations to form an approximate simplex ETF geometry. In addition, to suppress intra-modal feature drift, we design a Cross-modal Relation Preserving loss, which leverages complementary modalities to preserve cross-modal similarity relations, providing stable relational supervision for feature updates. By jointly addressing non-cooperative feature drift across modalities and intra-modal feature drift, StructAlign effectively alleviates catastrophic forgetting in CTVR. Extensive experiments on benchmark datasets demonstrate that our method shows competitive advantages over state-of-the-art continual retrieval approaches.

Generative and LLM-based Retrieval · *Courtyard 1+2* ·

Wednesday 13:00–14:30

Is It Novel and Why? Fine-Grained Patent Novelty Prediction Based on Passage Retrieval [Full]

Valentin Knappich (Bosch Center for AI); Anna Hätty (Bosch Center for AI); Simon Razniewski (ScaDS.AI & TU Dresden); Annemarie Friedrich (University of Augsburg)

Novelty assessment is a critical yet complex task in the examination process for patent acceptance, requiring examiners to determine whether an invention is disclosed in a prior art document. The process involves intricate matching between specific features of a patent claim and passages in the prior art. While prior work has approached novelty prediction primarily as a binary classification task at the claim level, we argue that this formulation is susceptible to spurious correlations and lacks the granularity required for practical application. In this work, we introduce FiNE-Patents (Fine-grained Novelty Examination of Patents), a novel dataset comprising 3,658 first patent claims annotated with fine-grained, feature-level prior art references extracted from European Search Opinion (ESOP) documents. We propose shifting the evaluation paradigm from simple binary classification to a joint retrieval and abstract reasoning task at the feature level, requiring models to identify specific passages from a prior art document that disclose individual claim features, and to identify which features of a claim make it novel. We implement and evaluate LLM-based workflows that decompose claims into features, analyze each feature against prior art, and finally derive a claim-level novelty prediction. Our experiments demonstrate that these workflows outperform embedding-based baselines on passage retrieval and novel feature identification. Furthermore, we show that unlike trained classifiers, LLMs are robust against spurious correlations present in the claim-level novelty classification task. We release the dataset and code to foster further research into transparent and granular patent analysis.

Unifying Search and Recommendation in LLMs via Gradient Multi-Subspace Tuning [Full]

Jujia Zhao (Leiden University); Zihan Wang (CISPA Helmholtz Center for Information Security); Shuaiqun Pan (Leiden University); Suzan Verberne (Leiden University); Zhaochun Ren (Leiden University)

Search and recommendation (S&R) are two integral components of modern online platforms, both aiming to model and satisfy user information needs. This shared objective motivates a unified modeling paradigm that enables richer user modeling and improves the effectiveness of both tasks. Recent attempts to unify S&R formulate item ranking in both tasks as conditional generation. While this paradigm is promising, existing methods rely on full fine-tuning, which is computationally expensive and limits scalability. Parameter-efficient fine-tuning (PEFT) offers a more practical alternative but faces two critical challenges in unifying S&R: (1) gradient conflicts across tasks due to divergent optimization objectives, and (2) shifts in user intent understanding caused by overfitting to fine-tuning data, which distort general-domain knowledge and weaken LLM reasoning. To address these issues, we propose Gradient Multi-Subspace Tuning (GEMS), a novel framework that unifies S&R with LLMs while alleviating gradient conflicts and preserving general-domain knowledge. GEMS introduces (1) Multi-Subspace Decomposition, which disentangles shared and task-specific optimization signals into complementary low-rank subspaces, thereby reducing destructive gradient interference, and (2) Null-Space Projection, which constrains parameter updates to a subspace orthogonal to the general-domain knowledge space, mitigating shifts in user intent understanding. Extensive experiments on benchmark datasets show that GEMS consistently outperforms the state-of-the-art baselines across both search and recommendation tasks, and the gains remain consistent when scaling to billion-parameter LLMs.

Are LLM-Based Retrievers Worth Their Cost? An Empirical Study of Efficiency, Robustness, and Reasoning Overhead [Reproducibility]

Abdelrahman Abdallah (University of Innsbruck); Jamie Holdcroft (UNSW Sydney); Mohammed Ali (University of Innsbruck); Adam Jatowt (University of Innsbruck)

Large language model retrievers improve performance on complex queries, but their practical value depends on efficiency, robustness, and reliable confidence signals in addition to accuracy. We reproduce a reasoning-intensive retrieval benchmark (BRIGHT) across 12 tasks and 14 retrievers, and extend evaluation with cold-start indexing cost, query latency distributions and throughput, corpus scaling, robustness to controlled query perturbations, and confidence use (AUROC) for predicting query success. We also quantify reasoning overhead by comparing standard queries to five provided reasoning-augmented variants, measuring accuracy gains relative to added latency. We find that some reasoning-specialized retrievers achieve strong effectiveness while remaining competitive in throughput, whereas several large LLM-based bi-encoders incur substantial latency for modest gains. Reasoning augmentation incurs minimal latency for sub-1B encoders but exhibits diminishing returns for top retrievers and may reduce performance on formal math/code domains. Confidence calibration is consistently weak across model families, indicating that raw retrieval scores are unreliable for downstream routing without additional calibration. We release all code and artifacts for reproducibility.

A Parametric Memory Head for Continual Generative Retrieval [Full]

Kidist Amde Mekonnen (University of Amsterdam); Yubao Tang (University of Amsterdam); Maarten de Rijke (University of Amsterdam)

Generative information retrieval (GenIR) consolidates retrieval into a single neural model that decodes document identifiers (docids) directly from queries. While this model-as-index paradigm offers architectural simplicity, it is poorly suited to dynamic document collections. Unlike modular systems, where indexes are easily updated, GenIR's knowledge is parametrically encoded in its weights; consequently, standard adaptation methods such as full and parameter-efficient fine-tuning can induce catastrophic forgetting. We show that sequential adaptation improves retrieval on newly added documents but substantially degrades performance on earlier slices, exposing a pronounced stability--plasticity trade-off. To address this, we propose post-adaptation memory tuning (PAMT), a memory-only stabilization stage that augments an adapted model with a modular parametric memory head (PMH). PAMT freezes the backbone and attaches a product-key memory with fixed addressing. During prefix-trie constrained decoding, decoder hidden states sparsely query PMH to produce residual corrections in hidden space; these corrections are mapped to score adjustments via the frozen output embedding matrix, computed only over trie-valid tokens. This guides docid generation while keeping routing and backbone parameters fixed. To limit cross-slice interference, PAMT updates only a fixed budget of memory values selected using decoding-time access statistics, prioritizing entries frequently activated by the current slice and rarely used in prior sessions. Experiments on MS~MARCO and Natural Questions under sequential, disjoint corpus increments show that PAMT substantially improves retention on earlier slices with minimal impact on retrieval performance for newly added documents, while modifying only a sparse subset of memory values per session.

Lost in Decoding? Reproducing and Stress-Testing the Look-Ahead Prior in Generative Retrieval [Reproducibility]

Kidist Amde Mekonnen (University of Amsterdam); Yongkang Li (University of Amsterdam); Yubao Tang (University of Amsterdam); Simon Lupart (University of Amsterdam); Maarten de Rijke (University of Amsterdam)

Generative retrieval (GR) ranks documents by autoregressively generating document identifiers. Because many GR methods rely on trie-constrained beam search, they are vulnerable to early pruning of relevant prefixes under finite-beam decoding. Planning Ahead in Generative Retrieval (PAG) mitigates this failure mode by using simultaneous decoding to compute a document-level look-ahead prior that guides subsequent sequential decoding. We reproduce PAG at inference time and stress-test its decoding behavior. Using the authors' released checkpoint and identifier/trie artifacts under the reported decoding setup, we reproduce the main effectiveness results on MS~MARCO Dev and TREC-DL 2019/2020, and corroborate the reported beam-size--latency trade-off in our hardware setting. Beyond reproduction, we introduce plan drift diagnostics that quantify how intent-preserving query variations, including misspellings, reordering, synonym substitutions, paraphrases, and naturality shifts, alter the planner's top-n candidate set and highest-weight planner tokens, and how these changes affect guided decoding. We find that PAG's planning signal is brittle under lexical surface-form variation: intent-preserving typos can trigger plan collapse, where the planned candidate pool shifts enough that the look-ahead bonus provides little useful guidance, effectively reverting decoding toward weaker unguided search. We further evaluate fixed-index cross-lingual robustness using non-English mMARCO queries against an English index, and assess query-side mitigation strategies that require no re-indexing; query translation provides the strongest recovery in our setting. Overall, our results confirm PAG's reported effectiveness and the benefit of planning-guided decoding under the released inference setup, while showing that these gains depend on the stability of the planning signal under realistic query variation and query--document mismatch. Code available at https://github.com/kidist-amde/lost-in-decoding.

Beyond Chunk-Then-Embed: A Comprehensive Taxonomy and Evaluation of Document Chunking Strategies for Information Retrieval [Reproducibility]

Yongjie Zhou (The University of Queensland); Shuai Wang (The University of Queensland); Bevan Koopman (The University of Queensland); Guido Zuccon (The University of Queensland)

Document chunking is a critical preprocessing step in dense retrieval systems, yet the design space of chunking strategies remains poorly understood. Recent research has proposed several concurrent approaches, including LLM-guided methods (e.g., DenseX and LumberChunker) and contextualized strategies (e.g., Late Chunking), which generate embeddings before segmentation to preserve contextual information. However, these methods emerged independently and were evaluated on benchmarks with minimal overlap, making direct comparisons difficult. This paper reproduces prior studies in document chunking and presents a systematic framework that unifies existing strategies along two key dimensions: (1) segmentation methods, including structure-based methods (fixed-size, sentence-based, and paragraph-based) as well as semantically-informed and LLM-guided methods; and (2) embedding paradigms, which determine the timing of chunking relative to embedding (pre-embedding chunking vs. contextualized chunking). Our reproduction evaluates these approaches in two distinct retrieval settings established in previous work: in-document retrieval (needle-in-a-haystack) and in-corpus retrieval (the standard information retrieval task). Our comprehensive evaluation reveals that optimal chunking strategies are task-dependent: simple structure-based methods match or outperform LLM-guided alternatives for in-corpus retrieval, while LumberChunker performs best for in-document retrieval. Contextualized chunking improves in-corpus effectiveness but generally degrades

in-document retrieval. We also find that chunk size correlates moderately with in-document but weakly with in-corpus effectiveness, suggesting segmentation method differences are not purely driven by chunk size. Our code and evaluation benchmarks are publicly available at https://github.com/ielab/Reproduce-chunking-2026.

LLMs for Sequential Recommendation · *Goldfields* ·

Wednesday 16:00–17:30

Multimodal Large Language Models with Adaptive Preference Optimization for Sequential Recommendation [Full]

Yu Wang (Hefei University of Technology); Yonghui Yang (National University of Singapore); Le Wu (Hefei University of Technology); Yi Zhang (Anhui University); Fei Liu (Hefei University of Technology); Richang Hong (Hefei University of Technology)

Recent advances in Large Language Models (LLMs) have opened new avenues for sequential recommendation by enabling natural language reasoning over user behavior sequences. A common approach formulates recommendation as a language modeling task, where interaction histories are transformed into prompts and user preferences are learned via supervised fine-tuning. However, these methods operate solely in the textual modality and often miss users' fine-grained interests, especially when shaped by rich visual signals such as product images or movie posters. Multimodal Large Language Models (MLLMs) offer a promising alternative by aligning text and vision in a shared semantic space. A prevalent training paradigm applies Supervised Fine-Tuning (SFT) followed by Direct Preference Optimization (DPO) to model user preferences. Yet, two core challenges remain: 1) Imbalanced sample hardness, where random negative sampling causes overfitting on easy examples and under-training on hard ones; 2) Cross-modal semantic bias, where the fixed reference model in DPO prevents the policy model from correcting modality misalignments—especially over long sequences. To address these issues, we propose a Multimodal LLM framework that integrates Hardness-aware and Noise-regularized preference optimization for Recommendation (HaNoRec). Specifically, HaNoRec dynamically adjusts optimization weights based on both the estimated hardness of each training sample and the policy model's real-time responsiveness, prioritizing harder examples. It further introduces Gaussian-perturbed distribution optimization on output logits to enhance cross-modal semantic consistency and reduce modality bias inherited from the reference model. Experiments on three benchmarks demonstrate its superiority over state-of-the-art methods.

Think, But Don't Tell: Implicit Reasoning for LLM-based Sequential Recommendation via Multi-Teacher Distillation [Full]

Weihai Lu (Peking University); Xiaoxi Cui (Takeway.AI); Chenke Yin (Xi'an Jiaotong-Livepool University)

Large Language Models (LLMs) demonstrate significant potential in sequential recommendation, and leveraging their Chain-of-Thought (CoT) reasoning capabilities can further unlock profound user preference understanding. However, deploying explicit CoT reasoning in real-world systems faces prohibitive challenges: (i) the conflict between the large model scale required for high-fidelity reasoning and the resource constraints of online services, and (ii) the excessive latency introduced by auto-regressive rationale generation. To address these issues, we propose I Reasoning via Multi-Teacher Distillation (IRMD), a novel framework that 'compiles' the reasoning abilities of large teacher LLMs into a lightweight student Small Language Model (SLM). IRMD first employs a Multi-Teacher CoT Synthesis with Dual-Constraint Rejection Sampling module to generate a high-quality, diverse set of reasoning paths. Subsequently, our Annealing-Scheduled Reasoning Distillation strategy progressively trains the student to internalize this logic, transitioning from mimicking explicit CoT to performing purely implicit reasoning. Extensive experiments on multiple benchmark datasets demonstrate that IRMD significantly outperforms state-of-the-art baselines in both recommendation accuracy and inference efficiency. Our code is accessible at https://github.com/Cxx-0/IRMD.

SpecTran: Spectral-Aware Transformer-based Adapter for LLM-Enhanced Sequential Recommendation [Full]

Yu Cui (Zhejiang University); Feng Liu (OPPO Research Institute); Zhaoxiang Wang (OPPO Research Institute); Changwang Zhang (OPPO Research Institute); Jun Wang (OPPO Research Institute); Can Wang (Zhejiang University); Jiawei Chen (Zhejiang University)

Traditional sequential recommendation (SR) models learn low-dimensional item ID embeddings from user-item interactions, often overlooking textual information such as item titles or descriptions. Recent advances in Large Language Models (LLMs) have inspired a surge of research that encodes item textual information with high-dimensional semantic embeddings, and designs transformation methods to inject such embeddings into SR models. These embedding transformation strategies can be categorized into two types, both of which exhibits notable drawbacks: 1) adapter-based methods suffer from pronounced dimension collapse, concentrating information into a few dominant dimensions; 2) SVD-based methods are rigid and manual, considering only a few principal spectral components while discarding rich information in the remaining spectrum. To address these limitations, we propose SpecTran, a spectral-aware transformer-based adapter that operates in the spectral domain, attending to the full spectrum to select and aggregates informative components. A learnable spectral-position encoding injects singular-value cues as an inductive bias, guiding attention toward salient spectral components and promoting diversity across embedding dimensions. Across four real-world datasets and three SR backbones, it consistently outperforms strong baselines, achieving an average improvement of 9.17%.

Fusion and Alignment Enhancement with Large Language Models for Tail-item Sequential Recommendation [Full]

Zhifu Wei (Northeastern University); Yizhou Dang (Northeastern University); Guibing Guo (Northeastern University); Chuang Zhao (The Hong Kong University of Science and Technology); Zhu Sun (Singapore University of Technology and Design)

Sequential Recommendation (SR) learns user preferences from their historical interaction sequences and provides personalized suggestions. In real-world scenarios, most items exhibit sparse interactions, known as the tail-item problem. This issue limits the model's ability to accurately capture item transition patterns. To tackle this, large language models (LLMs) offer a promising solution by capturing semantic relationships between items. Despite previous efforts to leverage LLM-derived embeddings for enriching tail items, they still face the following limitations: 1) They struggle to effectively fuse collaborative signals with semantic knowledge, leading to suboptimal item embedding quality. 2) Existing methods overlook the structural inconsistency between the ID and LLM embedding spaces, causing conflicting signals that degrade recommendation accuracy. In this work, we propose a \textbf{F}usion and \textbf{A}lignment \textbf{E}nhancement framework with LLMs for Tail-item Sequential \textbf{Rec}ommendation (\textbf{FAERec}), which improves item representations by generating coherently-fused and structurally consistent embeddings. For the information fusion challenge, we design an adaptive gating mechanism that dynamically fuses ID and LLM embeddings. Then, we propose a dual-level alignment approach to mitigate structural inconsistency. The item-level alignment establishes correspondences between ID and LLM embeddings of the same item through contrastive learning, while the feature-level alignment constrains the correlation patterns between corresponding dimensions across the two embedding spaces. Furthermore, the weights of the two alignments are adjusted by a curriculum learning scheduler to avoid premature optimization of the complex feature-level objective. Extensive experiments across three widely used datasets with multiple representative SR backbones demonstrate the effectiveness and generalizability of our framework. The codes are provided at \url{https://github.com/ZhifuWei/FAERec}.

SPRINT: Scalable and Predictive Intent Refinement for LLM-Enhanced Session-based Recommendation [Full]

Gyuseok Lee (University of Illinois at Urbana-Champaign); Wonbin Kweon (University of Illinois Urbana-Champaign); Zhenrui Yue (University of Illinois Urbana-Champaign); Yaokun Liu (University of Illinois at Urbana-Champaign); Yifan Liu (University of Illinois at Urbana-Champaign); Susik Yoon (Korea University); Dong Wang (University of Illinois at Urbana-Champaign); SeongKu Kang (Korea University)

Large language models (LLMs) have enhanced conventional recommendation models via user profiling, which generates representative textual profiles from users' historical interactions. However, their direct application to session-based recommendation (SBR) remains challenging due to severe session context scarcity and poor scalability. In this paper, we propose SPRINT, a scalable SBR framework that incorporates reliable and informative intents while ensuring high efficiency in both training and inference. SPRINT constrains LLM-based profiling with a global intent pool and validates inferred intents based on recommendation performance to mitigate noise and hallucinations under limited context. To ensure scalability, LLMs are selectively invoked only for uncertain sessions during training, while a lightweight intent predictor generalizes intent prediction to all sessions without LLM dependency at inference time. Experiments on real-world datasets show that SPRINT consistently outperforms state-of-the-art methods while providing more explainable recommendations.

An Action-Aware Generative Sequence Modeling for Short Video Recommendation [Full]

Wenhao Li (Kuaishou Technology); Zihan Lin (Kuaishou Technology); ZhengXiao Guo (Kuaishou Technology); Jie Zhou (Kuaishou Technology); Shukai Liu (Unaffiliated); Yongqi Liu (Kuaishou Technology); Chuan Luo (Beihang University)

With the rapid development of the Internet, users have increasingly higher expectations for the recommendation accuracy of online content consumption platforms (e.g., short video platforms). However, short videos often contain diverse segments, and users may not hold the same attitude toward all of them (e.g., music enthusiasts may not enjoy all songs in a medley). Traditional binary-classification recommendation models, which treat a video as a single holistic entity, face limitations in accurately capturing such nuanced preferences. Considering that user consumption is a temporal process, this paper demonstrates that the timing of user actions can represent diverse intentions through statistical analysis and examination of action patterns. Based on this insight, we propose a novel modeling paradigm: Action-Aware Gen erative Sequence Network (A2Gen), which refines user actions (e.g., Like and Follow, etc.) along the temporal dimension and chains them into sequences for unified processing and prediction. First, we introduce the Context-aware Attention Module (CAM) to model action sequences enriched with item-specific contextual features. Building upon this, we develop the Hierarchical Sequence Encoder (HSE) to learn temporal action patterns from users' historical actions. Finally, through leveraging CAM, we design a module for action sequence generation: the Action-seq Autoregressive Generator (AAG). Extensive offline experiments on the Kuaishou's dataset and the Tmall public dataset demonstrate the superiority of our proposed model. Furthermore, through large-scale online

A/B testing deployed on Kuaishou's platform, our model achieves significant improvements over baseline methods in multi-task prediction by leveraging sequential information. Specifically, it yields increases of 0.34% in user watch time, 8.1% in interaction rate, and 0.162% in overall user retention (LifeTime-7), leading to successful deployment across all traffic, serving over 400 million users every day.

Agentic Search · *Room 110 · Wednesday 16:00–17:30*

Towards Knowledgeable Deep Research: Framework and Benchmark [Full]

Wenxuan Liu (State Key Laboratory of AI Safety); Zixuan Li (State Key Laboratory of AI Safety); Long Bai (State Key Laboratory of AI Safety); Chunmao Zhang (State Key Laboratory of AI Safety); Fenghui Zhang (State Key Laboratory of AI Safety); Zhuo Chen (State Key Laboratory of AI Safety); Wei Li (State Key Laboratory of AI Safety); Yuxin Zuo (State Key Laboratory of AI Safety); Fei Wang (State Key Laboratory of AI Safety); Bingbing Xu (State Key Laboratory of AI Safety); Xuhui Jiang (State Key Laboratory of AI Safety); Jin Zhang (State Key Laboratory of AI Safety); Xiaolong Jin (State Key Laboratory of AI Safety); Jiafeng Guo (State Key Laboratory of AI Safety); Tat-Seng Chua (National University of Singapore); Xueqi Cheng (State Key Laboratory of AI Safety)

Deep Research (DR) requires LLM agents to autonomously perform multi-step information seeking, processing, and reasoning to generate comprehensive reports. In contrast to existing studies that mainly focus on unstructured web content, a more challenging DR task should additionally utilize structured knowledge to provide a solid data foundation, facilitate quantitative computation, and lead to in-depth analyses.In this paper, we refer to this novel task as Knowledgeable Deep Research (KDR), which requires DR agents to generate reports with both structured and unstructured knowledge.Furthermore, we propose the Hybrid Knowledge Analysis framework (HKA), a multi-agent architecture that reasons over both kinds of knowledge and integrates the texts, figures, and tables into coherent multimodal reports.The key design is the Structured Knowledge Analyzer, which utilizes both coding and vision-language models to produce figures, tables, and corresponding insights.To support systematic evaluation, we construct KDR-Bench, which covers 9 domains, includes 41 expert-level questions, and incorporates a large number of structured knowledge resources (e.g., 1,252 tables). We further annotate the main conclusions and key points for each question and propose three categories of evaluation metrics including general-purpose, knowledge-centric, and vision-enhanced ones.Experimental results demonstrate that HKA consistently outperforms most existing DR agents on general-purpose and knowledge-centric metrics, and even surpasses the Gemini DR agent on vision-enhanced metrics, highlighting its effectiveness in deep, structure-aware knowledge analysis.Finally, we hope this work can serve as a new foundation for structured knowledge analysis in DR agents and facilitate future multimodal DR studies.

Deep Search with Hierarchical Meta-Cognitive Monitoring Inspired by Cognitive Neuroscience [Full]

Zhongxiang Sun (Renmin University of China); Qipeng Wang (Renmin University of China); Weijie Yu (University of International Business and Economics); Jingxuan Yang (Tencent); Haolang Lu (Beijing University of Posts and Telecommunications); Jun Xu (Renmin University of China)

Deep search agents powered by large language models have demonstrated strong capabilities in multi-step retrieval, reasoning, and long-horizon task execution. However, their practical failures often stem from the lack of mechanisms to monitor and regulate reasoning and retrieval states as tasks evolve. Insights from cognitive neuroscience suggest that human metacognition is hierarchically organized, integrating fast anomaly detection with selectively triggered, experience-driven reflection. In this work, we propose Deep Search with Meta-Cognitive Monitoring (DS-MCM), a deep search framework augmented with an explicit hierarchical metacognitive monitoring mechanism. DS-MCM integrates a Fast Consistency Monitor, which performs lightweight checks on the alignment between external evidence and internal reasoning confidence, and a Slow Experience-Driven Monitor, which is selectively activated to guide corrective intervention based on experience memory from historical agent trajectories. By embedding monitoring directly into the reasoning–retrieval loop, DS-MCM determines both when intervention is warranted and how corrective actions should be informed by prior experience. Experiments across multiple deep search benchmarks and backbone models demonstrate that DS-MCM consistently improves performance and robustness.

DLLM-Searcher: Adapting Diffusion Language Model for Efficient Search Agents [Full]

Jiahao Zhao (Renmin University of China); Shaoxuan Xu (Renmin University of China); Zhongxiang Sun (Renmin University of China); Fengqi Zhu (Renmin University of China); Jingyang Ou (Renmin University of China); Yuling Shi (Shanghai Jiao Tong University); Chongxuan Li (Renmin University of China); Xiao Zhang (Renmin University of China); Jun Xu (Renmin University of China)

Recently, Diffusion Large Language Models (dLLMs) have demonstrated unique efficiency advantages, enabled by their inherently parallel decoding mechanism and flexible generation paradigm. Meanwhile, despite the rapid advancement of Search Agents, their practical deployment is constrained by a fundamental limitation, termed as 1) Latency Challenge : the serial execution of multi-round reasoning, tool calling, and tool response waiting under the ReAct agent paradigm induces severe end-to-end latency. Intuitively, dLLMs can leverage their distinctive strengths to optimize the operational efficiency of agents under the ReAct agent paradigm. Practically, existing dLLM backbones face the 2) Agent Ability Challenge. That is, existing dLLMs exhibit remarkably weak reasoning and tool-calling capabilities, preventing these advantages from being effectively realized in practice. In this paper, we propose DLLM-Searcher, an optimization framework for dLLM-based Search Agents. To solve the Agent Ability Challenge, we design a two-stage post-training pipeline encompassing Agentic Supervised Fine-Tuning (Agentic SFT) and Agentic Variance-Reduced Preference Optimization (Agentic VRPO), which enhances the backbone dLLM's information seeking and reasoning capabilities. To mitigate the Latency Challenge, we leverage the flexible generation mechanism of dLLMs and propose a novel agent paradigm termed Parallel-Reasoning and Acting (P-ReAct). P-ReAct guides the model to prioritize decoding tool_call instructions, thereby allowing the model to keep thinking while waiting for the tool's return. Experimental results demonstrate that DLLM-Searcher achieves performance comparable to mainstream LLM-based search agents and P-ReAct delivers approximately 15% inference acceleration. Our code is available at https://github.com/bubble65/DLLM-Searcher

Learning to Retrieve from Agent Trajectories [Full]

Yuqi Zhou (Renmin University of China); Sunhao Dai (Renmin University of China); Changle Qu (Renmin University of China); Liang Pang (Institute of Computing Technology, Chinese Academy of Sciences); Jun Xu (Renmin University of China); Ji-Rong Wen (Renmin University of China)

Information retrieval (IR) systems have traditionally been designed and trained for human users, with learning-to-rank methods relying heavily on large-scale human interaction logs such as clicks and dwell time. With the rapid emergence of large language model (LLM) powered search agents, however, retrieval is increasingly consumed by agents rather than human beings, and is embedded as a core component within multi-turn reasoning and action loops. In this setting, retrieval models trained under human-centric assumptions can be mismatched with the way agents issue intermediate queries and consume results. In this work, we argue that retrieval models for agentic search should be trained directly from agent interaction data. We study learning to retrieve from agent trajectories as a trajectory-supervised training setting, where supervision is derived from multi-step agent interactions. Through a systematic analysis of search agent trajectories, we identify key behavioral signals that reveal document utility, including browsing actions, unbrowsed rejections, and post-browse reasoning traces. Guided by these insights, we propose LRAT, a simple yet effective framework that mines high-quality retrieval supervision from agent trajectories and incorporates relevance intensity through weighted optimization. To instantiate this setting at scale, we deploy the Tongyi-DeepResearch-30B model on 10K InfoSeekQA queries with four retrievers, collecting 26,482 agent trajectories and constructing 91,713 training pairs. Extensive experiments on both in-domain and out-of-domain deep research benchmarks demonstrate that retrievers trained with LRAT consistently improve evidence recall, end-to-end task success, and execution efficiency across diverse agent architectures and scales. Our results highlight agent trajectories as a practical and scalable supervision source for retrieval in agentic search.

SmartSearch: Process Reward-Guided Query Refinement for Search Agents [Full]

Tongyu Wen (Renmin University of China); Guanting Dong (Renmin University of China); Zhicheng Dou (Renmin University of China)

Large Language Model (LLM)-based search agents have proven promising for addressing knowledge-intensive problems by incorporating information retrieval capabilities. Existing works largely focus on optimizing the reasoning paradigms of search agents, yet the quality of intermediate search queries during reasoning remains overlooked. As a result, the generated queries often remain inaccurate, leading to unexpected retrieval results and ultimately limiting search agents' overall effectiveness. To mitigate this issue, we introduce SmartSearch, a framework built upon two key mechanisms: (1) Process Rewards, which provide fine-grained supervision for the quality of each intermediate search query through Dual-Level Credit Assessment. (2) Query Refinement, which promotes the optimization of query generation by selectively refining low-quality search queries and regenerating subsequent search rounds based on these refinements. To enable the search agent to progressively internalize the ability to improve query quality under the guidance of process rewards, we design a three-stage curriculum learning framework. This framework guides the agent through a progression from imitation, to alignment, and ultimately to generalization. Experimental results show that SmartSearch consistently surpasses existing baselines, and additional quantitative analyses further confirm its significant gains in both search efficiency and query quality. The code is available at https://github.com/RUC-NLPIR/SmartSearch.

HiRA: Decoupling Planning and Execution with Hierarchical Reasoning in Deep Search [Full]

Jiajie Jin (Gaoling School of Artificial Intelligence, Renmin University of China); Xiaoxi Li (Gaoling School of Artificial Intelligence, Renmin University of China); Yuyao Zhang (Gaoling School of Artificial Intelligence, Renmin University of China); Guanting Dong (Gaoling School of Artificial Intelligence, Renmin University of China); Zhao Yang (Gaoling School of Artificial Intelligence, Renmin University of China); Yutao Zhu (Gaoling School of Artificial Intelligence, Renmin University of China); Zhicheng Dou (Gaoling School of Artificial Intelligence, Renmin University of China)

Complex information needs in real-world search scenarios demand deep reasoning and knowledge synthesis across diverse sources, which traditional

retrieval-augmented generation (RAG) pipelines struggle to address effectively. Current reasoning-based approaches face a key architectural challenge: they employ a single model to handle both high-level planning and detailed execution, resulting in inefficient reasoning and limited scalability. In this paper, we introduce HiRA, a hierarchical framework that separates strategic planning from specialized execution. Our approach decomposes complex search tasks into multiple subtasks, assigns each subtask to a domain-specific agent equipped with external tools and reasoning capabilities, and coordinates the results through a structured integration mechanism. This separation prevents execution details from disrupting high-level reasoning while enabling the system to leverage specialized expertise for different types of information processing. Experiments on four complex, cross-modal deep search benchmarks show that HiRA significantly outperforms state-of-the-art RAG and agent-based systems, highlighting the effectiveness of decoupled planning and execution for multi-step information seeking tasks. The code is available at https://github.com/RUC-NLPIR/HiRA.

Computational Advertising and Large-Scale Serving

· Eureka 1 · Wednesday 16:00–17:30

Generative Bid Shading in Real-Time Bidding Advertising [Full]

Yinqiu Huang (Meituan); Hao Ma (Chongqing University); Wenshuai Chen (Meituan); Zongwei Wang (Chongqing University); Shuli Wang (Meituan); Yongqiang Zhang (Meituan); Xue Wei (Meituan); Yinhua Zhu (Meituan); Haitao Wang (Meituan); Xingxing Wang (Meituan)

Bid shading plays a crucial role in Real-Time Bidding (RTB) by adaptively adjusting the bid to avoid advertisers overspending. Existing mainstream two-stage methods, which first model bid landscapes and then optimize surplus using operations research techniques, are constrained by unimodal assumptions that fail to adapt for non-convex surplus curves and are vulnerable to cascading errors in sequential workflows. Additionally, existing discretization models of continuous values ignore the dependence between discrete intervals, reducing the model's error correction ability, while sample selection bias in bidding scenarios presents further challenges for prediction. To address these issues, this paper introduces Generative Bid Shading (GBS), which comprises two primary components: 1) an end-to-end generative model that utilizes an autoregressive approach to generate shading ratios by stepwise residuals, capturing complex value dependencies without relying on predefined priors; and 2) a reward preference alignment system, which incorporates a channel-aware hierarchical dynamic network (CHNet) as the reward model to extract fine-grained features, along with modules for surplus optimization and exploration utility reward alignment, ultimately optimizing both short-term and long-term surplus using group relative policy optimization (GRPO). Extensive experiments on both offline and online A/B tests validate GBS's effectiveness. Moreover, GBS has been deployed on the Meituan DSP platform, serving billions of bid requests daily.

Generative Auto-Bidding with Unified Modeling and Exploration [Full]

Mingming Zhang (Wuhan University); Feiqing Zhuang (Taobao & Tmall Group of Alibaba); Na Li (Wuhan University); Shengjie Sun (Taobao & Tmall Group of Alibaba); Xiaowei Chen (Taobao & Tmall Group of Alibaba); Junxiong Zhu (Alibaba Group); Fei Xiao (Taobao & Tmall Group of Alibaba); Keping Yang (Taobao & Tmall Group of Alibaba); Lixin Zou (Wuhan University); Chenliang Li (Wuhan University)

Automated bidding has become a core component of modern digital advertising. Early methods were primarily rule-based, while easy to implement, they struggled to adapt to rapidly changing environments. Subsequent Reinforcement Learning methods modeled bidding as a Markov Decision Process but were limited in their ability to capture long-term dependencies. While recent generative models have shown encouraging progress, they generally lack explicit mechanisms to balance exploration and safety. They often rely solely on simple action perturbations or trajectory guidance to foster bidding exploration, and critically, they lack a safety fallback mechanism. This limitation leads to inefficient exploration and significantly increases the financial risk for advertising platforms. To bridge this gap, we propose a new framework named Generative Auto-Bidding with Uni fied Mod eling and Exploration (\baby), which synergistically integrates directed exploration with a safe fallback mechanism. \baby utilizes a Decision Transformer (DT) to jointly model historical bidding actions and environmental state transitions. A Q-value module guides the DT's exploration through regularization constraints. Concurrently, an Inverse Dynamics Module (IDM) leverages the future states predicted by the DT to infer robust and behaviorally consistent actions, thereby providing a safe policy fallback. The Q-value module then adaptively selects the final action from these two options, balancing exploration and safety. Together, these three components form an integrated "explore--safeguard--select" pipeline, unifying efficiency and safety. We conduct comprehensive experiments on public datasets, in simulated auction environments, and through a large-scale online deployment on Taobao, a leading advertising platform in China. The results demonstrate that \baby consistently outperforms state-of-the-art (SOTA) baseline methods across all scenarios. In real-world online deployment, \baby achieves remarkable improvements: +4.10% in ad GMV, +1.40% in ad clicks, +1.66% in ad cost, and +3.52% in ad ROI, demonstrating its effectiveness and strong industrial applicability.

TAR: Generative Auto-Bidding and Budget Pacing via Multi-Scale Trajectory Modeling [Full]

Liang Shi (Peking University); Longxiang Xu (Taobao & Tmall Group of Alibaba); Zhengju Tang (Peking University); Yundu Huang (Taobao & Tmall Group of Alibaba); Jian Xu (Taobao & Tmall Group of Alibaba); Zhi Yang (Peking University)

Auto-bidding and budget pacing are formulated as sequential decision-making tasks. While flexible, such a framework faces a fundamental granularity mismatch: decisions are made at a fine temporal scale, while their performance feedback is fully and reliably observable at a much coarser resolution. This manifests as sparse reward signals and delayed feedback, forcing agents to learn from locally noisy and incomplete signals. We address this core challenge by introducing the \underline{T}rajectory \underline{A}uto-\underline{R}egressive Model (\textbf{TAR}), a generative framework that aligns planning resolution with feedback dynamics. Motivated by the insight that coarser temporal aggregation yields denser rewards and less scattered feedback, TAR generates trajectories in a coarse-to-fine manner. It incorporates three key innovations: (1) progressive trajectory generation across multiple temporal scales; (2) latent-space compression via a multi-scale VQVAE to handle heterogeneous feature types; and (3) state-action integration that captures long-term dependencies without auxiliary inverse models. Comprehensive experiments in both sparse-reward and delayed-feedback settings demonstrate that TAR consistently outperforms strong baselines in offline simulations and online deployment, validating its effectiveness in overcoming the granularity mismatch for more stable and robust advertising optimization.

Optimizing Marketing Subsidies via Counterfactual Learning with Asymmetric Reward Function [Full]

Xiang Li (Peking University); Yanghao Xiao (University of Chinese Academy of Sciences); Chunyuan Zheng (Peking University); Qian Zou (Meituan); Cheng Bing (Meituan); Wei Lin (Meituan); Haoxuan Li (Peking University); Zhouchen Lin (Peking University)

In marketing, optimizing subsidy allocation to maximize overall profits is of substantial economic importance. Prior research has employed treatment effect estimation techniques to identify subsidy-sensitive items and design corresponding allocation strategies. However, more accurate treatment effect estimations do not necessarily lead to better allocations, underscoring the critical influence of decision boundaries in decision-making. This paper argues that optimal allocation fundamentally depends on predicting the expected optimal subsidy, a challenge distinct from conventional treatment effect estimation or causal decision-making, which existing approaches fail to address. To fill this gap, we introduce a two-stage Counterfactual optimal subsidy Learning method with an Asymmetric reward (CoLA). In the first stage, we derive a coarse estimate of the expected subsidy threshold by exploiting order information and the conditional independence between expected and observed subsidies. In the second stage, we refine these estimates using an asymmetric loss function, leading to more robust predictions. Under practical budget constraints, we prioritize candidates based on their Sharpe ratios to determine the final subsidy allocation strategy. Experiments on three public datasets and an online A/B test show that our method achieves significant performance improvements, yielding the highest total profit and incremental leverage ratios.

Balanced Co-Clustering of Users and Items for Embedding Table Compression in Recommender Systems [Full]

Runhao Jiang (Hong Kong Baptist University); Renchi Yang (Hong Kong Baptist University); Donghao Wu (The Chinese University of Hong Kong)

Recommender systems have advanced markedly over the past decade by transforming each user/item into a dense embedding vector with deep learning models. At industrial scale, embedding tables constituted by such vectors of all users/items demand a vast amount of parameters and impose heavy compute and memory overhead during training and inference, hindering model deployment under resource constraints. Existing solutions towards embedding compression either suffer from severely compromised recommendation accuracy or incur considerable computational costs. To mitigate these issues, this paper presents \algo, a fast and effective framework for compressing embedding tables. Unlike traditional ID hashing, \algo is built on the idea of exploiting collaborative signals in user-item interactions for user and item groupings, such that similar users/items share the same embeddings in the codebook. Specifically, we formulate a balanced co-clustering objective that maximizes intra-cluster connectivity while enforcing cluster-volume balance, and unify canonical graph clustering techniques into the framework through rigorous theoretical analyses. To produce effective groupings while averting codebook collapse, \algo instantiates this framework with a principled weighting scheme for users and items, an efficient label propagation solver, as well as secondary user clusters. Our extensive experiments comparing \algo against full models and 18 baselines over benchmark datasets demonstrate that \algo cuts embedding parameters by over 75% with a drop of at most 1.85% in recall, while surpassing the strongest baselines by being up to 346× faster.

SilverTorch: A Unified Model-based System to Democratize Large-Scale Recommendation on GPUs [Full]

Bi Xue (Meta Platforms, Inc); Hong Wu (Meta Platforms, Inc); Lei Chen (Meta Platforms, Inc); Chao Yang (Meta Platforms, Inc); Yiming Ma (Meta Platforms, Inc); Fei Ding (Meta Platforms, Inc); Zhen Wang (Meta Platforms, Inc); Liang Wang (Meta Platforms, Inc); Xiaoheng Mao (Meta Platforms, Inc); Ke Huang (Meta Platforms, Inc); Xialu Li (Meta Platforms, Inc.); Peng Xia (Meta Platforms, Inc.); Rui Jian (Meta Platforms, Inc.); Yanli Zhao (Meta Platform, Inc.); Yanzun Huang (Meta Platform, Inc.); Yijie Deng (Meta Platform, Inc.); Harry Tran (Meta Platform, Inc.); Ryan Chang (Meta Platform, Inc.); Min Yu (Meta Platform, Inc.); Eric Dong (Meta Platform, Inc.); Jiazhou Wang (Meta Platform, Inc.); Qianqian Zhang (Meta Platform, Inc.); Keke Zhai (Meta Platform, Inc.); Hongzhang Yin (Meta Platform, Inc.); Pawel Garbacki

(Meta Platform, Inc.); Zheng Fang (Meta Platform, Inc.); Yiyi Pan (Meta Platform, Inc.); Min Ni (Meta Platform, Inc.); Yang Liu (Meta Platform, Inc.)
Serving deep learning based recommendation models (DLRM) at scale is challenging. Existing approaches rely on dedicated ANN indexing and filtering services on CPUs, suffering from non-negligible costs and missing co-design opportunities. Such inefficiency makes them difficult to support complex model architectures, such as learned similarities and multi-task retrieval. In this paper, we present SilverTorch, a model-based serving system that brings all components into one unified model. It unifies model serving by replacing standalone indexing and filtering services with model layers. We propose a model-based GPU Bloom index for feature filtering and a fused Int8 ANN kernel for nearest neighbor search. Through co-design of the ANN search and feature filtering, we reduce GPU memory usage and eliminate computation. Benefiting from this design, we scale up retrieval by introducing an OverArch scoring layer and a multi-task retrieval with a Value Model to aggregate scores. These advancements improve the retrieval accuracy and enable future studies for serving more complex models. Our evaluation on industry-scale datasets shows that SilverTorch achieves up to 23.7× higher throughput compared to the state-of-the-art approaches. We also demonstrate that SilverTorch's solution is 13.35× more cost-efficient than CPU-based solution while improving accuracy via serving more complex models.

Adaptive Retrieval for Reasoning · *Eureka 2* ·

Wednesday 16:00–17:30

Hierarchical Embedding Fusion for Retrieval-Augmented Code Generation [Full]

Nikita Sorokin (MWS AI); Ivan Sedykh (MWS AI); Valentin Malykh (MWS AI)
Retrieval-augmented code generation commonly conditions a decoder on large retrieved snippets, which couples online cost to repository size and introduces long-context noise. We present Hierarchical Embedding Fusion (HEF), a two-stage repository representation for code completion: (i) an offline cache that compresses repository chunks into a reusable hierarchy of dense vectors using a small fuser model, and (ii) an online interface that maps a small number of retrieved vectors into learned pseudo-tokens consumed by a code generator. This replaces thousands of retrieved tokens with a fixed pseudo-token budget while retaining access to repository-level information. On RepoBench and RepoEval, HEF with a 1.8B pipeline attains comparable exact-match accuracy to snippet-based retrieval baselines while operating at sub-second median latency on a single A100. Compared to graph- and iterative-retrieval systems in our setup, HEF reduces median end-to-end latency by 13×–26×. We additionally introduce a utility-weighted likelihood signal to filter training contexts and report ablations over pseudo-token budget, embedding models, and robustness to harmful retrieval. These results support hierarchical dense caching as an effective mechanism for low-latency repository-aware code completion.

A Reproducibility Study of Metacognitive Retrieval-Augmented Generation [Reproducibility]

Gabriel Iturra Bocaz (University of Stavanger); Petra Galuš■áková (University of Stavanger)
Recently, Retrieval Augmented Generation (RAG) has shifted focus to multi-retrieval approaches to tackle complex tasks such as multi-hop question answering. However, these systems struggle to decide when to stop searching once enough information has been gathered. To address this, \citet{zhou2024metacognitive} introduced Metacognitive Retrieval Augmented Generation (MetaRAG), a framework inspired by metacognition that enables Large Language Models to critique and refine their reasoning. In this reproducibility paper, we reproduce MetaRAG following its original experimental setup and extend it in two directions: (i) by evaluating the effect of PointWise and ListWise rerankers, and (ii) by comparing with SIM-RAG, which employs a lightweight critic model to stop retrieval. Our results confirm MetaRAG's relative improvements over standard RAG and reasoning-based baselines, but also reveal lower absolute scores than reported, reflecting challenges with closed-source LLM updates, missing implementation details, and unreleased prompts. We show that MetaRAG is partially reproduced, gains substantially from reranking, and is more robust than SIM-RAG when extended with additional retrieval features.

Question-Adaptive Graph Learning for Multi-hop Retrieval Augmented Generation [Full]

Yuchen Yan (Department of Computer Science, National University of Singapore); Peiyan Zhang (Department of Computer Science and Engineering, Hong Kong University of Science and Technology); Zhihua Liu (Samsung R&D institute China-Beijing); Hao Wang (Samsung R&D institute China-Beijing); Yatao Bian (Department of Computer Science, National University of Singapore); Weiming Li (Samsung R&D institute China-Beijing); Xiaoshuai Hao (Xiaomi EV)
Retrieval-augmented generation (RAG) has demonstrated its ability to enhance Large Language Models (LLMs) by integrating external knowledge sources. However, multi-hop questions, which require the identification of multiple knowledge targets to form a synthesized answer, raise new challenges for RAG systems. Under the multi-hop settings, existing methods often struggle to fully understand the questions with complex semantic structures and are susceptible to irrelevant noise during the retrieval of multiple information targets. To address these limitations, we propose a novel graph representation learning framework for multi-hop question retrieval. We first introduce a Multi-information Level Knowledge Graph (Multi-L KG) to model various information levels for a more comprehensive understanding of multi-hop questions. Based on this, we design a Question-Adaptive Graph Neural Network (Quest-GNN) for representation learning on the Multi-L KG. Quest-GNN employs intra/inter-level message passing mechanisms, and in each message passing the information aggregation is guided by the question, which not only facilitates multi-granular information aggregation but also significantly reduces the impact of noise. To enhance its ability to learn robust representations, we further propose two synthesized data generation strategies for pre-training the Quest-GNN. Extensive experimental results demonstrate the effectiveness of our framework in multi-hop scenarios, especially in high-hop questions the improvement can reach 33.8%. The code is available at: https://github.com/Jerry2398/QSGNN.

Uncertainty Quantification for Retrieval-Augmented Reasoning [Full]

Heydar Soudani (Radboud University); Hamed Zamani (University of Massachusetts Amherst); Faegheh Hasibi (Radboud University)
Retrieval-augmented reasoning (RAR) is a recent evolution of retrieval-augmented generation (RAG) that employs multiple reasoning steps for retrieval and generation. While effective for some complex queries, RAR remains vulnerable to errors and misleading outputs. Uncertainty quantification (UQ) offers methods to estimate the confidence of systems' outputs. These methods, however, often handle simple queries with no retrieval or single-step retrieval, without properly handling RAR setup. Accurate estimation of UQ for RAR requires accounting for all sources of uncertainty, including those arising from retrieval and generation. In this paper, we account for these sources and introduce Retrieval-Augmented Reasoning Consistency (R2C), a novel UQ method for RAR. The core idea of R2C is to perturb the multi-step reasoning process by applying various actions to reasoning steps. These perturbations alter the retriever's input, which shifts its output and consequently modifies the generator's input at the next step. Through this iterative feedback loop, the retriever and generator continuously reshape each other's inputs, enabling us to capture uncertainty arising from both components. Experiments on five popular RAR systems across diverse QA datasets show that R2C improves AUROC by over 5% on average compared to the state-of-the-art UQ baselines. Extrinsic evaluations using R2C as an external signal further confirm its effectiveness for two downstream tasks: in the Abstention task, it achieves ~5% gains in both F1Abstain and AccAbstain; in Model Selection, it improves exact match by ~7% over single models and ~3% over selection methods. Code is available on https://github.com/HeydarSoudani/R2C.

Search for Coverage: Learning Coverage-Aware Retrieval with Augmented Sub-Question Answerability [Full]

Jia-Huei Ju (University of Amsterdam); Eugene Yang (Johns Hopkins University); Trevor Adriaanse (Johns Hopkins University); Suzan Verberne (Leiden University); Andrew Yates (Johns Hopkins University)
Long-form Retrieval-Augmented Generation (RAG) brings the challenge of coverage-based ranking, because ranking methods must ensure the inclusion of comprehensive relevant nuggets (i.e., facts), which can thereby be synthesized into a comprehensive output. In this work, we propose CoveR, a dense retrieval method optimized for coverage-aware retrieval scenarios. CoveR is a bi-encoder trained with the coverage-based contrastive and distillation objectives, which enables CoveR to capture diverse aspects of information needs. To train CoveR, we create the SCOPE dataset, which comprises 90K training pairs from Researchy Questions with synthetic coverage signals augmented from sub-question answerability judgments generated by LLMs. Our empirical experiments show that CoveR enhances nugget coverage by 10% over strong dense retrieval baselines without sacrificing its relevance-based retrieval capability. Further ablation studies validate the importance of our proposed learning method, showing that CoveR achieves a superior trade-off between relevance- and coverage-based ranking, which is essential for long-form RAG.

Learning to Rank · *Eureka 3* · *Wednesday 16:00–17:30*

Explainable, Effective, and Efficient Learning-to-Rank Models Using ILMART [TOIS]

Claudio Lucchese (Ca' Foscari University of Venice, Venice, Italy); Franco Maria Nardini (ISTI-CNR, Pisa, Italy); Salvatore Orlando (Ca' Foscari University of Venice, Venice, Italy); Raffaele Perego (ISTI-CNR, Pisa, Italy); Alberto Veneri (Ca' Foscari University of Venice, Venice, Italy; ISTI-CNR, Pisa, Italy)
Learning ranking models that are both explainable and effective is an emerging topic within the research area of explainable AI. Several Learning-to-Rank (LtR) algorithms have been recently proposed that build models that are simple to explain and, at the same time, almost as effective as their state-of-the-art, black-box counterparts. In this work, we propose ILMART (Interpretable LambdaMART), a novel framework with different strategies to constrain the state-of-the-art LtR LambdaMART algorithm to generate interpretable models, i.e., ensembles whose trees can use either single features (main effects) or a limited number of interacting features (interaction effects). ILMART facilitates a straightforward trade-off between model explainability and effectiveness by precisely tuning the quantity of main and interaction effects during the learning phase. We show that slightly increasing their number allows ILMART models to reach ranking performances at par with full-complexity LambdaMART ones. Furthermore, reproducible experiments conducted on publicly available LtR datasets demonstrate that ILMART can improve nDCG@10 by up to 10% compared to state-of-the-art competitors while preserving an explainable structure. Finally, we explore the relationship between model explainability and

inference efficiency by introducing a novel and easy-to-implement scoring algorithm for ILMART ranking models, achieving up to a 100× speedup compared to the baseline.

OPS: An Order-Preserving Sorting Network for Information Retrieval [Full]

Chao Wang (Kuaishou Technology); Yongxiang Tang (Kuaishou Technology); Guikai Luan (Kuaishou Technology); Kaiyuan Li (Kuaishou Technology); Yanhua Cheng (Kuaishou Technology); Xialong Liu (Kuaishou Technology); Shu Wu (Institute of Automation, Chinese Academy of Sciences); Peng Jiang (Kuaishou Technology)

Learning-to-rank (LTR) is a fundamental component of modern large-scale information retrieval (IR) systems, playing an essential role across various stages of the ranking pipeline. Recently, differentiable sorting networks have attracted increasing attention for LTR as a permutation-level learning paradigm, enabling end-to-end optimization directly on ranking structure. However, existing approaches suffer from two critical limitations: (i) permutation-matrix fidelity, i.e., the predicted soft permutation matrix may deviate from the exact hard permutation matrix required by permutation-level objectives; and (ii) uncertainty in target ordering arising from coarse or tied relevance labels, where the ground-truth order is set-valued rather than unique. To address the first issue, we propose an Order-Preserving Sorting network (OPS), which combines a relaxation operator with random hard swaps to explicitly control the approximation error between the predicted soft permutation matrix and the exact hard permutation matrix. To handle target order uncertainty, we further introduce a novel Variational Order Uncertainty (VOU) loss that enables effective optimization under incomplete or partially ordered supervision. We evaluate our approach through comprehensive experiments on both public benchmarks and real-world systems. On public datasets, including a synthetic four-digit sorting task based on MNIST and LTR benchmarks such as WEB30K and Istella, our method consistently outperforms state-of-the-art baselines. Moreover, we deploy OPS in the advertising pre-ranking stage of a large-scale industrial retrieval system, where it achieves a statistically significant 1.74% revenue uplift in online A/B tests and is currently serving full production traffic. These results demonstrate the effectiveness and robustness of our approach across both academic benchmarks and real-world industrial scenarios. The experimental code is available at https://github.com/lx1563371/OPS_SIGIR2026.

Logit Inflation in ListMLE: Theoretical Analysis and Mitigation Strategies [Full]

Riyaz Ahmad Bhat (IBM Research); Jaydeep Sen (IBM Research)

Modern learning-to-rank methods often rely on listwise objectives that directly model and optimize relative document order over entire permutations. While these objectives improve ranking quality, they frequently produce models with highly inflated relevance scores whose magnitudes exceed what is necessary for meaningful document separation, leading to poor probabilistic calibration. In this work, we analyze this phenomenon in the context of ListMLE [34]. We show that score inflation is a structural consequence of the objective's overlapping softmax normalizers and shift invariance, which generate persistent gradient contributions even after the correct ordering is achieved. As a result, absolute logits drift to increasingly large values, producing saturated and overconfident predictions. We analyze this inflation at the gradient level and identify two complementary mechanisms to address it. Training-time temperature scaling bounds the effective logit scale allowed by shift invariance, while position-dependent discounting suppresses asymmetric tail gradients induced by overlapping suffix normalizers. Together, these modifications prevent persistent score inflation and yield bounded and well-separated scores. We evaluate our approach in a text-based retrieval setting using a state-of-the-art embedding model, and demonstrate substantial improvements in performance under calibration metrics alongside competitive ranking performance on the BEIR benchmark[30].

The Relevance of (Relevant) Entity Presence [Reproducibility]

Norbert Boudens (Radboud University); Chris Kamphuis (Spinque); Arjen P. de Vries (Radboud University)

Query-Specific Document and Entity Representations (QDER) has shown state-of-the-art effectiveness on multiple information retrieval benchmarks. This work attempts to reproduce this promising approach. In an initial attempt, we were unable to find results that resemble the convincing effects measured in the original work. A deep dive into the published artifacts, plus a comparative study of similar work by the same authors show inconsistencies between the descriptions in the paper and their implementation in the published artifacts, and a leakage of relevance assessments due to oversampling entities that are known to appear in relevant documents only. Due to dependencies in the ranking pipeline upon relevance assessments, the final ranking models gain knowledge about the relevance of entities. We conclude that knowing the relevance of entities for an information need can be very valuable, but that the methods described are insufficient to determine this relevance information from the documents and queries alone. The resulting pipeline cannot be deployed effectively in practice, limiting the impact of the research results.

Diagnosing Identifiability in Two-Tower Models for Unbiased Learning to Rank [Full]

Stan Fris (University of Amsterdam); Philipp Hager (University of Amsterdam)

Two-tower models are a popular unbiased learning-to-rank (ULTR) approach for mitigating position bias in clicks. A major challenge in two-tower models is identifiability: whether relevance and position bias can be uniquely determined from clicks. Recent work shows that two-tower models can be identified when the same documents (or documents with similar features) are observed across positions. However, these conditions are defined in infinite data. In practice, we lack methods for diagnosing identifiability in real-world datasets with small sample sizes and limited positional variability. In this work, we present a practical identifiability diagnostic for two-tower models. We quantify whether individual bias parameters are uniquely determined by shifting them away from their optimal values, retraining parts of the model, and measuring whether click-prediction performance degrades significantly. Identified parameters exhibit measurable performance loss when shifted, while unidentified parameters can be freely adjusted without affecting model fit. Our method can distinguish between cases in which identifiability fails due to limitations in model assumptions or data collection and cases caused by a limited sample size. We demonstrate, through simulation, how non-deterministic logging policies, feature overlap, and increased sample size affect identifiability. Applying our method to the real-world Baidu-ULTR dataset, we find that despite large amounts of click data, some parameters of two-tower models remain unidentified, highlighting the practical need for diagnostic methods. All code, data, and results are available at: https://github.com/stanfris/practical-identifiability-ultr

Diagnosing and Mitigating Mid-Sequence Degradation in Recommender Systems [Full]

Linjiang Guo (University of Technology Sydney); Nitin Bisht (University of Technology Sydney); Shiqing Wu (City University of Macau); Huan Huo (University of Technology Sydney); Xianzhi Wang (University of Technology Sydney); Guandong Xu (The Education University of Hong Kong)

Although attention-based sequential recommenders have achieved strong accuracy, they can exhibit a systematic failure mode, i.e., lost-in-the-middle, where attention disproportionately concentrates on sequence boundaries to informative middle behaviors. Through controlled perturbations, we find that this U-shaped bias persists even when collaborative signals are moved, indicating an architectural effect rather than a positional-encoding artifact. To mitigate this bias, we propose Disentangling and Calibrating Position Bias framework, which is a lightweight, plug-and-play calibration procedure that combines bias-adaptive masking during training with counterfactual logit calibration at inference to remove residual position-only effects. Experiments on diverse datasets demonstrate consistent improvements in ranking quality. Further ablations verify the contributions of both masking and calibration components.

Low-Resource IR · *Courtyard 1+2 · Wednesday 16:00–17:30*

Toward a Telugu Question Answering Dataset for Agricultural IR to Serve Marginal Farmers in Drought-Prone Rayalaseema: A Gap Analysis and Roadmap [Low-Resource]

Piyush Joshi (India); Priyansh Singhal (Independent Researcher)

Rayalaseema in southern India is one of the country's most drought-vulnerable regions, experiencing drought in 15 of 18 years between 2000-2018. Over 80% of its 2.7 million farmers are marginal or small landholders practicing rainfed agriculture. Despite pressing agricultural information needs around water management, crop selection, and drought adaptation, these farmers face compounding barriers: 36% rural internet penetration, 60% literacy rates, and information systems designed for English or Hindi speakers. We examine the current landscape of agricultural information retrieval resources for Telugu-speaking farmers and find a critical gap: while agricultural QA datasets exist for Hindi, Tamil, and Bengali, no Telugu agricultural QA benchmark exists, let alone one addressing drought-specific contexts. We document available information sources, identify where they fall short, and propose a methodology for constructing a Telugu Agricultural QA dataset using Kisan Call Center records and regional extension materials. This work charts a path toward inclusive agricultural IR for over 74 million Telugu speakers.

Towards Building a Standard Benchmark for Low-Resource Nepali Information Retrieval [Low-Resource]

Praveen Acharya (Dublin City University); Bal Krishna Bal (Kathmandu University)

Despite being spoken by over 26 million people, Nepali lacks a standardized Information Retrieval (IR) benchmark, preventing systematic evaluation of retrieval models in this low-resource setting. While progress has been made in several Nepali NLP tasks, no publicly available test collection with queries and relevance judgments exists for ad-hoc retrieval task. We propose the creation of a first standardized Nepali IR benchmark supporting monolingual (Nepali queries \rightarrow Nepali documents), cross-lingual (English queries \rightarrow Nepali documents), and code-mixed (Nepali-English queries \rightarrow Nepali documents) retrieval. We argue that Nepali provides a compelling testbed for studying morphology-aware retrieval, code-mixed query processing, and multilingual embedding transfer.

Graph-Enhanced Sentence Retrieval for Multi-Document Summarization in Low-Resource Languages [Low-Resource]

Xuan-Hung Le (VNU University of Engineering and Technology); Thi Toan Do (Independent Researcher); Hoang-Quynh Le (VNU University of Engineering and Technology)

Multi-document summarization for low-resource languages faces a critical trade-off: large language models are computationally prohibitive for most institutions, while smaller models suffer from severe hallucination in abstractive generation. We address this through extractive sentence retrieval, which guarantees faithfulness while operating within constrained computational budgets. Our approach combines language-adaptive mixture-of-experts embeddings with graph neural networks that model

discourse structure, addressing linguistic challenges across typologically diverse low-resource languages. With only 3.2M trainable parameters, our model requires 28 times less training time than comparable transformer-based approaches, making it practical for single-GPU environments typical in resource-limited settings. We demonstrate applicability to some main Southeast Asia (SEA) countries including Vietnam, Thailand, Laos, Indonesia, and Malaysia, representing three distinct language families: Austroasiatic, Kra-Dai, and Austronesian.

ViCSR: A Large-scale Benchmark and Lightweight Two-Stage Framework for Vietnamese Case-to-Statute Retrieval

[Low-Resource]

Minh-Hien Nguyen (VNU University of Engineering and Technology); Khanh Huyen Nguyen (VNU University of Engineering and Technology); Tan-Minh Nguyen (Japan Advanced Institute of Science and Technology); Hoang-Quynh Le (VNU University of Engineering and Technology); Thi-Hai-Yen Vuong (VNU University of Engineering and Technology)

Case-to-Statute Retrieval—identifying applicable statutes from case facts—is essential for judicial efficiency but remains underexplored in low-resource languages like Vietnamese, where annotated legal corpora are scarce and domain-specific challenges persist. To advance research in this setting, we introduce ViCSR, a new benchmark of 10,000 Vietnamese criminal judgments and 1,122 statutory articles with citation-based relevance labels. We further propose a lightweight two-stage framework that performs efficient candidate retrieval with a fine-tuned Vietnamese bi-encoder and leverages citation structure for reranking with a compact heterogeneous GNN. Experiments show clear improvements over strong sparse and dense baselines, highlighting the importance of both domain adaptation and structure-aware modeling in low-resource legal IR. We will publicly release the benchmark, code, and checkpoints.

Bridging the Language Gap in Text-to-SQL: Adapting LLMs for Chichewa in a Low-Resource Setting [Low-Resource]

John Emeka Eze (African Institute for Mathematical Sciences); Dunstan Matekenya (The World Bank Group); Evance Mathewe (The World Bank Group)

Recent advances in Large Language Models (LLMs) have significantly improved Text-to-SQL performance in high-resource languages. However, their effectiveness in low-resource language settings remains largely underexplored. In this work, we investigate the adaptation of LLMs for Text-to-SQL generation in Chichewa, a low-resource Bantu language spoken by over 12 million people in Malawi and neighboring regions. We construct a structured Chichewa Text-to-SQL benchmark consisting of 400 manually curated natural language–SQL pairs grounded in a unified relational database covering agriculture, commodity prices, population statistics, market data, and food insecurity. We systematically evaluate five open-source LLMs under zero-shot, random 5-shot, and retrieval-augmented 5-shot prompting, in both English and Chichewa. We then apply parameter-efficient fine-tuning (QLoRA) to selected models and, crucially, evaluate the combined effect of QLoRA fine-tuning with retrieval-augmented prompting. QLoRA alone improves English execution accuracy to 78.3\% and Chichewa execution accuracy to 41.7\%. When combined with retrieval-augmented prompting, QLoRA achieves 53.3\% execution accuracy in Chichewa, representing the best reported result for this language on this benchmark and narrowing the English–Chichewa gap to 23.4 percentage points. Our findings offer practical guidance for deploying database interfaces in linguistically underserved environments.

Speak Beyond English: Multilingual Prompts Improve Query Classification in Small Language Models [Low-Resource]

Pratyay Banerjee (Department of Computer Science and Engineering, Indian Institute of Information Technology, Guwahati); Panthadeep Bhattacharjee (Department of Computer Science and Engineering, National Institute of Technology, Rourkela); Angshuman Jana (Department of Computer Science and Engineering, Indian Institute of Information Technology, Guwahati)

With recent advancements, Small language models (SLMs) are increasingly used as preprocessors to handle query classification, routing, and candidate selection in retrieval pipelines, but they are nearly always prompted in English, even when users search in Hindi, Bengali, or code-mixed forms. We test whether prompting the same (frozen) SLM in three typologically diverse languages and aggregating the outputs can improve classification without retraining or translation. Nine decoder-only models (1B--9B parameters) evaluated on four public benchmarks show that confidence-weighted fusion of English, Hindi, and Bengali predictions raises macro-F1 by 3--5 points over English-only baselines, with the strongest gains on binary and coarse intent tasks. Parallel execution keeps latency within 1.2--1.4× of the single-language baseline. A paraphrase-only ensemble under identical conditions reaches only +1.4~F1 on average, suggesting that cross-lingual diversity rather than surface-level input variation drives the gain. Because no additional data, training, or translation services are required, our method may be useful when scaling to larger models is out of reach.

Cross-Domain Recommendation · Room 110 · Thursday 10:30–12:00

FLAME: Condensing Ensemble Diversity into a Single Network for Efficient Sequential Recommendation [Full]

WooJoo Kim (Pohang University of Science and Technology); JunYoung Kim (Pohang University of Science and Technology); JaeHyung Lim (Pohang University of Science and Technology); SeongJin Choi (Pohang University of Science and Technology); SeongKu Kang (Korea University); HwanJo Yu (Pohang University of Science and Technology)

Sequential recommendation requires capturing diverse user behaviors, which a single network often fails to capture. While ensemble methods mitigate this, training multiple networks from scratch incurs high computational cost and instability from noisy mutual supervision. We propose Frozen and Learnable networks with Aligned Modular Ensemble (FLAME), a novel framework that condenses ensemble-level diversity into a single network for efficient sequential recommendation. During training, FLAME simulates exponential diversity using only two networks via modular ensemble, which dynamically combines sub-modules (e.g., layers) of each network to generate a rich space of diverse representation patterns. To stabilize training, FLAME pretrains and freezes one network as a semantic anchor and employs guided mutual learning to align diverse representations into the space of remaining learnable network. At inference, FLAME utilizes only the learnable network, achieving ensemble-level performance with zero overhead compared to a single network. Experiments on six datasets show that FLAME outperforms state-of-the-art baselines, achieving up to 7.69× faster convergence and 9.70% improvement in NDCG@20. Our code is available at https://github.com/woo-joo/FLAME_SIGIR26.

Bridging Behavior and Semantics for Time-aware Cross-Domain Sequential Recommendation [Full]

Zhida Qin (School of Computer Science, Beijing Institute of Technology); Zemu Liu (School of Computer Science, Beijing Institute of Technology); Haoyan Fu (School of Computer Science, Beijing Institute of Technology); Chong Zhang (School of Computer Science and Technology, Xi'an Jiaotong University); Tianyu Huang (School of Computer Science, Beijing Institute of Technology); Yidong Li (School of Computer Science and Technology, Beijing Jiaotong University); Gangyi Ding (School of Computer Science, Beijing Institute of Technology)

Cross-domain sequential recommendation (CDSR) alleviates interaction sparsity by jointly modeling user behaviors across multiple domains. While current studies have made some progresses, they still neglect two issues that severely impact recommendation performance: (i) ignoring domain-specific interaction frequencies and interest decay rates at identical time intervals; (ii) treating semantic preferences as time-invariant during cross-domain transfer. To address these, we propose a novel framework that bridges Behavior and Semantics for Time-aware Cross-Domain Sequential Recommendation (BST-CDSR). Specifically, we design a behavioral preference evolution module that decouples long-term interests and short-term intentions, and models continuous-time preference via a neural ordinary differential equation (ODE) with event-driven updates. Additionally, to capture time-aware semantic preferences, we introduce a temporal counterfactual-enhanced semantic generator that discretizes temporal interval tokens and leverages large language models (LLMs) to extract robust temporal semantics, where counterfactual perturbations enhance the time sensitivity of semantic preferences. Furthermore, we propose a time-preference guided domain transfer module to adaptively control transfer weights and mitigate negative transfer. Extensive experiments on real-world datasets demonstrate that BST-CDSR consistently outperforms baselines.

From Clues to Generation: Language-Guided Conditional Diffusion for Cross-Domain Recommendation [Full]

Ziang Lu (Anhui University); Lei Sang (Anhui University); Lin Mu (Anhui University); Yiwen Zhang (Anhui University)

Cross-domain Recommendation (CDR) exploits multi-domain correlations to alleviate data sparsity. As a core task within this field, inter-domain recommendation focuses on predicting preferences for users who interact in a source domain but lack behavioral records in a target domain. Existing approaches predominantly rely on overlapping users as anchors for knowledge transfer. In real-world scenarios, overlapping users are often scarce, leaving the vast majority of users with only single-domain interactions. For these users, the absence of explicit alignment signals makes fine-grained preference transfer intrinsically difficult. To address this challenge, this paper proposes Language-Guided Conditional Diffusion for CDR (LGCD), a novel framework that integrates Large Language Models (LLMs) and diffusion models for inter-domain sequential recommendation. Specifically, we leverage LLM reasoning to bridge the domain gap by inferring potential target preferences for single-domain users and mapping them to real items, thereby constructing pseudo-overlapping data. We distinguish between real and pseudo-interaction pathways and introduce additional supervision constraints to mitigate the semantic noise brought by pseudo-interaction. Furthermore, we design a conditional diffusion architecture to precisely guide the generation of target user representations based on source-domain patterns. Extensive experiments demonstrate that LGCD significantly outperforms state-of-the-art methods in inter-domain recommendation tasks.

LLM-EDT: Large Language Models Enhanced Cross-domain Sequential Recommendation with Dual-phase Training [Full]

Ziwei Liu (City University of Hong Kong); Qidong Liu (Xi'an Jiaotong University); Wanyu Wang (City University of Hong Kong); Yejing Wang (City University of Hong Kong); Pengyue Jia (City University of Hong Kong); Tong Xu (University of Science and Technology of China); Wei Huang (Independent Researcher); Chong Chen (Tsinghua University); Xiangyu Zhao (City University of Hong Kong)

Cross-domain Sequential Recommendation (CDSR) has been proposed to enrich user-item interactions by incorporating information from various domains. Despite current progress, the domain imbalance issue and domain transition issue hinder further development of CDSR. The former presents a

phenomenon where interactions in one domain dominate the entire behavior, leading to difficulty in capturing domain-specific features in the other domain. The latter points to the difficulty in capturing users' cross-domain preferences within the mixed interaction sequence, resulting in poor next-item prediction performance for specific domains. With world knowledge and powerful reasoning abilities, Large Language Models (LLMs) partially alleviate the above issues by functioning as both a generator and an encoder. However, current LLMs-enhanced CDSR methods are still under exploration, which fail to recognize the irrelevant noise and rough profiling problems. Thus, to address the aforementioned challenges, we propose an LLMs Enhanced Cross-domain Sequential Recommendation with Dual-phase Training (LLM-EDT). To address the domain imbalance issue while minimizing irrelevant noise, we propose the transferable item augmenter to adaptively generate possible cross-domain behaviors for users. Then, to alleviate the domain transition issue, we introduce a dual-phase training strategy to empower the domain-specific thread with a domain-shared background. As for the rough profiling problem, we devise a domain-aware profiling module to summarize the user's preference in each domain and adaptively aggregate them to generate comprehensive user profiles. The experiments on three public datasets validate the effectiveness of our proposed LLM-EDT. To ease reproducibility, we have released the detailed code online.
https://github.com/Applied-Machine-Learning-Lab/SIGIR26_LLM-EDT.

Set-Based Cross-Domain Recommendation [Full]

Kyunglim Kim (Samsung Research); James Russell Geraci (Samsung Research)

Cross-domain recommendation is a well-known technique for improving recommendations in a target domain, especially under sparse data or cold-start conditions. A common strategy is to train user embeddings separately in the source and target domains and learn a transfer function between them. In contrast, we propose SetCDR, which constructs more effective user representations in the target domain by directly incorporating each user's source and target history. We additionally introduce a lightweight domain indicator that preserves data–domain relational information. These histories, composed of item–rating pairs, are represented as variable-length sets and processed using a permutation-invariant neural architecture. This differs from conventional neural networks, which do not naturally handle unordered inputs, and is well suited to recommender systems where user history sizes vary greatly among users. The use of a permutation-invariant architecture ensures consistent embeddings regardless of input order, improving robustness to real-world variability and training efficiency. We demonstrate SetCDR in two forms: a simple sum-pooling method and an extended multihead attention–based method that captures more complex dependencies within user histories. Moreover, because SetCDR operates directly on sets of user history records, it provides a natural way to examine how histories influence user representations. Finally, SetCDR adapts immediately to new interactions without additional retraining, enabling on-the-fly performance improvement. Experimental results across multiple cross-domain benchmarks confirm that SetCDR consistently outperforms strong baselines in recommendation quality.

Cross-Domain Interest Representation Learning for Scenario- and Task-Aware Recommendation [Full]

Bokai Lin (Shanghai Jiao Tong University); Naijun Gao (Shanghai Jiao Tong University); Yucen Gao (Northeastern University); Heng Chang (Tsinghua University); Cheng Hu (Independent Researcher); Zhinan Zhang (Independent Researcher); Xiaofeng Gao (Shanghai Jiao Tong University)

Many internet companies operate multiple flagship applications, each of which can be regarded as a distinct business domain, covering areas such as video, reading, and gaming. Within each domain, diverse recommendation scenarios coexist, and users engage in various tasks with heterogeneous behaviors. In our industrial setting, we observe three key phenomena that existing methods rarely address: (i) users' Cross-Domain Interests (CDI) are weakly exploited, since behavior sequences are often pooled for efficiency, losing transferable cross-domain dependencies; (ii) multi-domain, scenario, and task variations are under-modeled, making it difficult to capture fine-grained complementarities; and (iii) multimodal features remain misaligned with ID features, especially when different domains emphasize different modalities. These gaps hinder cross-product collaboration. We propose STAR-CDR, a Scenario- and Task-Aware Cross-Domain Recommendation architecture. Following the modular decomposition of industrial recommenders, STAR-CDR introduces three innovations: (i) a CDI Extractor that enhances sequential modeling with domain indicators and other key features; (ii) a CDI Injector that recalibrates activations and enables domain-/scenario-/task-aware expert routing; and (iii) a CDI Multimodal Adapter that aligns image, text, and ID features under CDI-guided gating. Compared to the strongest baseline in our setting, STAR-CDR improves offline AUC 3.5% across five domains and yields +3.17% to +9.24% relative KPI gains in online A/B tests. The system now supports multiple domains (flagship applications), scenarios (recommendation scenarios), and tasks (user behaviors) at scale, serving over 280 million daily active users. We provide a GitHub repository. https://github.com/Rescomk/STAR-CDR with STAR-CDR's code, training scripts, and a synthetic data construction pipeline.

Multimodal Representation Learning · *Eureka 1* ·

Thursday 10:30–12:00

Robust Multimodal Recommendation via Graph Retrieval-Enhanced Modality Completion [Full]

Yuan Li (National University of Singapore); Jun Hu (National University of Singapore); Jiaxin Jiang (National University of Singapore); Bryan Hooi (National University of Singapore); Bingsheng He (National University of Singapore)

Multimodal data plays a critical role in web-based recommendation systems, where information from diverse modalities such as vision and text enhances representation learning. However, real-world multimodal datasets often suffer from modality incompleteness due to sensor failures, annotation scarcity, or privacy constraints, which substantially degrade model performance and reliability. One effective solution to address this issue is modality completion, which reconstructs missing features to provide modality-complete graphs for downstream tasks. Given a query node with missing multimodal features, existing modality completion methods typically infer information from the node itself or its neighbors to reconstruct the missing modality. However, these methods may overlook semantically relevant context in the graph, which contains valuable cues that are non-trivial to capture through simple methods like neighborhood aggregation. In this work, we propose GRE-MC, a Graph Retrieval–Enhanced Modality Completion framework, to overcome these limitations. By introducing a modality-aware subgraph retrieval mechanism, GRE-MC selects semantically relevant subgraphs from the entire graph, providing richer contextual information for completing missing modalities. Subsequently, a graph transformer jointly encodes the query node and the retrieved subgraph via global attention to complete the missing features, while a learnable sparse-routing codebook regularizes latent embeddings into compact bases for improved robustness. Extensive experiments on multimodal recommendation benchmarks demonstrate that GRE-MC consistently outperforms state-of-the-art methods, validating the effectiveness of subgraph retrieval and joint-encoding graph transformer for robust modality completion.

Discrete Preference Learning for Personalized Multimodal Generation [Full]

Yuting Zhang (Hong Kong University of Science and Technology (Guangzhou)); Ying Sun (Hong Kong University of Science and Technology (Guangzhou)); Dazhong Shen (Nanjing University of Aeronautics and Astronautics); Ziwei Xie (OPPO Research Institute); Feng Liu (OPPO Research Institute); Changwang Zhang (OPPO Research Institute); Xiang Liu (OPPO Internet Services System); Jun Wang (OPPO Research Institute); Hui Xiong (Hong Kong University of Science and Technology (Guangzhou))

The emergence of generative models enables the creation of texts and images tailored to users' preferences. Existing personalized generative models have two critical limitations: lacking a dedicated paradigm for accurate preference modeling, and generating unimodal content despite real-world multimodal-driven user interactions. Therefore, we propose personalized multimodal generation, which captures modal-specific preferences via a dedicated preference model from multimodal interactions, and then feeds them into downstream generators for personalized multimodal content. However, this task presents two challenges: (1) Gap between continuous preferences from dedicated modeling and discrete token inputs intrinsic to generator architectures; (2) Potential inconsistency between generated images and texts. To tackle these, we present a two-stage framework called Discrete Preference learning for Personalized Multimodal Generation (DPPMG). In the first stage, to accurately learn discrete modal-specific preferences, we introduce a modal-specific graph neural network (a dedicated preference model) to learn users' modal-specific preferences, which preferences are then quantized into discrete preference tokens. In the second stage, the discrete modal-specific preference tokens are injected into downstream text and image generators. To further enhance cross-modal consistency while preserving personalization, we design a cross-modal consistent and personalized reward to fine-tune token-associated parameters. Extensive experiments on two real-world datasets demonstrate the effectiveness of our model in generating personalized and consistent multimodal content.

Reconstructing Content with Collaborative Attention for Universal Multimodal Representation Learning [Full]

Jiahan Chen (State Key Laboratory of AI Safety); Da Li (State Key Laboratory of AI Safety); Hengran Zhang (State Key Laboratory of AI Safety); Yinqiong Cai (Baidu Inc.); Lixin Su (Baidu Inc.); Jiafeng Guo (State Key Laboratory of AI Safety); Daiting Shi (Baidu Inc.); Dawei Yin (Baidu Inc.); Keping Bi (State Key Laboratory of AI Safety)

Multimodal embedding models, rooted in multimodal large language models (MLLMs), have yielded significant performance improvements across diverse tasks such as retrieval and classification. However, most existing approaches rely heavily on large-scale contrastive learning and offer limited exploration of how the architectural and training paradigms of MLLMs affect embedding quality. While effective for generation, the causal attention and next-token prediction paradigm of MLLMs does not explicitly encourage the formation of globally compact representations, limiting their effectiveness as multimodal embedding backbones. To address this, we propose CoCoA, a Content reconstruction pre-training paradigm based on Collaborative Attention for universal multimodal representation learning. Specifically, we restructure the attention flow and introduce an EOS-based reconstruction task, encouraging the model to reconstruct input from the corresponding (EOS) embeddings. This drives the multimodal model to compress the semantic information of

the input into the (EOS) token, laying the foundations for subsequent contrastive learning. Extensive experiments on MMEB-V1 demonstrate that CoCoA built upon Qwen2-VL and Qwen2.5-VL significantly improves embedding quality. Results validate that content reconstruction serves as an effective strategy to maximize the value of existing data, enabling multimodal embedding models to generate compact and informative representations, raising their performance ceiling. Our project is available at https://github.com/Trustworthy-Information-Access/CoCoA.

RE-TRIANGLE: Does TRIANGLE Enable Multimodal Alignment Beyond Cosine Similarity in Retrieval? [Reproducibility]

Arijit Ghosh (University of Amsterdam); Aritra Bandyopadhyay (University of Amsterdam); Chiranjeev Bindra (University of Amsterdam); Jingfen Qiao (University of Amsterdam)

Multimodal alignment is critical for bridging the semantic gap in information retrieval. However, traditional pairwise strategies introduce a geometric blind spot: while they align anchor modalities (e.g., text) with others, they lack constraints to enforce mutual consistency between peripheral modalities (e.g., video and audio). The TRIANGLE framework addresses this by minimizing the area of modality triplets on a hypersphere to enforce holistic alignment. In this reproducibility study, we verify the robustness of this geometric objective for retrieval tasks. We confirm that TRIANGLE outperforms pairwise baselines in zero-shot settings, achieving Recall@1 gains of up to +8.7 points, though benefits are domain-dependent. However, we fail to reproduce the reported learning-from-scratch results. Analysis using a synthetic toy dataset attributes this to instability when jointly optimizing geometric alignment with Data-Text Matching (DTM) loss. Furthermore, we find that cosine regularization primarily stabilizes text-to-video retrieval, and fine-tuning with domain supervision amplifies geometric benefits but reduces cross-dataset generalization. Our findings support the efficacy of geometric alignment while highlighting critical optimization sensitivities.

NeuCon-ICE: Neuron-Level Controllable In-Context Editing for Multimodal Large Language Models [Full]

Chao Jiang (National University of Defense Technology); Jinzhi Liao (National University of Defense Technology); Xiang Zhao (National University of Defense Technology)

Multimodal knowledge editing (MKE) aims to efficiently rectify outdated or incorrect knowledge in multimodal large language models (MLLMs) while preserving reliability, generality, and locality. Recently, in-context editing (ICE) has emerged as a prevalent paradigm for MKE, focusing on inference-time context manipulation. While ICE mitigates the side effects of intrinsic interventions on MLLMs, it still suffers from an out-of-control limitation. Specifically, previous methods rely excessively on the implicit contextualization of MLLMs and regard the MKE process as a matter of chance. Based on this observation, we further identify the corresponding challenges as localizing the responsible key units and defining the triggering conditions within MLLMs. To address these challenges, we pursue a controllable ICE approach and refer to the multimodal neurons. Consequently, we propose a neuron-level controllable ICE framework for MKE, namely NeuCon-ICE. It consists of (1) a multimodal contextual neuron identification module that aims to determine where to edit by identifying the tightly coupled multimodal contextual neurons, and (2) a context-aware neuron editing module that aims to determine how to edit by selectively injecting context-aware updates into the identified neurons. Experiments on three representative MLLMs (BLIP-2, MiniGPT-4, and LLaVA 1.5) on ComprehendEdit and E-VQA demonstrate that NeuCon-ICE consistently achieves state-of-the-art overall performance, delivering an overall gain of at least 10.79% across baselines and datasets. The code is available at https://github.com/jc4357/NeuCon-ICE.

Seeing the Whole Through the Parts: Discovering Objects through Semantic Part Mining in Weak Supervision [Full]

Shucheng Li (Central South University); Weixuan Xu (Central South University); Le Jiang (Central South University); Hao Wu (National Key Lab for Novel Software Technology, Nanjing University); Fengyuan Xu (National Key Lab for Novel Software Technology, Nanjing University); Fan Wu (Central South University); Feng Lyu (Central South University)

Weakly Supervised Object Detection (WSOD) is fundamentally limited by instance ambiguity, manifesting as either part domination (focusing on discriminative fragments) or merged detection (confusing objects with context). Unlike existing approaches that rely on object-level patterns, we draw inspiration from human cognition, where objects are perceived structurally by integrating constituent parts. Building on this perspective, we propose P2WDet (Part-to-Whole Detection), a novel framework that shifts WSOD from traditional instance selection to semantic reconstruction. P2WDet comprises three systematic stages: (1) constructing a comprehensive visually-grounded part vocabulary leveraging large language models; (2) training dedicated part detectors by mining consistent patterns from cross-image proposal clusters; and (3) assembling detected parts into complete objects to enforce structural consistency. Extensive experiments show that P2WDet not only outperforms state-of-the-art methods on standard benchmarks but also demonstrates superior generalization in open-world settings. Furthermore, by decoupling detection from fixed category labels, P2WDet enables the flexible detection of novel or ambiguous objects defined solely by their components.

Explainability, Alignment, and Debiasing · *Eureka 2* ·

Thursday 10:30–12:00

Understanding Internal Representations of Recommendation Models with Sparse Autoencoders [TOIS]

Jiayin Wang (Tsinghua University, Beijing, China); Xiaoyu Zhang (Tsinghua University, Beijing, China); Weizhi Ma (Tsinghua University, Beijing, China); Zhiqiang Guo (Tsinghua University, Beijing, China); Min Zhang (Tsinghua University, Beijing, China)

Recommendation model interpretation aims to reveal the relationships between inputs, model internal representations, and outputs to enhance the transparency, interpretability, and trustworthiness of recommendation systems. However, the inherent complexity and opacity of deep learning models pose challenges for model-level interpretation. Moreover, most existing methods for interpreting recommendation models are tailored to specific architectures or model types, limiting their generalizability across different types of recommenders. In this article, we propose RecSAE, a generalizable probing framework that interprets Recommendation models with Sparse AutoEncoders. The framework extracts interpretable latents from the internal representations of recommendation models and links them to semantic concepts for interpretations. It does not alter original models during interpretations and also enables targeted tuning to models. Experiments on three types of recommendation models (general, graph-based, sequential) with four widely used public datasets demonstrate the effectiveness and generalization of the RecSAE framework. The interpreted concepts are further validated by human experts, showing strong alignment with human perception. Overall, RecSAE serves as a novel step in both model-level interpretations to various types of recommendation models without affecting their functions and offering potential for targeted tuning of models.

Towards End-to-End Alignment of User Satisfaction via Questionnaire in Video Recommendation [Full]

Na Li (Kuaishou Technology); Jiaqi Yu (Kuaishou Technology); Minzhi Xie (Kuaishou Technology); Tiantian He (Kuaishou Technology); Xiaoxiao Xu (Kuaishou Technology); Zixiu Wang (Kuaishou Technology); Lantao Hu (Kuaishou Technology); Yongqi Liu (Kuaishou Technology); Han Li (Kuaishou Technology); Kaiqiao Zhan (Kuaishou Technology); Kun Gai (Kuaishou Technology)

Short-video recommender systems typically optimize ranking models using dense user behavioral signals, such as clicks and watch time. However, these signals are only indirect proxies of user satisfaction and often suffer from noise and bias. Recently, explicit satisfaction feedback collected through questionnaires has emerged as a high-quality direct alignment supervision, but is extremely sparse and easily overwhelmed by abundant behavioral data, making it difficult to incorporate into online recommendation models. To address these challenges, we propose a novel framework which is towards End-to-End Alignment of user Satisfaction via Questionnaire, named EASQ, to enable real-time alignment of ranking models with true user satisfaction. Specifically, we first construct an independent parameter pathway for sparse questionnaire signals by combining a multi-task architecture and a lightweight LoRA module. The multi-task design separates sparse satisfaction supervision from dense behavioral signals, preventing the former from being overwhelmed. The LoRA module pre-inject these preferences in a parameter-isolated manner, ensuring stability in the backbone while optimizing user satisfaction. Furthermore, we employ a DPO-based optimization objective tailored for online learning, which aligns the main model outputs with sparse satisfaction signals in real time. This design enables end-to-end online learning, allowing the model to continuously adapt to new questionnaire feedback while maintaining the stability and effectiveness of the backbone. Extensive offline experiments and large-scale online A/B tests demonstrate that EASQ consistently improves user satisfaction metrics across multiple scenarios. EASQ has been successfully deployed in a production short-video recommendation system, delivering significant and stable business gains.

From Top-1 to Top-K: A Reproducibility Study and Benchmarking of Counterfactual Explanations for Recommender Systems [Reproducibility]

Quang-Huy Nguyen (VNU University of Engineering and Technology); Thanh-Hai Nguyen (VNU University of Engineering and Technology); Khac-Manh Thai (VNU University of Engineering and Technology); Duc-Hoang Pham (VNU University of Engineering and Technology); Huy-Son Nguyen (Delft University of Technology); Cam-Van Thi Nguyen (VNU University of Engineering and Technology); Masoud Mansoury (Delft University of Technology); Duc-Trong Le (VNU University of Engineering and Technology); Hoang-Quynh Le (VNU University of Engineering and Technology)

Counterfactual explanations (CEs) provide an intuitive way to understand recommender systems by identifying minimal modifications to user-item interactions that alter recommendation outcomes. Existing CE methods for recommender systems, however, have been evaluated under heterogeneous protocols, using different datasets, recommenders, metrics, and even explanation formats, which hampers reproducibility and fair comparison. In this paper, we systematically reproduce, re-implement, and re-evaluate eleven state-of-the-art CE methods for recommender systems, covering both native explainers (e.g., LIME-RS, SHAP, PRINCE, ACCENT, LXR, GREASE) and specific graph-based explainers originally proposed for GNNs. Building on these implementations, a unified benchmarking framework is proposed to assess explainers along three dimensions: explanation format (implicit vs. explicit), evaluation level (item-level vs.

list-level), and perturbation scope (user interaction vectors vs. user-item interaction graphs). Our evaluation protocol includes effectiveness, sparsity, and computational complexity metrics, and extends existing item-level assessments to top-K list-level explanations. Through extensive experiments on three real-world datasets and six representative recommender models, we analyze how well previously reported strengths of CE methods generalize across diverse setups. We observe that the trade-off between effectiveness and sparsity depends strongly on the specific method and evaluation setting, particularly under the explicit format; in addition, explainer performance remains largely consistent across item level and list level evaluations, and several graph-based explainers exhibit notable scalability limitations on large recommender graphs. Our results refine and challenge earlier conclusions about the robustness and practicality of CE generation methods in recommender systems, and provide a reproducible reference point for future work: \href{https://github.com/L2R-UET/CFExpRec}{https://github.com/L2R-UET/CFExpRec}.

Minimal-Perturbation Counterfactuals through Guided Denoising Diffusion for Recommender Systems Explanation [Full]

Amir Reza Mohammadi (University of Innsbruck); Andreas Peintner (University of Innsbruck); Michael M. Müller (University of Innsbruck); Eva Zangerle (University of Innsbruck)

Debiased Recommendation Beyond the Positive Propensity Assumption [Full]

Yanghao Xiao (University of Chinese Academy of Sciences); Hao Wang (Zhejiang University); Xiang Li (Peking University); Qian Zou (Meituan); Cheng Bing (Meituan); Wei Lin (Meituan); Haoxuan Li (Peking University); Zhouchen Lin (Peking University)

Post-click conversion rate (CVR) prediction is a central task in recommender systems, yet selection bias creates a severe distributional gap between the clicked training samples and the entire inference space. To address selection bias, propensity-based methods such as inverse propensity scoring (IPS) and doubly robust (DR) have been adopted, which aim to estimate the unbiased learning objective from biased training samples. However, these approaches assume strictly positive propensities, implying every user-item pair has a nonzero probability of interaction. In practice, such positivity assumption maybe violated, for example, in food-delivery platforms, some restaurants located more than 10 kilometers away will be blocked for recommendation. In this study, we theoretically show that when such zero-propensity samples, termed extrapolation samples exist, both IPS and DR estimators become biased. To overcome this limitation, we propose ExtraDebias method, which enables debiased recommendation in both non-extrapolation and extrapolation samples. Specifically, we first train a propensity model to identify extrapolation samples with extremely small propensity estimates, then estimate their pseudo-label intervals, and derive an upper bound of the learning objective for extrapolation samples. By minimizing the derived upper bound, debiased learning on extrapolation samples is ensured, while unbiased learning on non-extrapolation samples is achieved by standard IPS. Experiments on four real-world offline datasets and one online A/B test show that ExtraDebias effectively minimizes prediction errors on extrapolation samples and achieves optimal performance.

Debiased Multimodal Personality Understanding through Dual Causal Intervention [Full]

Yangfu Zhu (Capital Normal University); Zitong Han (Capital Normal University); Nianwen Ning (Henan Univeristy); Yuting Wei (University of International Relations); Yuandong Wang (Capital Normal University); Feng Hang (Capital Normal University); Zhenzhou Shao (Capital Normal University)

\begin{abstract}Multimodal personality understanding plays a critical role in human-centered artificial intelligence. Previous work mainly focus on learning rich multimodal representations for video personality understanding. However, they often suffer from potential harm caused by subject bias (e.g., observable age and unobservable mental states), as subjects originate from diverse demographic backgrounds. Learning such spurious associations between multimodal features and traits may lead to unfair personality understanding. In this work, we construct a Structural Causal Model (SCM) to analyze the impact of these biases from a causal perspective, and propose a novel \textbf{D}ual \textbf{C}ausal \textbf{A}djustment Network \textbf{(DCAN)} to mitigate the interference of subject attributes on personality understanding. Specifically, we design a Back-door Adjustment Causal Learning (BACL) module to block spurious correlations from observable demographic factors via a prototype-based confounder dictionary, and subsequently apply a Front-door Adjustment Causal Learning (FACL) module to address latent and unobservable biases through a learned mediator dictionary intervention, thereby achieving causal disentanglement of representations for deconfounded reasoning. Importantly, we construct a \textbf{D}emographic-annotated \textbf{M}ultimodal \textbf{S}tudent \textbf{P}ersonality \textbf{(DMSP)} dataset to support the analysis and discussion of fairness-related factors.Extensive experiments on the benchmark dataset CFI-V2 and our DMSP dataset demonstrate that DCAN consistently improves prediction accuracy, reaching 92.11\% and 92.90\%, respectively.Meanwhile, the improvements in the fairness metrics of equal opportunity and demographic parity are 6.57\% and 7.97\% on CFI-V2, and 15.38\% and 20.06\% on the DMSP dataset. Our code and DMSP dataset are availableat \url{https://github.com/Sabrina-han/DCAN}\end{abstract}

Query Reformulation and Tool Use · *Eureka 3* ·

Thursday 10:30–12:00

MCP-Focus: Leveraging Function-Oriented Document Enhancement for MCP Server Retrieval [Full]

Wenchun Jing (Peking University); Haiyang Shen (Peking University); Haoran Wang (Peking University); Qi Liu (Renmin University of China); Ningyuan Li (Beijing University of Technology); Chaoran Luo (National Key Laboratory of Dataspace Technology and System); Ning Zhang (Xidian University); Yun Ma (Peking University)

From Translation to Retrieval: Evaluating LLM-Based Information Retrieval for Hausa and Fongbe [Low-Resource]

Mahounan Pericles Adjovi (Carnegie Mellon University Africa); Roald Eiselen (North-West University); Prasenjit Mitra (Carnegie Mellon University Africa)

LLM-based reranking has been evaluated for some African languages, but whether LLM-based query expansion helps or hurts retrieval for low-resource African languages remains an open question. Adeyemi et al. evaluated cross-lingual LLM reranking for Hausa with English queries, yet to our knowledge no published work has evaluated LLM-based query expansion for Hausa or Fongbe specifically, and no IR evaluation resources were found for Fongbe. This study builds upon our prior work on LLM translation quality evaluation and data augmentation for corpus expansion in Hausa and Fongbe. We propose experiments that compare LLM reranking and query expansion against BM25 and multilingual dense retrieval baselines (mDPR, mContriever) for Hausa and Fongbe using three commercial LLMs. Our completed translation quality assessment confirms a large LLM-capability gap between the two languages (best BLEU: Hausa 15.75 vs. Fongbe 7.18; human scores 4.5/5 vs. 2.2/5), and our data augmentation experiments across three encoder models show that LLM-generated text consistently hurts downstream NER tasks while producing mixed effects on POS tagging, motivating careful language-specific IR evaluation. We plan to use the CIRAL test collection for Hausa and to construct a new cross-lingual test set derived from Fongbe Wikipedia data following the AfriCLIRMatrix methodology.

Morphology-Aware Retrieval for Low-Resource Environments: Advancing Information Retrieval for Shona Language [Low-Resource]

Tendai Mukande (Dublin City University); Noel O'Connor (Dublin City University); Ruvimbo Maud Munetsi (University of Zimbabwe)

Research on Information Retrieval (IR) has historically prioritised high-resource languages such as English and Chinese, with less attention given to many low-resource languages. For example, Shona, a Bantu language spoken by approximately 12 million people in Zimbabwe and neighbouring countries, remains under-explored in IR research despite its widespread societal use in Southern Africa. In this work, we present a preliminary study of Shona IR using sparse and dense retrieval models, demonstrating significant performance limitations due to morphological complexity and data scarcity. Based on these findings, we propose to develop a framework to advance Shona IR by developing a large-scale benchmark dataset to support morphology-aware retrieval. We hypothesise that improving Shona IR supports equitable access to digital information and enables language-inclusive AI technologies aligned with global development priorities such as accessibility to education, dissemination of healthcare information, and digital inclusion.

When More Reformulations Hurt: Avoiding Drift using Ranker Feedback [Full]

Venktesh V (Stockholm University); Mandeep Rathee (L3S Research Center); Avishek Anand (Delft University of Technology)

Modern retrieval pipelines increasingly rely on query reformulation and neural reranking to improve effectiveness, but this comes at a significant computational cost and introduces a fundamental tradeoff between recall and query drift. Generating many reformulated queries can substantially increase recall, yet naïvely merging or exhaustively reranking their results is prohibitively expensive. In this work, we argue that the core challenge is not reformulation generation itself, but the adaptive selection of reformulations and their retrieved documents under a strict inference budget. We propose ReformIR, a budget-aware retrieval framework that treats query reformulations as first-class features and performs online relevance estimation using a strong reranker as a teacher. Given multiple reformulated queries, ReformIR constructs a large candidate pool and learns a lightweight surrogate model that estimates document utility from reformulation-specific retrieval signals. Under a fixed reranking budget, the surrogate adaptively prioritizes both reformulations and documents, selectively querying a teacher reranker anchored to the original query. This process increases recall while actively suppressing drift through online feature selection over reformulations. We conduct extensive experiments on the MSMARCO passage corpora and TREC Deep Learning benchmarks (DL19–DL22), evaluating effectiveness under realistic reranking budgets. Our results show that ReformIR consistently outperforms existing reformulation strategies, particularly as the number of reformulations increases, where prior methods suffer from severe quality degradation due to drift. Our findings also suggest a shift in retrieval system design: rather than using large language models as rerankers, their capacity is more effectively leveraged in the reformulation stage with feedback-driven optimization. Our code: This paragraph is centered. https://github.com/VenkteshV/ReformIR

A Reproducibility Study of LLM-Based Query Reformulation [Reproducibility]

Amin Bigdeli (University of Waterloo); Radin Hamidi Rad (Mila – Quebec AI Institute); Hai Son Le (Toronto Metropolitan University); Mert Incesu (University of Toronto); Negar Arabzadeh (University of California, Berkeley); Charles L. A. Clarke (University of Waterloo); Ebrahim Bagheri (University of Toronto)

Large Language Models (LLMs) are now widely used for query reformulation and expansion in Information Retrieval, with many studies reporting substantial effectiveness gains. However, these results are typically obtained under heterogeneous experimental conditions, making it difficult to assess which findings are reproducible and which depend on specific implementation choices. In this work, we present a systematic reproducibility and comparative study of ten representative LLM-based query reformulation methods under a unified and strictly controlled experimental framework. We evaluate methods across two architectural LLM families at two parameter scales, three retrieval paradigms (lexical, learned sparse, and dense), and nine benchmark datasets spanning TREC Deep Learning and BEIR. Our results show that reformulation gains are strongly conditioned on the retrieval paradigm, that improvements observed under lexical retrieval do not consistently transfer to neural retrievers, and that larger LLMs do not uniformly yield better downstream performance. These findings clarify the stability and limits of reported gains in prior work. To enable transparent replication and ongoing comparison, we release all prompts, configurations, evaluation scripts, and run files through QueryGym, an open-source reformulation toolkit with a public leaderboard. https://leaderboard.querygym.com

Tool-Star: Empowering Multi-Tool Collaborative Web Agent via Reinforcement Learning [Full]

Guanting Dong (Renmin University of China); Yifei Chen (Renmin University of China); Xiaoxi Li (Renmin University of China); Jiajie Jin (Renmin University of China); Hongjin Qian (Beijing Academy of Artificial Intelligence); Yutao Zhu (Renmin University of China); Hangyu Mao (Microelectronics, Chinese Academy of Sciences); Guorui Zhou (Kuaishou Technology); Zhicheng Dou (Renmin University of China); Ji-Rong Wen (Renmin University of China)

Recently, Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities through reinforcement learning (RL). However, enabling LLM-based agents to effectively orchestrate multiple tools remains an open challenge. In this paper, we introduce Tool-Star, an end-to-end agentic post-training framework that empowers LLM-based web agents to strategically interact with external multi-tool environments. Tool-Star begins with a general tool-integrated data synthesis pipeline that combines two complementary sampling strategies to generate tool-use trajectories, followed by quality normalization and difficulty-aware curriculum construction to filter noisy samples and organize training data from easy to hard. We then introduce a two-stage training paradigm for multi-tool collaborative reasoning: (1) cold-start supervised fine-tuning with tool feedback to bootstrap long-horizon tool-augmented reasoning, and (2) a multi-tool self-critic RL algorithm with hierarchical rewards to reinforce effective tool coordination. Experiments across 13 benchmarks demonstrate Tool-Star's effectiveness. Further analyses provide practical insights for optimizing strategic tool use in web agents. The code is available at https://github.com/RUC-NLPIR/Tool-Star.

Evaluation Methodology · *Courtyard 1+2 · Thursday 10:30–12:00*

Decision-Theoretic Stopping Rules for Document Screening [Full]

Aaron H.A. Fletcher (University of Sheffield); Mark Stevenson (University of Sheffield)

Deciding when to stop reviewing the results of a search is a common problem with multiple applications. Existing stopping rules developed within Technology-Assisted Review (TAR) aim to achieve a pre-specified recall target and do not take into account the reason for examining the results, potentially leading to sub-optimal recommendations. This paper applies decision theory to the problem and uses it to derive three practical stopping policies based on the Expected Value of Perfect Information. The approach is applied to two professional search tasks: patent examining and systematic reviewing. Experiments on CLEF-IP and medical systematic review datasets show that the proposed approach generally produces more appropriate stopping decisions than existing methods, as demonstrated by higher net utility under the evaluated cost and payoff settings.

How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews [Full]

Riley Grossman (New Jersey Institute of Technology); Songjiang Liu (New Jersey Institute of Technology); Michael K. Chen (Nanyang Technological University); Mike Smith (Indiana University Bloomington); Cristian Borcea (New Jersey Institute of Technology); Yi Chen (New Jersey Institute of Technology)

Generative AI is being increasingly integrated into web search for the convenience it provides users. In this work, we aim to understand how generative AI disrupts web search by retrieving and presenting the information and sources differently from traditional search engines. We introduce a public benchmark dataset of 11,500 user queries to support our study and future research of generative search. We compare the search results returned by Google's search engine, the accompanying AI Overview (AIO), and Gemini Flash 2.5 for each query. We have made several key findings. First, we find that for 51.5% of representative, real-user queries, AIOs are generated, and are displayed above the organic search results. Controversial questions frequently result in an AIO. Second, we show that the retrieved sources are substantially different for each search engine (<0.2 average Jaccard similarity). Traditional Google search is significantly more likely to retrieve information from popular or institutional websites in government or education, while generative search engines are significantly more likely to retrieve Google-owned content. Third, we observe that websites that block Google's AI crawler are significantly less likely to be retrieved by AIOs, despite having access to the content. Finally, AIOs are less consistent when processing two runs of the same query, and are less robust to minor query edits. Our findings have important implications for understanding how generative search impacts website visibility, the effectiveness of generative engine optimization techniques, and the information users receive. We call for revenue frameworks to foster a sustainable and mutually beneficial ecosystem for publishers and generative search providers.

FACE: A Fine-Grained Reference-Free Evaluator for Conversational Information Access [Full]

Hideaki Joko (Radboud University); Faegheh Hasibi (Radboud University)

A systematic, reliable, and low-cost evaluation of Conversational Information Access (CIA) systems remains an open challenge. Existing reference-based evaluation methods have proven insufficient for evaluating the dynamic nature of information access conversations, while existing LLM-based reference-free methods suffer from evaluation bias and limited generalizability. We propose FACE: a Fine-grained, Aspect-based Conversation Evaluator that evaluates diverse turn- and dialogue-level aspects of conversations. FACE leverages beam search and bandit optimization to select optimized LLM instructions for each evaluation aspect. It assigns scores to atomic information units (particles) using the selected instructions and then aggregates them into a single score. We show that FACE achieves strong correlations with human judgments, outperforming state-of-the-art conversation evaluation methods by a large margin. We further demonstrate that its optimized instructions are transferable across various LLMs and datasets. Additionally, unlike existing LLM-based methods that provide single uninterpretable scores, FACE provides insights into the system performance and enables identifying and locating problems within conversations.

Evaluation Validity in Information Retrieval [Perspective]

Paul Thomas (Microsoft); Nick Craswell (Microsoft); Mark Sanderson (RMIT University); Seth Spielman (Microsoft); Robert Sim (Microsoft); Ryen W White (Microsoft)

Information retrieval has long relied on evaluations that measure system performance. Improvements on standard evaluation protocols are interpreted as progress in system effectiveness, on the understanding that improved metrics indicate a better experience. However, most evaluations are a drastic abstraction and simplification of that experience. It is reasonable to inquire after the validity of our evaluations, or the degree to which they do in fact represent phenomena we care about. If a metric improves, can we be sure there is a corresponding improvement in real-world effectiveness? We discuss practical ways to discuss, measure, and improve the validity of evaluations in a range of settings. By considering validity, we can make better choices in evaluation protocols; we have a chance to make progress if and when evaluating and retrieving collapse into each other entirely, e.g., with LLM-as-judge; and we can optimise towards systems that people actually want.

Evaluation of Agents under Simulated AI Marketplace Dynamics [Perspective]

To Eun Kim (Carnegie Mellon University); Alireza Salemi (University of Massachusetts Amherst); Hamed Zamani (University of Massachusetts Amherst); Fernando Diaz (Carnegie Mellon University)

Modern information access ecosystems consist of mixtures of systems, such as retrieval systems and large language models, and increasingly rely on marketplaces to mediate access to models, tools, and data, making competition between systems inherent to deployment. In such settings, outcomes are shaped not only by benchmark quality but also by competitive pressure, including user switching, routing decisions, and operational constraints. Yet evaluation is still largely conducted on static benchmarks with accuracy-focused measures that assume systems operate in isolation. This mismatch makes it difficult to predict post-deployment success and obscures competitive effects such as early-adoption advantages and market dominance. We introduce Marketplace Evaluation, a simulation-based paradigm that evaluates information access systems as participants in a competitive marketplace. By simulating repeated interactions and evolving user and agent preferences, the framework enables longitudinal evaluation and marketplace-level metrics, such as retention and market share, that complement and can extend beyond traditional accuracy-based metrics. We formalize the framework and outline a research agenda, motivated by business and economics, around marketplace simulation, metrics, optimization, and adoption in evaluation campaigns like TREC.

Stop Using the Wilcoxon Test: Myth, Misconception and Misuse in IR Research [Perspective]

Julián Urbano (Delft University of Technology)

In benchmarking of Information Retrieval systems, the Wilcoxon signed-rank test is often treated as a safer alternative to the t-test. This belief is fueled by textbooks and recommendations that portray Wilcoxon as the proper non-parametric alternative because metric scores are not normally distributed. We argue that this narrative is misleading and harmful. A careful review of Statistics textbooks reveals inconsistencies and omissions in how

the assumptions underlying these tests are presented, fostering confusion that has propagated into IR research. As a result, Wilcoxon has been routinely misapplied for decades, creating a false sense of safety against a threat that was never there to begin with, while introducing another one so severe that it virtually guarantees the test will break down and mislead researchers. Through a combination of systematic literature review, analysis and empirical demonstrations with TREC data, we show how and why the Wilcoxon test easily loses control of its Type I error rate in IR settings. We conclude that the continued use of Wilcoxon in IR evaluation is unjustified and that abandoning it would improve the methodological soundness of our field.

Industry Session I · *Goldfields · Thursday 10:30–12:00*

Improving Dating Outcomes by Predicting for Conversations at Hinge [Industry]

Frankie Lu (Hinge); Dan Turkel (Hinge); Sihan Chen (Hinge); Behrooz Afghahi (Hinge)

The ultimate measure of success for a dating platform is helping people go on in-person dates. In practice, ranking systems for such platforms rely on observable proxies, such as likes and matches, to predict whether an in-person date will occur. This paper introduces conversations, when both users have sent at least one message, as a stronger signal of dating intent. We model the dating journey as a probabilistic chain, decomposing the likelihood of a conversation into three sequential stages: the probability of a like, a match given a like, and a conversation given a match. By targeting conversations as a late-stage observable proxy for dating intent, this approach captures a higher-fidelity signal than using likes and matches alone. We present the production deployment of this system at Hinge via a multi-task two-tower model that jointly predicts three objectives with a shared embedding layer. User interaction sequences are encoded with a transformer architecture to capture temporal patterns. Our A/B test results demonstrate a +2.8% improvement in number of conversations and a +1.4% increase in exchange coverage.

Building a Production Shopping Agent at Scale [Industry]

Chen Luo (Amazon); Jason Choi (Amazon); Ziwei Dong (Amazon); Rahul Dua (Amazon); Cong Xu (Amazon); Xuejing Lei (Amazon); Yuchen Yan (Amazon); Xin Zhang (Amazon); Josef Valvoda (Amazon); Gaurang Sinkar (Amazon); Binit Jha (Amazon); Yi Liu (Amazon); Monica Cheng (Amazon)

Deploying a conversational shopping agent at production scale remains challenging despite rapid advances in large language models. Unlike research prototypes, production systems must satisfy strict requirements on accuracy, latency, reliability, and cost while serving millions of customers. We share our year-long journey of building a production shopping agent at scale and show that combining agentic reasoning with production-aware retrieval, tool orchestration, and system optimization enables a single shopping agent to support product discovery, shopping question answering, and agentic actions under real-world traffic. Our experience shows that LLM-based agents can be reliably operated at Amazon scale and provides practical design principles for industrial search and recommendation systems.

Probe-then-Plan: Environment-Aware Planning for Industrial E-commerce Search [Industry]

Mengxiang Chen (JD.com); Zhouwei Zhai (JD.com); Jin Li (JD.com)

Modern e-commerce search is evolving from simple keyword matching to resolving complex user intents. While large language models (LLMs) offer powerful reasoning capabilities, existing LLM-based search paradigms suffer from a fundamental blindness-latency dilemma: query rewriting methods are agnostic to retrieval tool capabilities and real-time inventory states, resulting in invalid plans; conversely, deep search agent approaches initially plan without environment awareness, then rely on iterative tool calls and reflection to perceive and correct failures, leading to seconds of latency, incompatible with the sub-second budget for the planning module in industrial e-commerce search. To resolve this conflict, we propose Environment-Aware Search Planning (EASP), a novel paradigm that reformulates search planning as a dynamic reasoning process grounded in environmental reality. EASP introduces a Probe-then-Plan mechanism: a lightweight Retrieval Probe first exposes the retrieval snapshot, enabling the Planner to diagnose execution gaps and generate grounded search plans. Our methodology unfolds in three stages: (1) Offline Data Synthesis: The Teacher Agent synthesizes diverse, execution-validated plans by diagnosing the retrieval environment exposed by the Retrieval Probe. (2) Planner Training and Alignment: The Planner is initialized via supervised fine-tuning (SFT) on the offline dataset to internalize the Teacher's diagnostic capabilities, followed by alignment with business outcomes (conversion rate) through reinforcement learning. (3) Adaptive Online Serving: A complexity-aware routing mechanism selectively activates the planning pipeline only for complex queries, ensuring optimal resource allocation.Extensive offline evaluations and online A/B testing on JD.com demonstrate that EASP significantly improves relevant recall and achieves substantial lifts in UCVR and GMV. EASP has been successfully deployed in JD.com's AI-Search system.

Policy-Guided RAG: Enforcing Verbatim and Controlled Synthesis [Industry]

Mina Naghash Asadi (Services Australia); Leila Tavakoli (Services Australia); Mustafa Bilgrami (Services Australia)

Retrieval-Augmented Generation (RAG) typically assumes retrieved text can be freely paraphrased, but this fails in regulated domains where some passages must be quoted verbatim. We propose a policy-guided transformation pattern for RAG, where retrieved segments carry sensitivity metadata and a lightweight router selects a transformation mode (Verbatim, Combined, or Synthesis). Quote-aware templates enforce constraints, and answer-side auditing verifies the final output using response--sensitive matching and a two-pass check for unquoted sensitive content. Experiments on an organisational corpus with simulated constraints show clear compliance--utility trade-offs. Compared to a Standard RAG baseline under the same retrieved context, policy-guided Combined substantially reduces verbatim exposure risk (VVR) across model families and generally reduces paraphrase-style sensitive reuse (PSR), while near-verbatim proximity can persist when sensitive and non-sensitive evidence are mixed. Under Verbatim, exposure decreases further, but some models pay a coverage cost via abstention; Synthesis preserves utility when sensitive evidence is removed. We summarise these outcomes (with emphasis on Combined) in a compact metric panel and distil engineering lessons for deploying auditable transformation governance in real RAG systems. We release the full evaluation protocol (routing rules, prompt templates, and compliance metrics) to enable replication on other corpora with similar policy constraints.

When Search Is Not Enough: Ranking Content for Query-Less Browse Surfaces [Industry]

Shreya Mahapatra (Adobe); Diksha Bhardwaj (Adobe); Anandita Chopra (Adobe); Tracy Holloway King (Adobe)

Adobe Express is a content creation platform that enables users to easily create and customize visual content. Browse categories, such as Flyers, play a critical role in content discovery, serving as influential entry points where users explore templates without issuing explicit search queries. These browse categories are organized into collections, which represent curated and themed groups of content. Effectively ranking content within these collections is crucial. A simple formulation of using a collection name or category label as a search query may not adequately capture all aspects of a collection. We developed a machine learned browse ranker based on browse session data. Through offline experiments, online A/B testing, and production deployment, our results demonstrate that machine-learned browse ranking produces rankings that differ from collection-name-as-query approaches and that these differences result in measurable improvements to user engagement and product success metrics.

Scaling Search Relevance: Augmenting App Store Ranking with LLM-Generated Judgments [Industry]

Evangelia Christakopoulou (Apple); Vivekkumar Patel (Apple); Hemanth Velaga (Apple); Sandip Gaikwad (Apple)

Large-scale commercial search systems optimize for relevance to drive successful sessions that help users find what they are looking for. To maximize relevance, we leverage two complementary objectives: behavioral relevance (results users tend to click or download) and textual relevance (a result's semantic fit to the query). A persistent challenge is the scarcity of expert-provided textual relevance labels relative to abundant behavioral relevance labels. We first address this by systematically evaluating LLM configurations, finding that a specialized, fine-tuned model significantly outperforms a much larger pre-trained one in providing highly relevant labels. Using this model as a force multiplier for human annotation, we generate millions of textual relevance labels to overcome the data scarcity. We show that augmenting our production ranker with these textual relevance labels leads to a significant outward shift of the Pareto frontier: offline NDCG improves for behavioral relevance while simultaneously increasing for textual relevance. These offline gains were validated by a worldwide A/B test on the App Store ranker, which demonstrated a statistically significant +0.24% increase in conversion rate, with the most substantial performance gains occurring in tail queries, where the new textual relevance labels provide a robust signal in the absence of reliable behavioral relevance labels.

Contrastive Collaborative Filtering · *Room 110 · Thursday 13:30–15:00*

Revisiting Collaborative Filtering by Unleashing the Power of Similarity [Full]

Mingyang Li (Beijing University of Posts and Telecommunications); Xinlang Yue (Kuaishou Technology); Chen Chen (Kuaishou Technology); Muyang Li (Kuaishou Technology); Yongqi Liu (Kuaishou Technology); Kaiqiao Zhan (Kuaishou Technology)

While deep learning-based approaches dominate the research on collaborative filtering (CF) in recommender systems, recent studies have re-examined this paradigm and found that deep models do not consistently outperform traditional CF alternatives on standard evaluation metrics. Notably, similarity-based methods exhibit substantial advantages. Inspired by these observations, we demonstrate that the performance of a minimalist similarity-based model can match or exceed leading CF approaches, highlighting its significant potential. Despite this, the expressiveness of high-performing similarity-based models is fundamentally confined to pairwise interactions within immediate neighborhoods, overlooking the combinatorial preference patterns in high-order connections. To overcome these limitations and fully unleash the potential of similarity-based models, we propose a novel bidirectional extension framework that systematically generalizes the modeling of similarity to group-level and multi-hop perspectives. However, directly computing these generalized similarities is prohibitively expensive due to their combinatorial complexity. To address this, we develop a MinHash-based estimation algorithm, which ensures the feasibility of the computation and reduces the complexity from exponential to linear. Extensive experiments on three public datasets show that our

framework achieves competitive or superior performance over strong baselines, yielding relative improvements ranging from 4\% to 11\% on key metrics, underscoring the enduring power of similarity-based methods in recommender systems.

Contrastive Flow Matching for Collaborative Filtering [Full]

Wangyu Jin (Soochow University); Jiansheng Qian (Soochow University); Wenwen Xia (Soochow University); Hongliang He (Soochow University); Guanfeng Liu (Macquarie University); Pengpeng Zhao (Soochow University)

Flow Matching (FM) has recently been introduced into generative collaborative filtering due to its potential to capture user preferences. However, existing FM-based methods are often trained with only the standard regression objective in flow matching and lack explicit inter-user supervision. This can drive the model toward an averaged preference pattern and generate indistinguishable predictions across users. A natural remedy is to introduce contrastive learning for inter-user separation, yet naive designs face two challenges: (1) In FM-based collaborative filtering, the output is a user's predicted interaction scores over items. Applying contrastive learning directly on this output forces different users to be separated even when they share the same preferences, leading to false repulsion. (2) Flow matching is highly sensitive to its input, so input-level perturbations for view construction can change the generated outcome, making positive pairs semantically mismatched. To address these issues, we propose CoFlowCF, a contrastive flow matching framework for collaborative filtering. Specifically, CoFlowCF introduces hidden contrastive learning by applying inter-user constraints before the flow model's final outputs. We decompose the flow model into a flow encoder and a prediction head, and apply contrastive regularization to the encoder output representations to avoid false repulsion in the prediction space. To build semantically consistent views, we keep the input unchanged and use independent dropout inside the encoder instead of input-level perturbations. Extensive experiments and ablation studies on multiple real-world datasets demonstrate that CoFlowCF consistently improves recommendation performance and robustness. The code is available at https://github.com/DevByQian/CoFlowCF.

Adaptive Rich-kernelized Contrastive Learning for Capacity Enhancement in Collaborative Filtering [Full]

Jie Yang (University of Melbourne); Ling Luo (University of Melbourne); Nestor Cabello (University of Melbourne); Lars Kulik (University of Melbourne)

Recent research has shown that single-vector embedding retrieval models face a fundamental bottleneck: the finite dimensionality of single-vector representations limits their capacity to represent arbitrary top-k relevant item combinations, even with perfect training. This inherent bottleneck substantially limits the expressive capacity of such models and reduces their ability to capture complex user-item interaction patterns. To overcome this bottleneck, we propose Adaptive Rich-kernelized Contrastive Learning (ARC), which enhances model expressiveness while maintaining the computational efficiency of single-vector retrieval. ARC replaces the fixed inner product with a learnable spherical kernel family parameterized by a truncated Gegenbauer expansion, thereby increasing the model's effective dimensionality while preserving single-vector efficiency and low-pass inductive bias for generalization. Specifically, we construct a positive-definite kernel on the unit sphere using a positive combination of Gegenbauer polynomials, and adopt a contrastive learning objective to jointly learn the polynomials weights and user–item embeddings. Through data-driven optimization, the adaptive kernel induces a more expressive representation space, enabling the model to better capture complex preferences. From a theoretical perspective, the learned kernel implicitly maps embeddings into a higher-dimensional Reproducing Kernel Hilbert Space, allowing the model to capture a broader range of top-k item combinations before reaching the geometric limit imposed by the embedding dimensionality. Consequently, the proposed ARC framework effectively alleviates the inherent representational bottleneck in traditional single-vector embedding models. Extensive experiments on four real-world datasets demonstrate the effectiveness of our method, showing consistent improvements over strong baselines and state-of-the-art models.

ACE: Semantically-Grounded Graph Alignment via Affective Contrastive Learning [Full]

Potito Aghilar (Politecnico di Bari); Sabino Roccotelli (Politecnico di Bari); Vito Walter Anelli (Politecnico di Bari); Alejandro Bellogin (Universidad Autónoma de Madrid); Michelantonio Trizio (Wideverse); Tommaso Di Noia (Politecnico di Bari)

Graph Contrastive Learning (GCL) methods for recommendation learn representations by propagating signals over user-item interaction graphs. However, modeling these graphs with homogeneous edges can lead to semantic-structural misalignment, where information is exchanged between structurally adjacent but semantically dissimilar items, adversely affecting retrieval quality. Existing solutions typically rely on auxiliary encoders or additional supervision, increasing model complexity and training cost. We propose ACE (Affective Contrastive Embeddings), a framework that improves representation alignment by incorporating affective semantics into contrastive learning. ACE encodes the affective dimensions of valence and arousal as a topological prior, encouraging consistency between learned embeddings and an affective semantic space distilled from large language models. To operationalize this alignment, we introduce a Semantically Weighted Noise Contrastive Estimation (SW-NCE) loss that modulates contrastive gradients according to users' affective preferences. Experiments on Amazon, Last.fm, and SiTunes demonstrate that ACE consistently improves top-K retrieval performance over 11 baselines while reducing computational overhead. These results indicate that affective geometric alignment is an effective and efficient mechanism for enhancing graph-based retrieval models.

DIAURec: Dual-Intent Space Representation Optimization for Recommendation [Full]

Yu Zhang (Anhui University); Yiwen Zhang (Anhui University); Yi Zhang (Anhui University); Lei Sang (Anhui University)

General recommender systems deliver personalized services by learning user and item representations, with the central challenge being how to capture latent user preferences. However, representations derived from sparse interactions often fail to comprehensively characterize user behaviors, thereby limiting recommendation effectiveness. Recent studies attempt to enhance user representations through sophisticated modeling strategies (e.g., intent or language modeling). Nevertheless, most works primarily concentrate on model interpretability instead of representation optimization. This imbalance has led to limited progress, as representation optimization is crucial for recommendation quality by promoting the affinity between users and their interacted items in the feature space, yet remains largely overlooked. To overcome these limitations, we propose DIAURec, a novel representation learning framework that unifies intent and language modeling for recommendation. DIAURec reconstructs representations based on the prototype and distribution intent spaces formed by collaborative and language signals. Furthermore, we design a comprehensive representation optimization strategy. Specifically, we adopts alignment and uniformity as the primary optimization objectives, and incorporates both coarse- and fine-grained matching to achieve effective alignment across different spaces, thereby enhancing representational consistency. Additionally, we further introduce intra-space and interaction regularization to enhance model robustness and prevent representation collapse in reconstructed space representation. Experiments on three public datasets against fifteen baseline methods show that DIAURec consistently outperforms state-of-the-art baselines, fully validating its effectiveness and superiority.

AsarRec: Adaptive Sequential Augmentation for Robust Self-supervised Sequential Recommendation [Full]

Kaike Zhang (Institute of Computing Technology, Chinese Academy of Sciences); Qi Cao (Institute of Computing Technology, Chinese Academy of Sciences); Fei Sun (Institute of Computing Technology, Chinese Academy of Sciences); Liu Xinran (Institute of Computing Technology, Chinese Academy of Sciences); Huawei Shen (Institute of Computing Technology, Chinese Academy of Sciences); Xueqi Cheng (Institute of Computing Technology, Chinese Academy of Sciences)

Real-world user behaviors are often noisy due to factors such as human errors, uncertainty, and behavioral ambiguity, which can lead to degraded recommendation performance. To address this issue, recent approaches widely adopt self-supervised learning (SSL), particularly contrastive learning, by generating perturbed views of user interaction sequences and maximizing their mutual information to improve model robustness. However, these methods heavily rely on their pre-defined static augmentation strategies (where the augmentation type remains fixed once chosen) to construct augmented views, leading to two critical challenges: (1) the optimal augmentation type can vary significantly across different scenarios; (2) inappropriate augmentations may even degrade recommendation performance, limiting the effectiveness of SSL. To overcome these limitations, we propose an adaptive augmentation framework. We first unify existing basic augmentation operations into a unified formulation via structured transformation matrices. Building on this formulation, we introduce AsarRec, an Adaptive Sequential Augmentation for Robust Sequential Recommendation. To enable stable end-to-end optimization of discrete and strongly constrained augmentations, AsarRec learns to generate transformation matrices by encoding user sequences into probabilistic transition matrices and projecting them into hard semi-doubly stochastic matrices via a differentiable Semi-Sinkhorn algorithm. To ensure that the learned augmentations benefit downstream performance, we jointly optimize three objectives: diversity (encouraging distinct views), semantic invariance (preserving semantic consistency among views), and informativeness (identifying augmentations most beneficial to recommendation). Extensive experiments on four benchmarks under varying noise levels validate the effectiveness of AsarRec, demonstrating its superior robustness and consistent improvements.

Knowledge Graph Reasoning · *Eureka 1* · *Thursday 13:30–15:00*

MECI: Multi-Element Collaborative Interaction for Multimodal Entity Linking [Full]

Peng Jie (National University of Defense Technology); Yongxue Shan (National University of Defense Technology); Yongfu Zha (National University of Defense Technology); Xiaodong Wang (National University of Defense Technology)

Multimodal Entity Linking (MEL) aims to disambiguate mentions in multimodal contexts by grounding them to specific entities in a knowledge base. A pivotal challenge in MEL is capturing multi-level correspondences: the semantic consistency between mention-entity pairs and the complementary correlations across modalities. However, existing methods often suffer from element dominance due to their reliance on coupled interactions or coarse global aggregations. In response, we propose the Multi-Element Collaborative Interaction (MECI) framework. First, to capture multi-element mention-entity correspondences, we develop a Multi-view

Experts Network that leverages a "divide-and-conquer" strategy for decoupled feature learning to mitigate element dominance, supported by a KL-guided routing mechanism that governs expert specialization and collaboration. Furthermore, to model cross-modal complementary correlations, we propose a Hierarchical Multimodal Interaction Module, where a dynamic modality-aware weighting network refines interactions across hierarchical semantic levels, thereby integrating multi-granular evidence to counteract element dominance. Finally, we incorporate a generative semantic refinement stage that utilizes large language models for zero-shot re-ranking. Extensive experiments on WikiDiverse, RichpediaMEL, and WikiMEL show that MECI consistently outperforms state-of-the-art baselines, improving Hits@1 by 1.95%, 7.30%, and 2.31%, respectively.

SSR: Structured Subgraph Retrieval for Temporal Knowledge Graph Question Answering with LLMs [Full]

Ying Zhang (Nankai University); Li Zhang (Nankai University); Wenya Guo (Nankai University); Shilong Ping (Nankai University); Xinying Qian (Nankai University)

Temporal Knowledge Graph Question Answering (TKGQA) aims to answer natural language questions based on quadruple facts stored in Temporal Knowledge Graphs (TKGs). Recent studies have integrated Large Language Models (LLMs) to handle the complex semantic reasoning required by temporal questions. They typically linearize TKG quadruples into plain text, reducing TKGQA to a top-n text retrieval and context-based question answering problem. However, this textualization process destroys the original quadruple structure and entangles structural and temporal information with text semantics. Moreover, selecting top-n facts based on semantic similarity inevitably introduces a large amount of irrelevant noise into the LLM input, which degrades reasoning performance. To address these limitations, we propose SSR, a Structured Subgraph Retrieval framework for TKGQA with LLMs. In SSR, a Temporal Question Parser is first introduced to extract structured subgraph patterns and temporal constraints from the input questions, leveraging background context obtained via text retrieval. The Subgraph Retrieval module is then applied to directly filter a relevant subgraph from the TKG that contains the facts required to answer the question. The retrieved subgraph is further temporally compressed and subsequently incorporated into the LLM context for final question answering. Experimental results on two benchmark datasets demonstrate that SSR consistently outperforms strong baselines by a clear margin, achieving state-of-the-art performance. Our code is available at https://github.com/zhangli-coding/SSR.

Exploration-and-Thinking: Agentic Reasoning over Knowledge Graphs via an LLM-RL Synergized Framework [Full]

Yi Xia (State Key Laboratory of Mathematical Engineering and Advanced Computing); Gang Zhou (State Key Laboratory of Mathematical Engineering and Advanced Computing); Jing Chen (State Key Laboratory of Mathematical Engineering and Advanced Computing); Xiaohui Chen (State Key Laboratory of Mathematical Engineering and Advanced Computing); Qinlong Fan (State Key Laboratory of Mathematical Engineering and Advanced Computing); Shunhang Li (State Key Laboratory of Mathematical Engineering and Advanced Computing)

While Knowledge Graphs (KGs) can ground Large Language Models (LLMs) in factual knowledge, existing LLM-KG integration methods for complex reasoning are plagued by computational inefficiency and semantic inconsistency. The tight coupling of LLM inference and KG traversal leads to prohibitive costs, while spurious reasoning paths often misguide learning-based agents, causing reward hacking. To this end, we propose EAT (Exploration-and-Thinking), a novel agentic framework that synergizes LLMs with Reinforcement Learning (RL) for efficient and faithful reasoning on KGs. EAT's core innovations are twofold: (1) an adaptive retrieval-augmented generation mechanism that decouples language comprehension from structured exploration, dramatically improving efficiency; and (2) an LLM-guided reward shaping strategy that explicitly penalizes semantically inconsistent paths and promotes logically valid trajectories grounded in the KG. Extensive experiments on benchmarks like WebQSP, CWQ, and GrailQA show that EAT achieves state-of-the-art performance. Notably, it surpasses the reasoning capability of GPT-4o while utilizing a significantly smaller 8B-parameter LLM.

Multi-Faceted Continual Knowledge Graph Embedding for Semantic-Aware Link Prediction [Full]

Jing Qi (Hangzhou Dianzi University); Yuxiang Wang (Hangzhou Dianzi University); Zhiyuan Yu (Hangzhou Dianzi University); Xiaoliang Xu (Hangzhou Dianzi University); Yuanshi Zheng (Xidian University); Tianxing Wu (Southeast University)

Continual Knowledge Graph Embedding (CKGE) aims to continually learn embeddings for new knowledge, i.e., entities and relations, while retaining previously acquired knowledge. Most existing CKGE methods mitigate catastrophic forgetting via regularization or replaying old knowledge. They conflate new and old knowledge of an entity within the same embedding space to seek a balance between them. However, entities inherently exhibit multi-faceted semantics that evolve dynamically as their relational contexts change over time. A shared embedding fails to capture and distinguish these temporal semantic variations, degrading lifelong link prediction accuracy across snapshots. To address this, we propose a Multi-Faceted CKGE framework (MF-CKGE) for semantic-aware link prediction. During offline learning, MF-CKGE separates temporal old and new knowledge into distinct embedding spaces to prevent knowledge entanglement and employs semantic decoupling to reduce semantic redundancy, thereby improving space efficiency. During online inference, MF-CKGE adaptively identifies semantically query-relevant entity embeddings by quantifying their semantic importance, reducing interference from query-irrelevant noise. Experiments on eight datasets show that MF-CKGE achieves an average (maximum) improvement of 1.7% (2.7%) and 1.4% (3.8%) in MRR and Hits@10, respectively, over the best baseline. Our source code and datasets are available at: https://anonymous.4open.science/r/MF-CKGE-04E5.

Towards Conflict-aware Selective Knowledge Unlearning for Continual Few-shot Knowledge Graph Completion [Full]

Junlin Zhu (University of Electronic Science and Technology of China); Bo Fu (University of Electronic Science and Technology of China); Guiduo Duan (University of Electronic Science and Technology of China)

Few-shot knowledge graph completion (FKGC) aims to enhance knowledge reasoning for newly emerging relations using only a limited number of supporting triples. Given the inherent growth and evolution of knowledge graphs (KGs), FKGC faces greater challenges in continual learning (CL) scenarios, as models tend to rely more heavily on historical knowledge in few-shot settings. Existing studies have mainly focused on preserving old knowledge to mitigate catastrophic forgetting. However, these strategies often overlook that not all knowledge is worth retaining. Some prior knowledge may conflict with newly acquired knowledge, either structurally or semantically, distorting entity representations and relation patterns and weakening few-shot inference for new relations. To bridge this research gap, we propose a novel FKGC framework named CSKU, which achieves dynamic knowledge adaptation by selectively unlearning structural and semantic conflicts. Specifically, to preserve entity representation consistency, we design a graph-structured dynamic progressive unlearning mechanism that adaptively mitigates structural conflicts during training. Meanwhile, we introduce a semantic-guided selective unlearning mechanism that dynamically attenuates the impact of conflicting relations. Experimental results on multiple real-world FKGC datasets demonstrate that CSKU significantly outperforms state-of-the-art baselines in both forgetting mitigation and few-shot completion, validating the effectiveness of the proposed selective knowledge unlearning strategy. Our code and datasets are publicly available. https://anonymous.4open.science/r/CSKU.

GCA-KBQA: A Step-Wise Logical Form Generation Approach for KBQA with Knowledge-Assisted Calibration [Full]

Ranran Bu (Shanghai Jiao Tong University); Jian Cao (Shanghai Jiao Tong University); Jianqi Gao (Shanghai University); Jinghua Tang (Shanghai Jiao Tong University); Shiyou Qian (Shanghai Jiao Tong University); Hongming Cai (Shanghai Jiao Tong University)

Knowledge base question answering (KBQA) aims to answer natural language questions using large-scale knowledge bases (KBs). Among various KBQA approaches, semantic parsing-based (SP-based) methods have demonstrated strong effectiveness by generating concise logical forms (LFs) that capture complex subgraph structures and semantic information. Recent research suggests that integrating large language models (LLMs) with SP can achieve significant improvements in the performance and efficiency of KBQA by facilitating the direct generation of LFs with minimal retrieval. However, generating complete LFs with LLMs continues to pose a challenge due to the complexity of the required graph structures and constraints, leading to the significant issue of non-executability. To address these challenges, we propose GCA-KBQA, a step-wise fine-tuned LLM-based framework that employs hop-wise generation, knowledge-assisted calibration, and path-level assembly to construct complete LFs for KBQA. Specifically, we decompose the complex SP process into manageable steps: first, we iteratively generate LFs for each topic entity one hop at a time using a fine-tuned LLM, leveraging KB knowledge to calibrate intermediate outputs and mitigate error propagation. Subsequently, we guide the LLM in assembling path-level LFs from different topic entities, resulting in optimized final LF. We evaluate the proposed method on four KBQA benchmarks spanning two distinct KBs, demonstrating its superior performance compared to state-of-the-art baselines. The code is available at https://github.com/pvfeldt/GCA-KBQA.

Retrieval and Understanding of Visual Documents and Video · *Eureka 2 · Thursday 13:30–15:00*

ProEchoMem: Enhancing Long Video Understanding via Multi-Trace Probe-Echo Memory [Full]

Derong Xu (University of Science and Technology of China); Yanxin Chen (City University of Hong Kong); Wanyu Wang (City University of Hong Kong); Pengyue Jia (City University of Hong Kong); Chao Zhang (University of Science and Technology of China); Maolin Wang (City University of Hong Kong); Yiqi Wang (Michigan State University); Jipeng Qiang (Yangzhou University); Xuetao Wei (Southern University of Science and Technology); Hongzhi Yin (University of Queensland); Tong Xu (University of Science and Technology of China); Xiangyu Zhao (City University of Hong Kong)

Large vision-language models (LVLMs) have shown significant progress in video understanding, but they struggle to scale to long videos due to limited context windows. Existing methods reduce input dimensionality via frame sampling and feature compression, yet discard details and incur high computational cost for post-training. In contrast, retrieval-augmented generation (RAG) that indexes long videos for query retrieval and memory-based methods that maintain evolving long-term stores, offer a lighter and deployment-friendly solution. Nevertheless, they rely on shallow retrieval that selects only top-ranked segments and fails to integrate information across multiple relevant video episodes. Inspired by Multiple-Trace Theory in cognitive psychology, we revisit long video

understanding from a probe-echo perspective, in which human episodic memories are activated and integrated in parallel. Building on this insight, we propose ProEchoMem, a cognitive-inspired framework that simulates the probe-echo mechanism: (1) Incremental Episodic Memory Construction builds structured knowledge graphs from video streams; (2) Probe-Driven Memory Activation generates probe signals from user queries to activate all stored traces simultaneously; (3) Memory Echo Synthesis integrates activated traces into a coherent and structured memory echo. Experiments on LongerVideos, LVBench, and cross-domain settings demonstrate the effectiveness of ProEchoMem, with multi-trace probing achieving up to 14.2% higher relevance and ablation studies validating the contribution of each module. The code is available at https://github.com/Applied-Machine-Learning-Lab/SIGIR26_ProEchoMem

ReAlign: Optimizing the Visual Document Retriever with Reasoning-Guided Fine-Grained Alignment [Full]

Hao Yang (Northeastern University); Yifan Ji (Northeastern University); Zhipeng Xu (Northeastern University); Zhenghao Liu (Northeastern University); Yukun Yan (Tsinghua University); Zulong Chen (Alibaba Group); Shuo Wang (Tsinghua University); Yu Gu (Northeastern University); Ge Yu (Northeastern University)

Visual document retrieval aims to retrieve a set of document pages relevant to a query from visually rich collections. Existing methods often employ Vision-Language Models (VLMs) to encode queries and visual pages into a shared embedding space, which is then optimized via contrastive training. However, during visual document representation, localized evidence is usually scattered across complex document layouts, making it difficult for retrieval models to capture crucial cues for effective embedding learning. In this paper, we propose Reasoning-Guided Alignment (ReAlign), a method that enhances visual document retrieval by leveraging the reasoning capability of VLMs to provide fine-grained visual document descriptions as supervision signals for training. Specifically, ReAlign employs a superior VLM to identify query-related regions on a page and then generates a query-aware description grounding the cropped visual regions. The retriever is then trained using these region-focused descriptions to align the semantics between queries and visual documents by encouraging the document ranking distribution induced by the region-focused descriptions to match that induced by the original query. Experiments on diverse visually rich document retrieval benchmarks demonstrate that ReAlign consistently improves visual document retrieval performance on both in-domain and out-of-domain datasets, achieving up to 2% relative improvements. Moreover, the advantages of ReAlign generalize across different VLM backbones by guiding models to better focus their attention on critical visual cues for document representation. All code and datasets are available at https://github.com/NEUIR/ReAlign.

Attention Grounded Enhancement for Visual Document Retrieval [Full]

Wanqing Cui (Alibaba Group); Wei Huang (University of Chinese Academy of Sciences); Yazhi Guo (Alibaba Group); Yibo Hu (Alibaba Group); Meiguang Jin (Alibaba Group); Junfeng Ma (Alibaba Group); Keping Bi (University of Chinese Academy of Sciences)

Visual document retrieval requires understanding heterogeneous and multi-modal content to satisfy implicit information needs. Recent advances use screenshot-based document encoding with fine-grained late interaction to encode holistic information and capture nuanced alignments, significantly improving retrieval performance. However, retrievers are still trained with coarse global relevance labels, without revealing which regions support the match. As a result, retrievers tend to rely on surface-level cues and struggle to capture implicit semantic connections, hindering their ability to handle non-extractive queries. To improve fine-grained relevance modeling, we propose a Attention-Grounded REtriever Enhancement (AGREE) framework. AGREE leverages cross-modal attention from multimodal large language models (MLLMs) as proxy supervision to guide the retriever in identifying relevant document regions. Specifically, AGREE extracts attention maps from the MLLM that highlight which document regions are attended to based on the query. These attention scores serve as local, region-level relevance signals. During training, AGREE combines local signals with the global document-level relevance label to jointly optimize the retriever. This dual-level supervision enables the model to learn not only whether documents match, but also which content drives relevance. Experiments on the challenging visual document retrieval benchmark, ViDoRe V2, show that AGREE significantly outperforms the global-supervision-only baseline by 12.82% and 5.03% in terms of average nDCG@1 and nDCG@5. Quantitative and qualitative analyses further demonstrate that AGREE promotes deeper alignment between query terms and document regions, moving beyond surface-level matching toward more accurate and interpretable retrieval. Our code is available at: https://github.com/VickiCui/AGREE.

RegionSLM: Region-aware Question Answering on Document Screenshots [Full]

Chao Wang (CSIRO); Hehe Fan (Zhejiang University); Huichen Yang (CSIRO); Sarvnaz Karimi (CSIRO); Lina Yao (University of New South Wales); Yi Yang (Zhejiang University)

Real-world document question-answering that relies on screenshots, such as bills and forms, requires evidence that is often spatially localised and visually cluttered. However, most Screenshot Language Models (SLMs) encode the entire page holistically and rely on implicit attention to "find" relevant content, which limits both accuracy and efficiency. We present RegionSLM, a region-aware SLM designed to explicitly connect the question to its supporting regions. RegionSLM has two key components: (1) a patch-relevance router that learns a query–region relevance distribution, enabling the model to produce a box-free relevance prior at inference; and (2) Relevance-Guided Region Pooling (RGRP), a query-conditional attention–pooling module that aggregates dense features into a small set of region tokens, which preserves grounding signals while reducing computational overhead. To support training and evaluation, we further curate ReDoc, a region-supervised corpus with 105k documents and 350k question-answer pairs, obtained via a question-guided two-step filtering procedure. Extensive experiments on 12 datasets demonstrate that explicitly learning query–region relevance and pooling it into compact region tokens is an effective and practical recipe for document retrieval and understanding.

Good Ranks Follow Good Answers: Unsupervised Answer-Driven Reranking for Multimodal Document QA [Full]

Keyu Zhu (University of Science and Technology of China); Shuanghong Shen (Institute of Artificial Intelligence, Hefei Comprehensive National Science Center); Xianquan Wang (University of Science and Technology of China); Kai Zhang (University of Science and Technology of China); Shijin Wang (iFLYTEK, Co., Ltd); Qi Liu (University of Science and Technology of China); Zhenya Huang (University of Science and Technology of China)

Multimodal Document Question Answering (MDQA) systems commonly follow a retrieve-then-answer paradigm; however, the retrieval stage often introduces substantial noise, making an effective reranking component indispensable.Existing reranker training frameworks in MDQA rely predominantly on proxy supervision derived from human annotations or large language model (LLM) outputs, which are frequently noisy and, more critically, misaligned with downstream answer quality.To overcome this limitation, we propose \textbf{AD-Reranker}, a novel framework that shifts reranker training from proxy imitation to \textbf{answer-driven} utility optimization.Specifically, we reformulate the reranker as an environment-grounded agent that interacts with a downstream reader, modeled as a deterministic environment.We further design a composite reward function that integrates answer correctness, thereby explicitly incentivizing ranking strategies aligned with downstream task performance.To optimize the agent, we adopt Group Relative Policy Optimization (GRPO), enabling stable and effective group-wise policy learning.Empirical results demonstrate that AD-Reranker achieves superior reranking quality and an optimal accuracy-efficiency trade-off.When integrated into standard MDQA pipelines, AD-Reranker improves accuracy by 1.9\%–5.0\% while reducing the reader's context token consumption by 15\%–52\%, providing strong evidence for the effectiveness of answer-driven reranker training.

DocQAC: Adaptive Trie-Guided Decoding for Effective In-Document Query Auto-Completion [Full]

Rahul Mehta (Microsoft Corporation); Kavin R V (Indian Institute of Technology Kharagpur); Indrajit Pal (Independent); Tushar Abhishek (Microsoft Corporation); Pawan Goyal (Indian Institute of Technology Kharagpur); Manish Gupta (Microsoft Corporation)

Query auto-completion (QAC) has been widely studied in the context of web search, yet remains underexplored for in-document search, which we term DocQAC. DocQAC aims to enhance search productivity within long documents by helping users craft faster, more precise queries, even for complex or hard-to-spell terms. Unlike traditional WebQAC systems, DocQAC can leverage rich document context, having access not only to the partially typed user query and global historical queries, but also the content of the current document itself, and crucially, the document-specific history of user query interactions. To address this setting, we propose a novel adaptive trie-guided decoding framework that uses user query prefixes to softly steer language models toward high-quality completions. Our approach introduces an adaptive penalty mechanism with tunable hyperparameters, enabling a principled trade-off between model confidence and trie-based guidance. To efficiently incorporate document context, we explore retrieval-augmented generation (RAG) and lightweight contextual document signals such as titles, keyphrases, and summaries. When applied to encoder–decoder models like T5 and BART, our trie-guided framework outperforms strong baselines and even surpasses much larger instruction-tuned models such as LLaMA-3 and Phi-3 in seen-query settings. This demonstrates its practicality for real-world DocQAC system deployments, where efficiency and scalability are critical. We evaluate our method on a newly introduced DocQAC benchmark derived from ORCAS, enriched with query–document pairs. We make both the DocQAC dataset https://bit.ly/3IGEkbH and code https://github.com/rahcode7/DocQAC publicly available.

Applications and Low-Resource IR · *Eureka 3* ·

Thursday 13:30–15:00

Beyond Exposure Diversity: Debiasing News Consumption With Topic-Locality Calibration and Personalized Preview Nudges [Full]

Ruixuan Sun (University of Minnesota - Twin Cities); Matthew Zent (University of Minnesota - Twin Cities); Minzhu Zhao (University of Minnesota - Twin Cities); Xinyi Li (Northwestern University); Thanmayee Boyapati (University of Minnesota - Twin Cities); Joseph A. Konstan (University of Minnesota - Twin Cities)

In this study, we applied the "personalized diversity nudge framework" with the goal of expanding readers' reading coverage in terms of news locality (i.e., domestic and world news). We designed a novel topic-locality dual calibration algorithmic nudge and a large language model-based news

personalization presentation nudge, then launched a 5-week real-user study with 120 U.S. news readers on the news recommendation experiment platform POPROX. With interaction logs and survey responses, we found that algorithmic nudges can successfully increase exposure and consumption diversity, while the impact of LLM-based presentation nudges varied. Participant-level topic interest is a strong predictor of individual clicks, while highlighting the relevance of news articles to prior read articles outperforms generic topic-based and no personalization. We also demonstrate that longitudinal exposure to calibrated news may shift readers' reading habits to value a balanced news digest from both domestic and world articles. Our results provide direction for future work on nudging for diverse consumption in news recommendation systems.

RES-MR: Risk-Aware Reasoning for Explainable and Safe Medication Recommendation [Full]

Cong Wang (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences); Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science); Jin Li (University of Technology Sydney); Shoujin Wang (University of Technology Sydney); Yishuo Li (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences); Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science); Huilin Gu (University of Technology Sydney); Wenpeng Lu (Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences); Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science)

Medication recommendations seek to deliver personalized drug combinations based on patients' clinical information, where ensuring transparent decision-making and personalized safety are of paramount importance. However, existing approaches predominantly operate as "black boxes", offering limited interpretability and hindering clinician trust. Furthermore, they typically rely on fixed, patient-agnostic safety constraints, neglecting the heterogeneity of risk tolerance across individuals, which poses significant risks to vulnerable populations. To address these gaps, we propose RES-MR, a novel Risk-aware Reasoning framework for Explainable and Safe Medication Recommendation using LLMs. RES-MR follows a two-stage training paradigm: (1) an explainable clinical reasoning distillation stage to elicit diagnostic capabilities by fine-tuning on synthesized reasoning trajectories, grounded in patient-specific knowledge from graph retrieval; and (2) a risk-aware policy optimization stage to dynamically balance therapeutic efficacy with personalized safety. This stage employs disentangled drug factors to construct patient-specific reference prototypes that capture desired therapeutic drug features and risk features to avoid. These prototypes enable risk boundary calibration and guide the model toward safe recommendations via reward shaping and clipping. Extensive experiments on benchmark datasets demonstrate that RES-MR significantly outperforms state-of-the-art baselines in terms of accuracy, safety, and explainability. Our code and data are available at: https://github.com/wangcong2001/RES-MR.

Demand-Calibrated Facet Diversification with Deployable Soft Quotas for E-Commerce Search [Full]

Shenghui Xu (eBay Inc.); Jiang Yu (Trovar.ai)

E-commerce search rankers optimized for engagement or conversion often homogenize first-page results, under-serving ambiguous or exploratory queries. Prior diversification methods typically assume per-query intent/aspect supervision or heavy re-ranking, which is costly and brittle under production constraints. We proposed a supervision-free, deployable framework for demand-calibrated facet diversification. From independent holdout traffic, we estimate a cohort-level target facet distribution using session-attributed downstream engagement, capturing how much diversity users actually demand. At serving time, we steer the first-page slate toward this target with a lightweight soft-quota score adjustment over a short candidate prefix, preserving the base ranker's within-facet ordering and adding negligible latency. Large-scale online experiments on high-volume traffic show consistent improvements in revenue and engagement while reducing abandonment. We further find that delivered first-page facet mixes move closer to the engagement-derived target, indicating effectiveness of the proposed mechanism.

Improving Amharic Information Retrieval with Translative and Multi-Agent Debate Retrieval Augmented Generation

[Low-Resource]

Abel Jotie (Carnegie Mellon University Africa); Prasenjit Mitra (Carnegie Mellon University Africa)

Retrieval-augmented generation (RAG) has been used to improve the accuracy and transparency of outputs produced by large language models (LLMs) by integrating external knowledge; however, applying RAG to low-resource languages presents unique challenges, including poor embedding representations, low retrieval quality, and semantic gaps caused by the scarcity of digital documents. In this proposal, we address these challenges for a selected low-resource language, Amharic, by using translative and debate-based RAG techniques to improve retrieval and reasoning. The proposal outlines the key problems and research gaps in applying RAG to low-resource languages and introduces a method to enhance RAG performance for Amharic. Additionally, we introduce the first comprehensive Amharic Retrieval-Augmented Generation Benchmark (ARGB), designed to capture grammatical, cultural, and writing-system-specific constraints of the Amharic language. ARGB evaluates not only retrieval and generation quality, but also noise robustness, counterfactual robustness, negative rejection, and multi-source information integration, providing a holistic assessment of RAG capabilities. This work aims to improve the reliability, accessibility, and inclusiveness of AI systems for Amharic speakers while providing a scalable framework for other low-resource languages. Current progress on the code and benchmark can be found on this GitHub link: \hrefhttps://github.com/AbelAlemu155/AmharicRAG link.

Detectability and Evaluation Risks of LLM-Generated Answers in Low-Resource Swahili Agricultural Information Retrieval

[Low-Resource]

Theofrida J. Maginga (Sokoine University of Agriculture); Farian S. Ishengoma (Sokoine University of Agriculture)

The deployment of information retrieval (IR) systems in agriculture remains particularly challenging in low-resource environments, where limited access to expert knowledge, data resources, and digital infrastructure constrains effective information access. Recent advances in large language models (LLMs) have enabled conversational interfaces that generate direct answers to user queries, increasingly functioning as retrieval outputs in IR systems. However, little is known about how such generated answers behave in low-resource language settings. This study examines Swahili maize-related question answering as a low-resource agricultural IR task and investigates whether LLM-generated responses are detectably different from human-authored answers. Using a dataset of 127 questions with responses from human experts, ChatGPT, and Gemini, we analyze textual characteristics including answer length, lexical diversity, and TF-IDF cosine similarity. A multi-class TF-IDF and logistic regression classifier is then used to evaluate source detectability. Results show that LLM-generated answers are systematically longer, less lexically diverse, and exhibit moderate similarity to human responses. The classifier achieves 0.98 accuracy, indicating that generated responses exhibit consistent and learnable patterns. These findings highlight detectability as a useful diagnostic signal for evaluating generative IR systems, particularly in low-resource environments where ground-truth evaluation is limited.

Building Foundations for Information Retrieval in Extremely Low-Resource Ethnic Languages: Evidence from Tanzania

[Low-Resource]

Joseph P. Telemala (Sokoine University of Agriculture); Onesmo S. Nyinondi (Sokoine University of Agriculture); Neema N. Lyimo (Sokoine University of Agriculture); Baraka W. Nyamtiga (Sokoine University of Agriculture); Farian S. Ishengoma (Sokoine University of Agriculture); Wema L. Msigwa (National Kiswahili Council)

Despite major advances in information retrieval (IR) and language technologies, speakers of extremely low-resource ethnic community languages (ECLs) remain largely excluded from digital information access. With Tanzania having over 120 languages, yet most IR systems and digital assistants primarily support English and, to a limited extent, Kiswahili. This disproportionately affects rural populations (over 65%), where ECLs are used alongside Kiswahili in daily life. This paper presents the ongoing Ethnic Voices initiative, which addresses this gap by building foundational text and speech resources explicitly relevant to IR research in extremely low-resource multilingual settings. The project targets six ECLs in Tanzania, with current data collection completed for Hehe and Luguru. Data collection from the two ECLs has produced over 800 Kiswahili-ECL parallel sentences per language and over 95 hours of speech. Preliminary survey results show strong demand for multilingual and voice-based access to information. It further identifies language barriers as a primary limitation of existing English-centric IR systems. This also paper discusses the IR implications of these findings and outlines directions toward inclusive retrieval and evaluation in extremely low-resource languages. Beyond resource creation, we position the collected dataset as IR-enabling infrastructure that supports the formulation of retrieval tasks, the development of baseline systems, and the design of evaluation protocols in benchmark-scarce, extremely low-resource multilingual settings.

Sparse Retrieval · *Courtyard 1+2 · Thursday 13:30–15:00*

From Tokens to Concepts: Leveraging SAE for SPLADE [Full]

Yuxuan Zong (CNRS, Sorbonne Université, ISIR); Mathias Vast (CNRS, Sorbonne Université, ISIR); Basile Van Cooten (ChapsVision); Laure Soulier (CNRS, Sorbonne Université, ISIR); Benjamin Piwowarski (CNRS)

Learned Sparse IR models, such as SPLADE, offer an excellent efficiency-effectiveness tradeoff. However, they rely on the underlying backbone vocabulary, which might hinder performance (polysemicity and synonymy) and pose a challenge for multi-lingual and multi-modal usages. To solve this limitation, we propose to replace the backbone vocabulary with a latent space of semantic concepts learned using Sparse Auto-Encoders (SAE). Throughout this paper, we study the compatibility of these 2 concepts, explore training approaches, and analyze the differences between our SAE-SPLADE model and traditional SPLADE models. Our experiments demonstrate that SAE-SPLADE achieves retrieval performance comparable to SPLADE on both in-domain and out-of-domain tasks while offering improved efficiency.

Why Advanced Encoders Lag on Sparse Retrieval? The Answer and an Approach to Bridging Vocabulary Gaps [Full]

Zhichao Geng (Amazon Web Service); Yang Yang (Amazon Web Service)

While advanced foundation models like ModernBERT significantly outperform older architectures in dense retrieval, they surprisingly lag behind the aging BERT-base baseline in learned sparse retrieval (LSR). We identify the root cause as the Vocabulary Gap : modern tokenizers utilize raw, case-sensitive vocabularies designed for lossless reconstruction, which map single semantic units to redundant surface forms, wasting model capacity on morphological noise and hindering lexical matching. We formalize this intuition through a theoretical framework, demonstrating that appropriate vocabulary coarse-graining can tighten the generalization bounds by reducing complexity of the hypothesis class, provided that semantic integrity is preserved. To resolve this, we propose Vocabulary Transfer (VT), a model-agnostic framework that migrates advanced encoders to sparse-friendly, normalized vocabularies with minimal computational cost. VT utilizes a novel Semantic Initialization via spatial topology to preserve geometric structure and an Activation Potential Calibration (APC) mechanism to align pre-trained manifolds with sparsity constraints, preventing the dead neuron and dense collapse observed in standard fine-tuning. Empirically, VT is universally effective: it enables ModernBERT to achieve state-of-the-art performance on the BEIR benchmark (52.4 nDCG, a +4.7 improvement), resuscitates failing models like RoBERTa-large, and generalizes seamlessly to inference-free architectures and specialized domains. These results confirm that the performance lag is not an architectural deficiency but a solvable vocabulary mismatch. We've released our code and models.. https://anonymous.4open.science/r/vocab-transfer/. All details included.

Adaptive Sparsity Optimization with Learnable Soft Top-K and Per-Term Thresholding for Efficient Retrieval [Full]

Wentai Xie (University of California, Santa Barbara); Parker Carlson (University of California, Santa Barbara); Shanxiu He (University of California, Santa Barbara); Tao Yang (University of California, Santa Barbara)

Recent work on neural sparse retrieval has demonstrated strong relevance by leveraging Large Language Models (LLMs) for semantic term expansion. However, learned models paired with previous sparsification techniques still yield overly long document and query vectors partly due to a large LLM vocabulary, imposing a serious challenge to retrieval time and space efficiency. This paper proposes a scheme for optimizing model sparsity through a synergy of adaptive strategies, including learnable soft top-??, per-term thresholding, and FLOPs regularization to increase the sparsity of query and document vectors. Experimental results with Lion-SP model on the MS MARCO and BEIR datasets demonstrate that the proposed scheme can outperform the baselines by significantly reducing the average query and document lengths. Our scheme can achieve much shorter retrieval latency and lower storage cost while maintaining highly competitive relevance.

Efficient Sparse Retrieval with Lightweight Superblock Pruning [Full]

Parker Carlson (University of California, Santa Barbara); Wentai Xie (University of California, Santa Barbara); Rohil Shah (University of California, Santa Barbara); Tao Yang (University of California, Santa Barbara)

Learned sparse retrieval (LSR) is a popular method for first-stage retrieval because it combines the semantic matching of language models with efficient CPU-friendly algorithms. Previous work aggregates blocks into "superblocks" to quickly skip the visitation of blocks during query processing by using an advanced pruning heuristic. This paper proposes a simple and effective superblock pruning scheme that reduces the overhead of superblock score computation while preserving competitive relevance. It combines this scheme with a compact index structure and a robust zero-shot configuration that is effective across LSR models and multiple datasets. This paper provides an analytical justification and evaluation on the MS MARCO and BEIR datasets, demonstrating that the proposed scheme can be a strong alternative for efficient sparse retrieval.

Insights into the Efficiency of Open-Source Score-at-a-Time Search Engines: A Reproducibility Study [Reproducibility]

Katelyn Harlan (University of Otago); Andrew Trotman (University of Otago); Veronica Liesaputra (University of Otago)

Score-at-a-Time (SaaT) retrieval, utilising impact-ordered indexes, remains a relatively overlooked search strategy with increasing importance. Within the context of learned sparse representations and approximate retrieval, SaaT has proven to be a competitive alternative to the popular Document-at-a-Time (DaaT) approach. Yet, there are only two notable implementations of SaaT search engines: JASSv2 and IOQP. We are interested in the differences between these systems and how those differences affect performance. We identify differences in postings lists due to indexing, ranking functions, and quantization. Thus, we introduce ciffTools for quantizing the ciff indexes used by both search engines; eliminating these differences. Then, in a reproducibility study we reproduce previous experiments investigating the efficiency gap between the systems. They also differ in their compression codecs, with IOQP using SIMD BP-128 and JASSv2 using Elias Gamma SIMD VB. We, unexpectedly, find in another reproducibility experiment that SIMD BP-128 and QMX outperform Elias Gamma SIMD VB in situ. Finally, we investigate the CPU effect, and find the engines are affected differently. Overall, JASSv2 has faster median latency on a server-grade CPU, while on a desktop-grade they are evenly matched. IOQP has faster tail latency regardless of CPU. Our work reduces the throughput difference between JASSv2 and IOQP and offers insights into which aspects affect the efficiency of SaaT.

Understanding Wacky Weights: A Dissection of SPLADE's Learned Term Importance [Reproducibility]

Gregory Polyakov (University of Tübingen); Harrisen Scells (University of Tübingen); Carsten Eickhoff (University of Tübingen)

Learned sparse retrieval models such as SPLADE combine the effectiveness of neural architectures with the efficiency of inverted indices. As these models assign weights to terms from a fixed vocabulary, interpretability is often touted as a major benefit of these models. However, the emergence of wacky weights, i.e., expansion terms that appear semantically unrelated to the input, limits interpretability. While prior research has anecdotally observed this phenomenon, there is a lack of systematic understanding regarding their origins, prevalence, and contribution to retrieval effectiveness. In this paper, we reproduce SPLADE-v2 to systematically investigate wacky weights across the SPLADE family of models. We present a comprehensive dissection of wacky weights, providing a formal definition of wackiness based on the lexical utility of expansion terms. Furthermore, we introduce a novel measure to compare the prevalence of these tokens across models with varying vocabularies and sparsity levels. Beyond reproducing the original SPLADE-v2, we train it with various loss functions, datasets, and backbone transformers to isolate the factors contributing to wackiness. Our results show that larger vocabularies are associated with a higher prevalence of wacky tokens, while stricter sparsity regularizers are associated with lower prevalence. Finally, we find that wacky weights are used primarily for in-domain effectiveness rather than out-of-domain generalization.

Industry Session II · *Goldfields · Thursday 13:30–15:00*

From Doxa to Logos in Scientific Peer Review [Industry]

Negar Arabzadeh (Berkeley); Sajad Ebrahimi (University of Toronto); Alireza Daghighfarsoodeh (Reviewerly); Soroush Sadeghian (Reviewerly); Seyed Mohammad Hosseini (Reviewerly); Hai Son Le (Reviewerly); Mahdi Bashari (Reviewerly); Ebrahim Bagheri (University of Toronto)

Peer review is central to scientific decision-making, yet it is rarely evaluated or audited at scale. Growing submission volumes and the increasing use of large language models (LLMs) in drafting reviews have introduced new challenges for transparency, accountability, and quality control. Today, peer reviews are often produced through hybrid human--AI workflows, where a reviewer may develop the core evaluative ideas while using an LLM to refine wording, restructure arguments, or improve fluency. This shift raises new questions beyond authorship detection alone: Are reviews constructive? Are reviewer claims grounded in the submitted paper? How can we quantify collaboration between human reasoning and AI-assisted writing, and distinguish whether the intellectual contribution or the surface text originates from humans or models?In this industry talk, we present Reviewerly's retrieval-centered infrastructure for auditing peer review at scale. We describe three deployed systems: \textit{Peeriscope}, which evaluates review quality across multiple interpretable dimensions; \textit{Peerispect}, which uses retrieval-augmented generation (RAG) to verify whether reviewer claims are supported by evidence in the manuscript; and \textit{PeerPrism}, which analyzes hybrid human--AI authorship by disentangling the origin of ideas from the origin of text in peer reviews, enabling measurement of how human reasoning and AI-generated writing interact within a review. We will share architectural design decisions, lessons learned from real-world deployment, and practical trade-offs among model complexity, interpretability, computational efficiency, and predictive accuracy. The talk will include short live demonstrations\footnote{\url{https://app.reviewer.ly/app/peerispect}} \footnote{\url{https://app.reviewer.ly/app/peeriscope}} of our systems to illustrate how the combination of retrieval systems and LLM-based pipelines can improve peer review quality and strengthen transparency and research integrity across high-volume decision-making environments.

CS3: Efficient Online Capability Synergy for Two-Tower Recommendation [Industry]

Lixiang Wang (Kuaishou Technology); Shaoyun Shi (Kuaishou Technology); Peng Wang (Kuaishou Technology); Wenjin Wu (Kuaishou Technology); Peng Jiang (Kuaishou Technology)

To balance effectiveness and efficiency in recommender systems, multi-stage pipelines commonly use lightweight two-tower models for large-scale candidate retrieval. However, the isolated two-tower architecture restricts representation capacity, embedding-space alignment, and cross-feature interactions. Existing solutions such as late interaction and knowledge distillation can mitigate these issues, but often increase latency or are difficult to deploy in online learning settings. We propose Capability Synergy (CS3), an efficient online framework that strengthens two-tower retrievers while preserving real-time constraints. CS3 introduces three mechanisms: (1) Cycle-Adaptive Structure for self-revision via adaptive feature denoising within each tower; (2) Cross-Tower Synchronization to improve alignment through lightweight mutual awareness between towers; and (3) Cascade-Model Sharing to enhance cross-stage consistency by reusing knowledge from downstream models. CS3 is plug-and-play with diverse two-tower backbones and compatible with online learning. Experiments on three public datasets show consistent gains over strong baselines, and deployment in a large-scale advertising system yields up to 8.36% revenue improvement across three scenarios while maintaining ms-level latency.

Negative Data Mining for Contrastive Learning in Dense Retrieval at IKEA.com [Industry]

Eva Agapaki (IKEA Retail (Ingka Group)); Amritpal Singh Gill (IKEA Retail (Ingka Group))
Contrastive learning is a core component of modern retrieval systems, but its effectiveness heavily relies on the quality of negative examples used during training. In this work, we present a systematic approach to improving dense retrieval for IKEA product search through structured negative sampling strategies and scalable LLM-as-a-judge relevance evaluation. Building on IKEA Search Engine's late-interaction retrieval architectures, we introduce two key contributions: (1) structured negative sampling strategies that leverage product hierarchical taxonomy and product attributes to generate semantically challenging negatives, and (2) a comprehensive LLM-based evaluation methodology for generating training data. Rather than relying on sparse human annotations or random sampling, our LLM-based evaluation system allocates a score for all candidate products against each query.Our methodology achieves +2.6\% average category accuracy on offline real user query experiments on the Canada market. However, our A/B test on long-tail queries showed no statistically significant differences in user engagement metrics between the improved and baseline models ($p > 0.05$). We trace this gap to user search behavior: 67\% of popular searches exhibit zero-click rates above 50\%, indicating that a substantial proportion of search sessions result in no product engagement regardless of result ranking. These findings underscore the importance of hard negative mining but also the need for grounding training data and offline evals in real user search behavior---including query intent distribution and zero-click patterns---to bridge the gap between offline retrieval quality and online user engagement.

WeSEAL: Well-calibrated Search for Eliminating Attention-sink Leakage [Industry]

Juyuan Wang (Tencent); Chenxing Wang (Tencent); Aolin Li (Tencent); Huiyun Hu (Tencent); Yuchen Fang (Tencent); Haijun Wu (Tencent); Jin Xu (South China University of Technology); Dongliang Liao (South China University of Technology)
Large Language Models (LLMs) have revolutionized relevance ranking, yet their deployment in high-concurrency search systems is stifled by a critical pathology: "Attention Sink Drift." In standard causal attention, models disproportionately allocate weight to the initial document regardless of relevance, creating a severe positional bias that traditionally necessitates costly test-time calibration, effectively halving system throughput. We challenge the consensus that this bias is immutable and present WeSEAL, a single-pass framework that internalizes calibration directly into model parameters. By employing a contrastive head selection mechanism and a novel attention-aligned Supervised Fine-Tuning objective, WeSEAL suppresses background noise and redirects discriminative signals to semantic retrieval heads, thereby intrinsically correcting positional bias. We deployed WeSEAL in the core search scene of WeChat, a large-scale mobile platform serving over 1 billion monthly active users. In rigorous online A/B testing, our approach not only reduced P99 latency to 26.56ms (a 3.3x speedup) and improved the Positive-to-Negative Ratio (PNR) to 2.08, but also delivered statistically significant gains in core engagement metrics—including a 0.402% increase in Click-Through Rate (CTR), a 0.185% growth in Average Video View Duration, and a 1.25% rise in overall user satisfaction—without increasing serving costs. This work establishes a new Pareto frontier for efficient LLM ranking, demonstrating that topologically optimized small-scale models can deliver robust, unbiased performance in strict industrial environments.

Agentic Multi-Source Grounding for Enhanced Query Intent Understanding: A DoorDash Case Study [Industry]

Emmanuel Aboah Boateng (DoorDash); Kyle MacDonald (DoorDash); Akshad Viswanathan (DoorDash); Sudeep Das (DoorDash)
Accurately mapping user queries to business categories is a fundamental Information Retrieval challenge for multi-category marketplaces, where context-sparse queries such as "Wildflower" exhibit intent ambiguity, simultaneously denoting a restaurant chain, a retail product, and a floral item. Traditional classifiers force a winner-takes-all assignment, while general-purpose LLMs hallucinate unavailable inventory. We introduce an Agentic Multi-Source Grounded system that addresses both failure modes by grounding LLM inference in (i) a staged catalog entity retrieval pipeline and (ii) an agentic web-search tool invoked autonomously for cold-start queries. Rather than predicting a single label, the model emits an ordered multi-intent set, resolved by a configurable disambiguation layer that applies deterministic business policies and is designed for extensibility to personalization signals. This decoupled design generalizes across domains, allowing any marketplace to supply its own grounding sources and resolution rules without modifying the core architecture. Evaluated on DoorDash's multi-vertical search platform, the system achieves +10.9pp over the ungrounded LLM baseline and +4.6pp over the legacy production system. On long-tail queries, incremental ablations attribute +8.3pp to catalog grounding, +3.2pp to agentic web search grounding, and +1.5pp to dual-intent disambiguation, yielding 90.7% accuracy (+13.0pp over baseline). The system is deployed in production, serving over 95% of daily search impressions, and establishes a generalizable paradigm for applications requiring foundation models grounded in proprietary context and real-time web knowledge to resolve ambiguous, context-sparse decision problems at scale.

Policy-Grounded Dynamic Facet Suggestions for Job Search [Industry]

Dan Xu (LinkedIn Corporation); Baofen Zheng (LinkedIn Corporation); Qianqi Shen (LinkedIn Corporation); Jianqiang Shen (LinkedIn Corporation); Wenqiong Liu (LinkedIn Corporation); Chunnan Yao (LinkedIn Corporation); Ping Liu (LinkedIn Corporation); Rajat Arora (LinkedIn Corporation); Kevin Kao (LinkedIn Corporation); Hsiang Lin (LinkedIn Corporation); Wanjun Jiang (LinkedIn Corporation); Yusuke Takebuchi (LinkedIn Corporation); Jingwei Wu (LinkedIn Corporation); Wenjing Zhang (LinkedIn Corporation)
Job seekers often initiate search with short, underspecified queries. At LinkedIn, over 80% of job-related queries contain three or fewer keywords, making accurate user intent inference and relevant job retrieval particularly challenging. We present dynamic facet suggestion (DFS), an interactive query-refinement mechanism that facilitates intent disambiguation by surfacing personalized semantic attributes conditioned on the joint user-query context in real time. We propose a policy-grounded, retrieval-augmented ranking framework for facet suggestion, comprising offline taxonomy curation, embedding-based retrieval of top-K candidates, and a distilled small language model (SLM) based candidate scoring. The system is optimized for real-time serving via point-wise single-token scoring and batching/prefix caching. Offline evaluation demonstrates high precision for generated suggestions, and online A/B tests show significant lifts in suggestion engagement and job search outcomes.

Generative Recommendation · *Room 110 · Thursday 16:15–17:30*

Large-Scale Online Learning for Generative List Recommendation in E-commerce: An Environment Policy Optimization Approach [Full]

Yuan Wang (Alibaba Group); Zhiyu Li (Alibaba Group); Ang Gao (Renmin University of China); Changshuo Zhang (Renmin University of China); Xiao Zhang (Renmin University of China); Jun Xu (Renmin University of China); Quan Lin (Alibaba Group)
Generative List Recommendation (GLR) models have shown superior performance in E-commerce by directly generating high-quality recommendation lists through sequential item selection. While Online Learning (OL) has proven valuable for streaming point-wise recommendation models in adapting to dynamic user preferences, its application to GLR remains largely unexplored, due in large part to the inefficiency and instability of conventional on-policy reinforcement learning algorithms used in existing GLR approaches. Existing approaches typically rely on surrogate losses, which provide indirect and biased gradient estimates, making them ill-suited for the rapid, subtle distribution shifts common in real-world E-commerce environments. In this paper, we propose Environment Policy Optimization (EPO), a novel GLR model that fundamentally reshapes policy learning by exploiting the differentiability of the environment within the Generator-Evaluator framework. EPO recognizes that the evaluator is a neural network, capable of providing gradient signals. By directly utilizing these gradients, EPO enables end-to-end optimization of the total list-wise reward—the true objective. To ensure differentiable list generation, EPO introduces a new indexing and exploration strategy based on NeuralSort and Gumbel noise, which relaxes discrete item selection into a continuous, gradient-friendly operation. EPO not only demonstrates strong performance in offline evaluations but also unlocks the potential of online learning for GLR at an industrial scale, yielding a 1.18% relative improvement in user clicks in online A/B tests. The results underscore the critical role of EPO in providing the sensitivity, stability, and gradient fidelity necessary for effective real-time adaptation in streaming and dynamic recommendation environments. Moreover, EPO reached the baseline performance with a training time deduction of 76% under identical hardware conditions.

GenRecEdit: Adapting Model Editing for Generative Recommendation with Cold-Start Items [Full]

Chenglei Shen (Renmin University of China); Teng Shi (Renmin University of China); Weijie Yu (University of International Business and Economics); Xiao Zhang (Renmin University of China); Jun Xu (Renmin University of China)
Generative recommendation (GR) has demonstrated substantial potential for sequential recommendation in an end-to-end generation paradigm. However, existing models suffer from severe cold-start collapse, i.e., recommendation accuracy on cold-start items drops to near zero. Current solutions rely on retraining with cold-start interactions, which faces sparse feedback, high computational cost, and delayed updates, diminishing practical utility in rapidly evolving catalogs in recommendation. Inspired by NLP model editing (enabling training-free knowledge injection into large language models), we explore applying this paradigm to generative recommendation, but face two fundamental challenges: (1)~sequential data in GR lacks explicit subject–object binding (a core NLP sentence structure), hindering targeted model edits; (2)~sequential data in GR has no fixed token co-occurrence patterns (unlike NLP's stable phrases), making multi-token injection unreliable. To address these, we propose GenRecEdit, the first model editing framework tailored for generative recommendation. Specifically, we: (1)~mitigate the absence of sentence structure by explicitly modeling the intrinsic relationship between the entire sequence context and the next-token (e.g., semantic IDs, SIDs) generation; (2)~adopt iterative token-level edits to effectively inject token bundles (e.g., items); and (3)~introduce a One-One trigger mechanism to avoid interactions among multiple token-level edits during inference. Extensive experiments across multiple datasets demonstrate that ørname substantially improves recommendation performance on cold-start items while preserving the model's original recommendation quality. Moreover, ørname achieves these gains with only approximately 9.5% of the training time required by retraining, significantly reducing computational cost and enabling efficient, frequent model updates.

Adaptive Mix Preference Optimization for Generative Recommendation [Full]

Junbo Qi (Waseda University); Yanyan Zou (JD.com); Xuanhua Yang (JD.com); Sulong Xu (JD.com); Ying Sun (Harbin Institute of Technology); Shengjie Li (JD.com)

Recommender systems aim to leverage user interaction signals to recommend items that users are likely to be interested in. Motivated by the success of Large Language Model (LLMs), generative recommendation (GR) has recently gained increasing attention, typically following a two-stage paradigm: supervised fine-tuning followed by preference alignment. However, aligning generative recommenders with users' personalized preferences remains challenging, as user feedback is inherently heterogeneous and uncertain. Different types of user interaction signals reflect varying levels of intent and should therefore be modeled differently. In this work, we propose Adaptive Mix Preference Optimization (AMPO), an adaptive alignment framework that mixes likelihood and preference objectives with self-calibrated, sample-wise confidence adjustment. AMPO introduces an adaptive target margin that leverages the model's own probability ratio to modulate optimization strength: confident pairs receive full margins that reinforce correct rankings, while uncertain pairs receive reduced margins that prevent overfitting to ambiguous signals. Additionally, AMPO incorporates negative log-likelihood regularization on preferred items to counteract likelihood displacement, a phenomenon where contrastive objectives cause preferred and non-preferred probabilities to collapse simultaneously. Such a design eliminates the need for a reference model, yielding up to 2.5x speedup and 35% memory reduction. Extensive experiments on public benchmarks and a large-scale industrial dataset demonstrate consistent improvements in ranking metrics. Online A/B tests on a major e-commerce platform further confirm statistically significant gains in click-through and conversion rates. The code is available at https://github.com/jumbo-q/ampo.

MVIGER: Multi-View Variational Integration of Complementary Knowledge for Generative Recommender [Full]

Tongyoung Kim (Yonsei University); SooJin Yoon (Yonsei University); SeongKu Kang (Korea University); Jinyoung Yeo (Yonsei University); Dongha Lee (Yonsei University)

Language Models (LMs) have been widely used in recommender systems to incorporate textual information of items into item IDs, leveraging their advanced language understanding and generation capabilities. Recently, generative recommender systems have utilized the reasoning abilities of LMs to directly generate index tokens for potential items of interest based on the user's interaction history. To inject diverse item knowledge into LMs, prompt templates with detailed task descriptions and various indexing techniques derived from diverse item information have been explored. This paper focuses on the inconsistency in outputs generated by variations in input prompt templates and item index types, even with the same user's interaction history. Our in-depth quantitative analysis reveals that preference knowledge learned from diverse prompt templates and heterogeneous indices differs significantly, indicating a high potential for complementarity. To fully exploit this complementarity and provide consistent performance under varying prompts and item indices, we propose MVIGER, a unified variational framework that models selection among these information sources as a categorical latent variable with a learnable prior. During inference, this prior enables the model to adaptively select the most relevant source or aggregate predictions across multiple sources, thereby ensuring high-quality recommendation across diverse template-index combinations. We validate the effectiveness of MVIGER on three real-world datasets, demonstrating its superior performance over existing generative recommender baselines through the effective integration of complementary knowledge.

BEAR: Towards Beam-Search-Aware Optimization for Recommendation with Large Language Models [Full]

Weiqin Yang (State Key Laboratory of Blockchain and Data Security, Zhejiang University); Bohao Wang (State Key Laboratory of Blockchain and Data Security, Zhejiang University); Zhenxiang Xu (State Key Laboratory of Blockchain and Data Security, Zhejiang University); Jiawei Chen (State Key Laboratory of Blockchain and Data Security, Zhejiang University); Shengjia Zhang (State Key Laboratory of Blockchain and Data Security, Zhejiang University); Jingbang Chen (The Chinese University of Hong Kong, Shenzhen); Canghong Jin (Hangzhou City University); Can Wang (State Key Laboratory of Blockchain and Data Security, Zhejiang University)

Recent years have seen a rapid surge in research leveraging Large Language Models (LLMs) for recommendation. These methods typically employ supervised fine-tuning (SFT) to adapt LLMs to recommendation scenarios, and utilize beam search during inference to efficiently retrieve B top-ranked recommended items. However, we identify a critical training-inference inconsistency: while SFT optimizes the overall probability of positive items, it does not guarantee that such items will be retrieved by beam search even if they possess high overall probabilities. Due to the greedy pruning mechanism, beam search can prematurely discard a positive item once its prefix probability is insufficient. To address this inconsistency, we propose BEAR (Beam-SEarch-Aware Regularization), a novel fine-tuning objective that explicitly accounts for beam search behavior during training. Rather than directly simulating beam search for each instance during training, which is computationally prohibitive, BEAR enforces a relaxed necessary condition: each token in a positive item must rank within the top-B candidate tokens at each decoding step. This objective effectively mitigates the risk of incorrect pruning while incurring negligible computational overhead compared to standard SFT. Extensive experiments across four real-world datasets demonstrate that BEAR significantly outperforms strong baselines. Code is available at https://github.com/Tiny-Snow/BEAR-SIGIR-2026.

Community and Subgraph Search · Eureka 1 · Thursday 16:15–17:30

One-for-All Community Search on Unseen Graphs [Full]

Mo Li (Liaoning University); Zhaosong Zhao (Liaoning University); Linlin Ding (Liaoning University); Renata Borovica-Gajic (University of Melbourne); Zhongming Yao (Aalborg University); Jianxin Li (Edith Cowan University)

Community search is a fundamental graph-based retrieval problem that aims to identify a query-dependent subgraph whose nodes exhibit strong internal connectivity. While recent learning-based methods improve retrieval effectiveness via graph representation learning, they follow a "one-use-one-train" paradigm that requires retraining or fine-tuning for each target graph, leading to high data dependency, high training costs, and limited generalization. To handle this, we propose OFA-CS, a "one-for-all" community search framework trained once on source datasets and directly deployed to arbitrary unseen graphs without retraining or fine-tuning, while preserving strong performance. Specifically, we introduce a Spectral-Aware Feature Alignment module to unify feature dimensionality and align cross-domain semantics in a community-aware manner. We further develop a Graph Diffusion Tokenized Transformer that constructs hybrid token sequences from local and global structural contexts for Transformer encoding, and applies diffusion-based refinement to mitigate distribution shifts on unseen graphs. With the unified representations, communities are efficiently retrieved via a modularity-driven search procedure. Extensive experiments on diverse real-world graphs demonstrate that OFA-CS achieves strong cross-domain generalization and competitive retrieval effectiveness against state-of-the-art methods, without requiring target-domain supervision.

GNN-based Anchor Embedding for Efficient Subgraph Retrieval [Full]

Bin Yang (Harbin Institute of Technology); Jianxiong Ye (Harbin Institute of Technology); Zhaonian Zou (Harbin Institute of Technology)

Several recent works utilize deep learning (DL) techniques for subgraph retrieval via matching, yet most only return approximate isomorphism relations between queries and data graphs--failing to retrieve all exact matching locations, a critical demand for structured graph retrieval in information retrieval. Unlike these DL-based approximate methods, we propose a learning-based framework for subgraph retrieval, called the graph neural network (GNN)-based anchor embedding framework (GNN-AE), which can efficiently retrieve all exact matching locations. In contrast to most traditional exact subgraph matching methods, which create auxiliary structures online for each query, our method has two core optimizations: (1) We construct offline, one-time-only efficient embedding indices for small feature subgraphs (namely, anchored subgraphs and anchored paths) in the data graph and obtain candidates for the query on these indexed feature subgraphs, trading space for time to reduce online query latency; (2) We leverage GNNs to perform graph isomorphism tests on indexed feature subgraphs and generate low-conflict embeddings for these feature subgraphs, yielding a high-quality, compact set of candidates that further enhances query efficiency. Beyond these core optimizations, we develop a parallel matching growth algorithm and design a cost-based DFS query strategy to retrieve all matching locations. Extensive experiments on both real and synthetic datasets validate the efficiency and effectiveness of our GNN-AE for exact subgraph retrieval.

Cohesive Group Discovery in Interaction Graphs under Explicit Density Constraints [Full]

Yu Zhang (Peking University); Yilong Luo (Peking University); Mingyuan Ma (Peking University); Yao Chen (Beijing Capital International Airport Co., Ltd.); Enqiang Zhu (Guangzhou University); Jin Xu (Peking University); Chanjuan Liu (Dalian University of Technology)

Discovering cohesive groups is a fundamental primitive in graph-based recommender systems, underpinning tasks such as social recommendation, bundle discovery, and community-aware modeling. In interaction graphs, cohesion is often modeled as the γ-quasi-clique, an induced subgraph whose internal edge density meets a user-defined threshold γ. This formulation provides explicit control over within-group connectivity while accommodating the sparsity inherent in real-world data. However, ensuring explicit density constraints while maintaining robustness remains challenging for existing heuristic approaches. This paper presents EDQC, an effective framework for cohesive group discovery under explicit density constraints. EDQC leverages a lightweight energy diffusion process to rank vertices for localizing promising candidate regions. Guided by this ranking, the framework extracts and refines a candidate subgraph to ensure the output strictly satisfies the target density requirement. Extensive experiments on 75 real-world graphs across varying density thresholds demonstrate that EDQC identifies the largest mean γ-quasi-cliques in the vast majority of cases, achieving lower variance than the state-of-the-art methods while maintaining competitive runtime, making it a robust and practical solution for cohesive group discovery in graph-based recommender systems.

Inductive Dual-Polarity Modeling via Static–Dynamic Disentanglement for Dynamic Signed Networks [Full]

Yikang Hou (Southwest University); Junjie Huang (Southwest University); Yijun Ran (Guizhou Normal University); Tao Jia (Southwest University)

Dynamic signed networks (DSNs) are common in online platforms, where time-stamped positive and negative relations evolve over time. A core task in

DSNs is dynamic edge prediction, which forecasts future relations by jointly modeling edge existence and polarity (positive, negative, or non-existent). However, existing dynamic signed network embedding (DSNE) methods often entangle positive and negative signals within a shared temporal state and rely on node-specific temporal trajectories, which can obscure polarity-asymmetric dynamics and harm inductive generalization, especially under cold-start evaluation. We study an inductive setting where each test edge contains at least one endpoint node held out from training, while its interactions prior to the prediction time are available as historical evidence. The model must therefore infer representations for unseen nodes solely from such limited history. We propose IDP-DSN, an Inductive Dual-Polarity framework for Dynamic Signed Networks. IDP-DSN maintains sign-selective memories to model positive and negative temporal dynamics separately, performs history-only neighborhood inference for unseen nodes (instead of learned node-wise trajectories), and enforces polarity-wise static--dynamic disentanglement via an orthogonality regularizer. Experiments on BitcoinAlpha, BitcoinOTC, Wiki-RfA, and Epinions demonstrate consistent improvements over the strongest baselines, achieving relative Macro-F1 gains of 16.8/23.4%, 16.9/24%, 30.1/25.5%, and 18.7/28.9% in the transductive/inductive settings, respectively. These results highlight the effectiveness of IDP-DSN on DSNs, particularly under inductive cold-start evaluation for dynamic signed edge prediction.

DCGL: Dual-Channel Graph Learning with Large Language Models for Knowledge-Aware Recommendation [Full]

Xinchi Zou (Huazhong University of Science and Technology); Tongzhenzhi Su (Huazhong University of Science and Technology); Jianjun Li (Huazhong University of Science and Technology); Yuan Fu (Huazhong University of Science and Technology); Chang Liu (Huazhong University of Science and Technology); Zhiying Deng (Central China Normal University); Zhiwei Shen (Huazhong University of Science and Technology)

Knowledge Graphs (KGs) have proven highly effective for recommendation systems by capturing latent item relationships, while recent integration of Large Language Models (LLMs) has further enhanced semantic understanding and addressed knowledge sparsity issues. Nevertheless, current KG&LLM-based methods still face three main limitations: 1) inadequate modeling of implicit semantics relationships beyond explicit KG links; 2) suboptimal single-channel fusion of ID and LLM embeddings, which often leads to signal interference and blurred representations; and 3) insufficient consideration of user-item interaction frequency variations in recommendation strategies. To address these challenges, we propose the Dual-Channel Graph Learning (DCGL) framework, featuring three key innovations: 1) a dual-channel architecture that structurally decoupling rich semantic information from user behavioral patterns, preventing early interference; 2) a multi-level contrastive learning mechanism that enhances robustness against KG noise through intra-view contrast and bridges semantic gaps between channels via inter-view alignment; and 3) a dynamic fusion mechanism that adaptively balances semantic generalization and behavioral specificity based on interaction frequency, resolving the cascading limitation. Extensive experiments on four real-world datasets show that DCGL consistently outperforms state-of-the-art methods, yielding substantial improvements in sparse scenarios while maintaining precision for active users. Our code is available at https://github.com/XinchiZou/DCGL.

Multimodal RAG · *Eureka 2 · Thursday 16:15–17:30*

Mixture-of-Retrieval Experts for Reasoning-Guided Multimodal Knowledge Exploitation [Full]

Chunyi Peng (Northeastern University); Zhipeng Xu (Northeastern University); Zhenghao Liu (Northeastern University); Yishan Li (ModelBest, Inc.); Yukun Yan (Tsinghua University); Shuo Wang (Tsinghua University); Yu Gu (Northeastern University); Minghe Yu (Northeastern University); Ge Yu (Northeastern University); Maosong Sun (Tsinghua University)

Multimodal Retrieval-Augmented Generation (MRAG) has shown promise in mitigating hallucinations in Multimodal Large Language Models (MLLMs) by incorporating external knowledge. However, existing methods typically adhere to rigid retrieval paradigms by mimicking fixed retrieval trajectories and thus fail to fully exploit the knowledge of different retrieval experts through dynamic interaction based on the model's knowledge needs or evolving reasoning states. To overcome this limitation, we introduce Mixture-of-Retrieval Experts (MoRE), a novel framework that enables MLLMs to collaboratively interact with diverse retrieval experts for more effective knowledge exploitation. Specifically, MoRE learns to dynamically determine which expert to engage with, conditioned on the evolving reasoning state. To effectively train this capability, we propose Stepwise Group Relative Policy Optimization (Step-GRPO), which goes beyond sparse outcome-based supervision by encouraging MLLMs to interact with multiple retrieval experts and synthesize fine-grained rewards, thereby teaching the MLLM to fully coordinate all experts when answering a given query. Experimental results on diverse open-domain QA benchmarks demonstrate the effectiveness of MoRE, achieving average performance gains of over 7% compared to competitive baselines. Notably, MoRE exhibits strong adaptability by dynamically coordinating heterogeneous experts to precisely locate relevant information, validating its capability for robust, reasoning-driven expert collaboration. All codes and data are released on https://github.com/OpenBMB/MoRE.

Towards Mixed-Modal Retrieval for Universal Retrieval-Augmented Generation [Full]

Chenghao Zhang (Renmin University of China); Guanting Dong (Renmin University of China); Xinyu Yang (Renmin University of China); Zhicheng Dou (Renmin University of China)

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing Large Language Models (LLMs) by retrieving relevant documents from an external corpus. However, existing RAG systems primarily focus on unimodal text documents, and often fall short in real-world scenarios where both queries and documents may contain mixed modalities (such as text and images). In this paper, we address the challenge of Universal Retrieval-Augmented Generation (URAG), which involves retrieving and reasoning over mixed-modal information to improve vision-language generation. To this end, we propose Nyx, a unified mixed-modal-to-mixed-modal retriever tailored for URAG scenarios. To mitigate the scarcity of realistic mixed-modal data, we introduce a four-stage automated pipeline for data generation and filtering, leveraging web documents to construct NyxQA, a dataset comprising diverse mixed-modal question-answer pairs that better reflect real-world information needs. Building on this high-quality dataset, we adopt a two-stage training framework for Nyx: we first perform pre-training on NyxQA along with a variety of open-source retrieval datasets, followed by supervised fine-tuning using feedback from downstream Vision-Language Models (VLMs) to align retrieval outputs with generative preferences. Experimental results demonstrate that Nyx not only performs competitively on standard text-only RAG benchmarks, but also excels in the more general and realistic URAG setting, significantly improving generation quality in vision-language tasks.

mKG-RAG: Leveraging Multimodal Knowledge Graphs in Retrieval-Augmented Generation for Knowledge-intensive VQA [Full]

Xu Yuan (The Hong Kong Polytechnic University); Liangbo Ning (The Hong Kong Polytechnic University); Qingqing Ye (The Hong Kong Polytechnic University); Wenqi Fan (The Hong Kong Polytechnic University); Qing Li (The Hong Kong Polytechnic University)

Retrieval-Augmented Generation (RAG) has emerged as an effective paradigm for expanding the knowledge capacity of Multimodal Large Language Models (MLLMs) by incorporating external knowledge sources into the generation process, and has been widely adopted for knowledge-based Visual Question Answering (VQA). Despite impressive advancements, vanilla RAG-based VQA methods that rely on unstructured documents and overlook the structural relations among knowledge elements frequently introduce irrelevant or misleading content, degrading answer accuracy and reliability. To overcome these challenges, a promising solution is to integrate multimodal knowledge graphs (KGs) into RAG-based VQA frameworks, thereby enhancing generation through structured multimodal knowledge. To this end, this paper proposes mKG-RAG, a novel retrieval-augmented generation framework built upon multimodal KGs for knowledge-intensive VQA tasks. Specifically, mKG-RAG leverages MLLM-driven graph extraction and vision-text matching to distill semantically consistent, modality-complementary entities and relations from multimodal documents, constructing high-quality multimodal KGs as structured knowledge representations. Furthermore, a dual-stage retrieval strategy equipped with a query-aware multimodal retriever is introduced to improve retrieval efficiency while progressively refining precision. Comprehensive experiments demonstrate that our approach significantly outperforms existing approaches and sets new state-of-the-art results for knowledge-based VQA. The code is available at https://github.com/xandery-geek/mKG-RAG.

Chain of Evidence: Pixel-Level Visual Attribution for Iterative Retrieval-Augmented Generation [Full]

Peiyang Liu (Peking University); Ziqiang Cui (City University of Hong Kong); Xi Wang (Peking University); Di Liang (Tencent Technology); Wei Ye (Peking University)

Iterative Retrieval-Augmented Generation (iRAG) has emerged as a powerful paradigm for answering complex multi-hop questions by progressively retrieving and reasoning over external documents. However, current systems predominantly operate on parsed text, which creates two critical bottlenecks: (1) Coarse-grained attribution, where users are burdened with manually locating evidence within lengthy documents based on vague text-level citations; and (2) Visual semantic loss, where the conversion of visually rich documents (e.g., slides, PDFs with charts) into text discards spatial logic and layout cues essential for reasoning. To bridge this gap, we present Chain of Evidence (CoE), a retriever-agnostic visual attribution framework that leverages Vision-Language Models to reason directly over screenshots of retrieved document candidates. CoE eliminates format-specific parsing and outputs precise bounding boxes, visualizing the complete reasoning chain within the retrieved candidate set. We evaluate CoE on two distinct benchmarks: Wiki-CoE, a large-scale dataset of structured web pages derived from 2WikiMultiHopQA, and SlideVQA, a challenging dataset of presentation slides featuring complex diagrams and free-form layouts. Experiments demonstrate that fine-tuned Qwen3-VL-8B-Instruct achieves robust performance, significantly outperforming text-based baselines in scenarios requiring visual layout understanding, while establishing a retriever-agnostic solution for pixel-level interpretable iRAG. Our code is available at https://github.com/PeiYangLiu/CoE.git.

Purifying Multimodal Retrieval: Fragment-Level Evidence Selection for RAG [Full]

Xihang Wang (Zhejiang University); Zihan Wang (Meituan LongCat Interaction Team); Chengkai Huang (University of New South Wales); Cao Liu (Meituan LongCat Interaction Team); Ke Zeng (Meituan LongCat Interaction Team); Quan Z. Sheng (Macquarie University); Lina Yao (University of New South Wales)

Multimodal Retrieval-Augmented Generation (MRAG) is widely adopted for Multimodal Large Language Models (MLLMs) with external evidence to reduce hallucinations. Despite its success, most existing MRAG frameworks treat retrieved evidence as indivisible documents, implicitly assuming that all content within a document is equally informative. In practice, however, sometimes only a small fraction of a document is relevant to a given query, while the remaining content introduces substantial noise that may lead to performance degradation. We address this fundamental limitation by reframing MRAG as a fine-grained evidence selection problem. We propose Fragment-level Evidence Selection for RAG (FES-RAG), a framework that selects atomic multimodal fragments rather than entire documents as grounding evidence. FES-RAG decomposes retrieved multimodal documents into sentence-level textual fragments and region-level visual fragments, enabling precise identification of evidence that directly supports generation. To guide fragment selection, we introduce Fragment Information Gain (FIG), a principled metric that measures the marginal contribution of each fragment to the MLLM's generation confidence. Based on FIG, we distill fragment-level utility judgments from a high-capacity MLLM into a lightweight selector, achieving accurate evidence selection with low inference overhead. Experiments on the M2RAG benchmark show that FES-RAG consistently outperforms state-of-the-art document-level MRAG methods, achieving up to 27% relative improvement in CIDEr. By selecting fewer yet more informative fragments, our approach substantially reduces context length while improving factual accuracy and generation coherence.

Query Performance and User Interaction Models ·

Eureka 3 · Thursday 16:15–17:30

Query Performance Prediction Using Relevance Judgments Generated by Large Language Models [TOIS]

Chuan Meng (University of Amsterdam, Amsterdam, Netherlands); Negar Arabzadeh (University of Waterloo, Waterloo, Ontario, Canada); Arian Askari (Leiden University, Leiden, Netherlands); Mohammad Aliannejadi (University of Amsterdam, Amsterdam, Netherlands); Maarten de Rijke (University of Amsterdam, Amsterdam, Netherlands)

Previous query performance prediction (QPP) methods typically return a single scalar value and do not require the predicted values to approximate a specific IR evaluation measure, leading to drawbacks: (i) a single scalar is insufficient to accurately represent different IR measures, especially when metrics do not correlate, and (ii) a single scalar limits the interpretability of QPP methods because solely using a scalar is insufficient to explain QPP results. To address these issues, we propose a QPP framework using automatically generated relevance judgments (QPP-GenRE), which decomposes QPP into independent subtasks of predicting the relevance of each item in a ranked list to a given query. This allows us to predict any IR measure using the generated relevance judgments as pseudo-labels. This also allows us to interpret predicted IR measures, and identify errors in generated relevance judgments to improve QPP quality. We predict an item's relevance by using open-source LLMs. We face two main challenges: (i) excessive computational costs of judging an entire corpus for predicting a metric considering recall, and (ii) limited performance in prompting open-source LLMs. To solve both, we devise an approximation strategy and propose to fine-tune open-source LLMs using human-labeled relevance judgments. Experiments on multiple ad-hoc and conversational search datasets show that QPP-GenRE achieves state-of-the-art QPP quality.

Can QPP Choose the Right Query variant? Evaluating Query Variant Selection for RAG Pipelines [Reproducibility]

Negar Arabzadeh (UC Berkeley); Andrew Drozdov (Databricks); Michael Bendersky (Databricks); Matei Zaharia (UC Berkeley)

Large Language Models (LLMs) have made query reformulation ubiquitous in modern retrieval and Retrieval-Augmented Generation (RAG) pipelines, enabling the generation of multiple semantically equivalent query variants. However, executing the full pipeline for every reformulation is computationally expensive, motivating selective execution: can we identify the best query variant before incurring downstream retrieval and generation costs?We investigate Query Performance Prediction (QPP) as a mechanism for variant selection across ad-hoc retrieval, and end-to-end RAG. Unlike traditional QPP, which estimates query difficulty across topics, we study intra-topic discrimination—selecting the optimal reformulation among competing variants of the same information need.Through large-scale experiments on TREC-RAG using both sparse and dense retrievers, we evaluate pre- and post-retrieval predictors under correlation- and decision-based metrics. Our results reveal a systematic divergence between retrieval and generation objectives: variants that maximize ranking metrics such as nDCG often fail to produce the best generated answers, exposing a "utility gap" between retrieval relevance and generation fidelity.Nevertheless, QPP can reliably identify variants that improve end-to-end quality over the original query. Notably, lightweight pre-retrieval predictors frequently match or outperform more expensive post-retrieval methods, offering a latency-efficient approach to robust RAG.

Towards a Relevance Posterior in Neural Information Access [Perspective]

Andrew Parry (University of Glasgow); Emmanouil Georgios Lionis (University of Glasgow); Debasis Ganguly (University of Glasgow); Sean MacAvaney (University of Glasgow)

Modern information retrieval systems typically operationalise relevance as a query-conditional score computed at inference time. This design choice has become dominant such that alternative decompositions of relevance are rarely discussed, despite the long history of document and query priors in probabilistic retrieval and large-scale search. As neural ranking models grow more computationally expensive and retrieval pipelines expand to include multi-stage ranking, recommendation, and retrieval-augmented generation, this monolithic view of query-time scoring becomes increasingly limiting. We argue that modern information access systems are more naturally understood as performing approximate posterior inference, in which relevance is refined through a staged combination of query-dependent likelihoods and query-independent priors. We extend classical probabilistic retrieval formalisms to contemporary learned systems and show how explicit likelihood–prior decomposition exposes new opportunities to shift computation offline while disentangling document-level and interaction-level beliefs. We present empirical evidence that incorporating query-independent document utility can complement existing rankers and improve effectiveness with minimal query-time computation (solely score fusion). Concretely, a learned prior improves first-stage retrieval through rank fusion (up to ? nDCG@10 ? 0.046 on TREC DL-2019 and ? 0.029 on TREC DL-2020) and also improves downstream re-ranking, with the largest gains observed for the LLM re-ranker RankZephyr (up to ? nDCG@10 ? 0.054 on TREC DL-2020). Finally, we discuss how this decomposition connects to broader information access and outline research directions for designing retrieval systems that explicitly allocate modelling capacity between offline priors and online interaction.

Task-Aware Automated User Profile Generation for Recommendation Simulation Using Large Language Models [Full]

Xinye Wanyan (RMIT University); Chenglong Ma (RMIT University); Danula Hettiachchi (RMIT University); Ziqi Xu (RMIT University); Jeffrey Chan (RMIT University)

Large Language Model (LLM)-based agent simulation has emerged as a promising approach to meet the increasing demand for real-time and rigorous evaluation in modern recommender systems. A typical LLM-driven simulation framework comprises three essential components: the profile module, memory module, and action module. However, existing studies have primarily concentrated on enhancing the memory and action modules, with limited attention to profile generation, which plays a pivotal role in ensuring realistic agent behaviours and aligning simulated interactions with real user dynamics. Moreover, the scarcity of datasets specifically designed for recommendation simulations has led to heavy reliance on manually crafted profiles, significantly limiting the scalability and generalisability of simulation frameworks across different datasets. To address these challenges, this work proposes an Automated Profile Generation Framework for Recommendation Simulation, APG4RecSim, that constructs realistic, coherent, and robust user profiles with minimal supervision. Extensive experiments on three benchmark datasets demonstrate that APG4RecSim achieves the best overall performance on discrimination, ranking, and rating tasks, improving ranking quality by up to 7% in nDCG@10 and reducing rating distribution divergence by 8% in JSD compared to existing profile-generation baselines. Beyond overall performance gains, our results show that APG4RecSim produces profiles that are resilient to popularity- and position-induced biases and maintain stable performance across datasets and different LLMs.

An Epistemic Position-Based Click Model: From Interactions to Epistemic Distributions of Relevance and Bias [Full]

Oscar Rolando Ramirez Milian (University of Amsterdam); Harrie Oosterhuis (University of Amsterdam)

User interactions with rankings are affected by both items' relevances and display positions. Accordingly, click probabilities are often modeled as a product of relevance and position factors; and for improving recommendation and search, one needs to disentangle relevance from position bias. However, existing click models only provide frequentist point-estimates that do not capture any measure of epistemic uncertainty. Consequently, there is no indication of how much confidence one should have in their predictions. In this work, we introduce the first evidential deep-learning approach to form an epistemic alternative to the important position-based click model. Our learned model takes as input item and position features and outputs a beta-distribution for every relevance and position-bias variable of the position-based model. These distributions capture epistemic uncertainty about click probabilities and the underlying effects of attraction and position bias. The main challenge of our approach is its optimization for which we propose approximation and conditioning techniques to provide numerical stability and variance reduction. Our experiments indicate that our approach captures epistemic uncertainty in predictions on previously-unseen data, whereas standard policy gradients fail to learn meaningful distributions. We believe our contribution of the first contextual epistemic click model constitutes an important step in incorporating Bayesian uncertainty into click modeling.

Low-Resource and Educational IR · *Courtyard 1+2* ·

Thursday 16:15–17:30

Improving Interpretability of Cognitive Diagnosis Models with LLM-based Semantic Augmentation [Full]

Youheng Bai (Jinan University); Jiaqi Zheng (Jinan University); Mingliang Hou (Jinan University); Teng Guo (Jinan University); Mi Tian (TAL Education Group); Xiangyu Zhao (City University of Hong Kong); Zitao Liu (Jinan University); Weiqi Luo (Jinan University)

Cognitive diagnosis aims to infer students' mastery levels over knowledge components from their learning interactions, supporting personalized

education applications. However, existing models encode responses as binary correctness labels, discarding information about which specific option a student selected. This input-level information loss limits their ability to distinguish qualitatively different error types. As a result, providing interpretable diagnostic outputs becomes challenging. To address these limitations, we propose SACD, a semantic-augmented cognitive diagnosis framework that integrates LLM-based semantic analysis with student behavioral modeling. SACD comprises the following key components. First, an LLM-based exercise diagnostic generator analyzes exercise content and produces structured semantic annotations for each answer choice, capturing the specific misconceptions each option represents. Second, a kernel-based alignment mechanism projects semantic embeddings and behavioral representations into a unified kernel space, enabling effective fusion of heterogeneous information. Third, an interpretable diagnosis layer predicts student performance and generates fine-grained mastery estimates, which LLMs further process to produce actionable learning plans. Extensive experiments on three real-world datasets demonstrate that SACD achieves superior prediction accuracy while enabling interpretable, actionable diagnostics.

Offline-First Information Retrieval for Curriculum-Aligned STEM Resources in Low-Resource Namibian Schools [Low-Resource]

Josephina Muntuumo (University of Eastern Finland); Wilbard Lazarus (Namibia University of Science and Technology); Naftali Indongo (Namibia University of Science and Technology)

Many Namibian schools face recurring connectivity and infrastructure constraints, yet ICT-in-education initiatives often assume always-on internet and cloud-based platforms for STEM teaching and learning. This paper asks what changes when the school, rather than the cloud, is treated as the primary locus of information retrieval. We describe the design of an offline-first IR prototype that hosts a curriculum-aligned STEM repository locally and supports keyword search, BM25 ranking, and curriculum-aware facets. Framing the system explicitly as IR raises questions that current offline educational platforms tend to leave under-specified: how should ranking and faceted search be designed when bandwidth, devices, storage, and power are scarce; how might teachers' and learners' retrieval behaviour change when queries remain inside the school network; and how can such systems be evaluated when large-scale logging and A/B testing are unlikely to be feasible in many classrooms? We outline the system architecture and a mixed-methods evaluation plan combining local interaction logs, teacher-rated relevance judgements, and qualitative classroom evidence.

Local Information Access in Marathi: Evaluating LLM-Native Web Retrieval in a Low-Resource Environment [Low-Resource]

Sahil Kale (Independent Researcher)

Web-augmented large language models (LLMs) promise end-to-end retrieval agents for time-sensitive information access, yet their reliability in low-resource environments remains unclear. We introduce MarathiWeb, a benchmark for evaluating LLM-native web retrieval in Marathi that separates translated global queries from natively authored, local information needs. We evaluate two frontier LLMs with built-in web search against a controlled Google-based retrieval baseline, and analyse retrieval depth, failure modes, and cost-accuracy trade-offs. Despite high tool invocation rates, accuracy drops sharply in Marathi, most severely for local queries, with the dominant failures arising from evidence integration rather than missing sources. We further find that shallow retrieval (one web call) is consistently the most accurate and cost-efficient strategy, while deeper multi-step search degrades performance. MarathiWeb exposes persistent gaps in local-language information access and provides an open benchmark for low-resource retrieval research.

Retrieval for User-Centered Translation: Lessons from RAG-based Tools for Low-Resource Domains [Low-Resource]

Raphael Merx (The University of Melbourne); Ekaterina Vylomova (The University of Melbourne)

Machine translation for low-resource languages suffers from domain-imbalanced corpora, causing quality degradation on technical text. However, in-context learning opens the possibility to rely on limited in-domain corpora to inform translation. We present lessons learned from Tulun, a retrieval-augmented system combining neural MT with LLM post-editing, guided by user-configurable translation memories and glossaries. Deployed for medical translation in Timor-Leste (Tetun) and disaster relief translation in Vanuatu (Bislama), the system achieves accuracy improvements over baseline MT by 16.90-22.41 ChrF++ points, while offering rapid adaptability and transparency to end-users. Key recommendations include: domain granularity matters more than broad categories; translation target audience should inform retrieval; and RAG-augmented MT is most effective for languages that lack domain corpora but remain within LLM pretraining distributions.

Mitigating Evidence Suppression: Bi-level Active Evidence Injection for Educational Video Understanding [Full]

Cheng Liu (Jinan University); Yiping Wang (Jinan University); Quanlong Guan (Jinan University); Chaobo He (South China Normal University); Xingyu Zhu (Jinan University); Liangda Fang (Jinan University)

Large Vision--Language Models (LVLMs) frequently fail on knowledge-intensive educational video QA despite the presence of requisite visual evidence. Through region-level analysis, we identify a systematic evidence-suppression pattern: task-critical tokens (e.g., diagrams) exhibit lower representational energy than distractors at the encoder output, rendering them prone to persistent attention neglect during decoding. While a controlled study shows that performance is sensitive to the strength and amount of injected candidate evidence signals, we find that rigid heuristics are insufficient due to sensitivity to token quality. To address this, we propose Bi-level Active Evidence Injection (BAEI), a decoding-time intervention that keeps the LVLM backbone frozen. BAEI employs a lightweight Injection Policy Network (IPN), optimized via GRPO, to dynamically select candidate evidence tokens and predict their token-wise injection strengths. The framework operates on two levels: increasing the contribution of candidate evidence-related signals in shallow layers and performing adaptive correction in deep layers triggered by predictive entropy. Experiments on educational benchmarks demonstrate consistent gains, validating decoding-time evidence intervention as an effective solution for factual alignment. The code is available at https://github.com/diojojolc-cell/BAEI.

Industry Panel Discussion · *Goldfields · Thursday 16:15–17:30*

AI Evals: Good News, Bad News, How to Win Big

For the IR community, there is both opportunity and challenge in the current AI evaluation boom: industry is highly motivated to improve evaluation practices, but assessing AI task success is far more complex than the ranked-list paradigms that have traditionally underpinned offline IR evaluation. The panel will explore how decades of IR research on conversational, task-based, and user-centred evaluation—and its culture of critically examining what is being measured and why—can help AI evaluation make substantial advances in rigour and effectiveness.

Poster Sessions

Poster Session 1 · *Tuesday 15:00–16:15 · Exhibition Hall 21B/22B · 118 boards*

P001 · **SurGE: A Benchmark and Evaluation Framework for Scientific Survey Generation** [Resource]

Weihang Su (Tsinghua University); Anzhe Xie (Tsinghua University); Qingyao Ai (Quan Cheng Laboratory); Jianming Long (Tsinghua University); Xuanyi Chen (Tsinghua University); Jiaxin Mao (Renmin University of China); Ziyi Ye (Fudan University); Yiqun Liu (Tsinghua University)

The exponential growth of scientific literature has created a pressing need for automated survey generation. Although recent LLM-based agents have shown promise in automating this task, current progress is hindered by the lack of a standardized, scalable evaluation protocol. Existing evaluation methods typically rely on either human evaluation or custom metrics designed to validate specific pipelines, which restricts scalability and hinders fair comparison. To address this, we introduce SurGE, a benchmark and evaluation framework tailored for scientific survey generation. SurGE provides a large-scale retrieval corpus of over one million papers and expert-validated ground-truth surveys. Furthermore, we propose a robust multi-dimensional evaluation protocol that integrates both objective metrics and LLM-based judgments, and empirically verify its high alignment with human experts. Our experiments reveal that while agentic pipelines outperform RAG baselines in fluency and structural quality, they still struggle with citation accuracy, highlighting key directions for future research.

P002 · **LiveRAG: A Diverse Q&A Dataset with Varying Difficulty Level for RAG Evaluation** [Resource]

David Carmel (Technology Innovation Institute (TII)); Simone Filice (Technology Innovation Institute (TII)); Guy Horowitz (Technology Innovation Institute (TII)); Yoelle Maarek (Technology Innovation Institute (TII)); Alexander Shtoff (Technology Innovation Institute (TII)); Oren Somekh (Technology Innovation Institute (TII)); Ran Tavory (Technology Innovation Institute (TII))

With Retrieval-Augmented Generation (RAG) becoming more and more prominent in generative AI solutions, there is an emerging need for systematically evaluating its effectiveness. We introduce the LiveRAG benchmark, a publicly available dataset of 895 synthetic questions and answers designed to support systematic evaluation of RAG-based Q&A systems. This synthetic benchmark is derived from the one used during the SIGIR'2025 LiveRAG challenge, where competitors were evaluated under strict time constraints. It is augmented with information that was not made available to competitors during the challenge, such as the ground-truth answers, together with their associated supporting claims which were used for evaluating competitors' answers. In addition, each question is associated with estimated difficulty and discriminability scores, derived from applying an Item Response Theory model to competitors' responses. Our analysis highlights the benchmark's question diversity, the wide range of difficulty levels, and their usefulness in differentiating between system capabilities. The LiveRAG benchmark will hopefully help the community advance RAG research, conduct systematic evaluation, and develop more robust Q&A systems.

P003 · **FinAuditing: A Financial Taxonomy-Structured Multi-Document Benchmark for Evaluating LLMs** [Resource]

Yan Wang (The Fin AI); Keyi Wang (Columbia University); Shanshan Yang (Stevens Institute of Technology); Jaisal Patel (Rensselaer Polytechnic Institute); Jeff Zhao (The University of Texas at Austin); Fengran Mo

(University of Montreal); Xueqing Peng (The Fin AI); Lingfei Qian (The Fin AI); Yankai Chen (Mohamed bin Zayed University of Artificial Intelligence); Víctor Gutiérrez-Basulto (Cardiff University); Jimin Huang (The Fin AI); Guojun Xiong (Harvard University); Xiao-Yang Liu (Columbia University); Jian-Yun Nie (University of Montreal)

Going beyond simple text processing, financial auditing requires detecting semantic, structural, and numerical inconsistencies across large-scale disclosures. As financial reports are filed in XBRL, a structured XML format governed by accounting standards, auditing becomes a structured information extraction and reasoning problem involving concept alignment, taxonomy-defined relations, and cross-document consistency. Although large language models (LLMs) show promise on isolated financial tasks, their capability in professional-grade auditing remains unclear. We introduce \textsc{\finauditing}, a taxonomy-aligned, structure-aware benchmark built from real XBRL filings. It contains 1,102 annotated instances averaging over 33k tokens and defines three tasks: \emph{Financial Semantic Matching} (FinSM), \emph{Financial Relationship Extraction} (FinRE), and \emph{Financial Mathematical Reasoning} (FinMR). Evaluations of 13 state-of-the-art LLMs reveal substantial gaps in concept retrieval, taxonomy-aware relation modeling, and consistent cross-document reasoning. These findings highlight the need for realistic, structure-aware benchmarks. We release the evaluation code\footnote{\url{https://github.com/The-FinAI/FinAuditing}} and dataset\footnote{\url{https://huggingface.co/collections/TheFinAI/finauditing}} publicly, and the task currently serves as the official benchmark of an ongoing public evaluation contest\footnote{\url{https://open-finance-lab.github.io/SecureFinAI_Contest_2026/}}.

P004 · **WeatherArchive: A Benchmark for Retrieval-Augmented Reasoning over Historical Weather Archives** [Resource]

Yongan Yu (McGill University); Xianda Du (University of Waterloo); Qingchen Hu (McGill University); Jiahao Liang (McGill University); Jingwei Ni (ETH Zürich); Dan Qiang (McGill University); Kaiyu Huang (Beijing Jiaotong University); Grant Mckenzie (McGill University); Renée Sieber (McGill University); Fengran Mo (Université de Montréal)

Historical news segments on weather events are collections of enduring primary source records that offer rich, untapped narratives of how societies have experienced and responded to extreme weather events. These qualitative accounts provide insights into societal vulnerability and resilience that are largely absent from meteorological records, making them valuable for climate scientists to understand societal responses. However, their large scale, noise in optical character recognition (OCR), and archaic language make it difficult to transform them into structured knowledge for climate research. To address this challenge, we introduce øurmethod, the first large-scale benchmark for evaluating end-to-end retrieval-augmented generation (RAG) systems on historical weather archives.
WeatherArchive-Bench comprises two tasks: WeatherArchive-Retrieval, which measures a system's ability to locate historically relevant news segments from over one million archival news segments, and WeatherArchive-Assessment, which evaluates whether Large Language Models (LLMs) can classify societal vulnerability and resilience indicators from extreme weather narratives and answer queries using the segments retrieved. Extensive experiments across sparse, dense, and re-ranking retrievers, as well as a diverse set of LLMs, reveal that dense retrievers often fail on historical terminology, while LLMs frequently misinterpret vulnerability and resilience concepts. These findings highlight key limitations in reasoning about complex societal indicators and provide insights for designing more robust climate-focused RAG systems from archival contexts. The constructed dataset and evaluation framework are available at: https://github.com/Weather-Archival-Rescue/WeatherArchive-Bench.

P005 · **JFinTEB: Japanese Financial Text Embedding Benchmark** [Resource]

Masahiro Suzuki (Amova Asset Management Co. Ltd.); Hiroki Sakaji (Hokkaido University)

We introduce JFinTEB, the first comprehensive benchmark specifically designed for evaluating Japanese financial text embeddings. Existing embedding benchmarks provide limited coverage of language-specific and domain-specific aspects found in Japanese financial texts. Our benchmark encompasses diverse task categories including retrieval and classification tasks that reflect realistic and well-defined financial text processing scenarios. The retrieval tasks leverage instruction-following datasets and financial text generation queries, while classification tasks cover sentiment analysis, document categorization, and domain-specific classification challenges derived from economic survey data. We conduct extensive evaluations across a wide range of embedding models, including Japanese-specific models of various sizes, multilingual models, and commercial embedding services. We publicly release JFinTEB datasets and evaluation framework at https://github.com/retarfi/JFinTEB to facilitate future research and provide a standardized evaluation protocol for the Japanese financial text mining community. This work addresses a critical gap in Japanese financial text processing resources and establishes a foundation for advancing domain-specific embedding research.

P006 · **MIRA: An LLM-Assisted Benchmark for Multi-Category Integrated Retrieval** [Resource]

Mehmet Deniz Türkmen (GESIS - Leibniz-Institute for the Social Sciences); Suchana Datta (University College Dublin); Dwaipayan Roy (Indian Institute of Science Education and Research); Daniel Hienert (GESIS - Leibniz-Institute for the Social Sciences); Philipp Mayr (GESIS - Leibniz-Institute for the Social Sciences); Derek Greene (University College Dublin)

Users increasingly expect modern search systems to offer a unified interface that seamlessly retrieves information from diverse data sources and formats. However, current information retrieval (IR) evaluation benchmarks have not kept pace with this development, primarily due to the lack of test collections that represent the diversity of contemporary search domains. We address this critical gap with MIRA, a novel benchmark based on a large-scale social science search platform. MIRA is designed for category-aware ranking across heterogeneous categories – Publications, Research Data, Variables, and Instruments & Tools – within a single, unified evaluation framework. The proposed collection is distinctive in several ways: (1) it is built upon real user queries, providing a more realistic basis for evaluation; (2) it covers scholarly items from four distinct categories, enabling multi-faceted evaluation; and (3) it leverages a Large Language Model to generate topic descriptions and narratives, as well as for relevance assessment with respect to these topics, substantially reducing the labor and cost of test collection generation. We release this resource to benefit the community by providing a foundational testbed for the research on multi-faceted, category-aware, integrated, or cross-category information retrieval.

P007 · **Answer First, Evidence Second? Uncovering Hidden Risks in Well-Structured AI Search Summaries** [Short]

Jinman Li (Institute of Software, Chinese Academy of Sciences); Xuanang Chen (Institute of Software, Chinese Academy of Sciences); Ruoxi Xu (Institute of Software, Chinese Academy of Sciences); Hongyu Lin (Institute of Software, Chinese Academy of Sciences); Yaojie Lu (Institute of Software, Chinese Academy of Sciences); Zecheng Fan (Institute of Software, Chinese Academy of Sciences); Xianpei Han (Institute of Software, Chinese Academy of Sciences); Le Sun (Institute of Software, Chinese Academy of Sciences)

As search engines increasingly adopt answer-centric interfaces, AI-generated summaries are often consumed as final answers. These summaries are typically well-structured and citation-rich, creating a strong appearance of reliability that can encourage user trust, even when evidential grounding is uncertain. Motivated by this gap between appearance and evidence, we analyze 14,175 real-world queries from MS MARCO by examining Google Search AI summaries and their cited sources. We find that reliable-looking summaries mask substantial evidential failures. Despite their credible appearance, 32.31% of summaries are incorrect, and 56.16% of these errors arise even when supporting evidence exists in the cited sources. Moreover, 31.08% of summaries exhibit citation inconsistencies, with conflict risk increasing as more sources are cited and as citations appear in more prominent positions. Our findings reveal systematic, user-facing risks in answer-centric search, highlighting the need for evaluation and design practices beyond surface-level reliability signals.
https://github.com/icip-cas/AISummary.

P008 · **RubricRAG: Towards Interpretable and Reliable LLM Evaluation via Domain Knowledge Retrieval for Rubric Generation** [Short]

Kaustubh Dhole (Emory University); Eugene Agichtein (Emory University)

LLMs are increasingly evaluated or trained using automated graders such as LLM-as-judges, but their scalar scores or preferences often provide little explanation of why an answer is good or bad. Rubric-based evaluation offers a more interpretable alternative by decomposing final scores into explicit criteria, enabling more consistent and controllable evaluation. However, manually creating such rubrics requires extensive domain knowledge, is labor-intensive, and difficult to scale. In this work, we study whether LLMs can generate useful instance-specific rubrics that align with human-authored rubrics and improve response evaluation. Across two rubric benchmarks and multiple prompting and post-training strategies, we find that off-the-shelf LLMs produce rubrics with limited alignment to human-written ones. In that regard, we propose RubricRAG, a simple inference-time strategy that retrieves rubrics from related queries to provide domain knowledge. RubricRAG improves both similarity to human-authored rubrics, and the accuracy of downstream LLM judges based on the generated rubrics, highlighting a promising path toward automated, scalable, and interpretable rubric-based evaluation.

P009 · **Towards Emotional Intelligence in Conversational AI: How Well Can LLMs Recognise Emotion in Conversations?** [Short]

Islam A. Hassan (School of Computer Science and Statistics, Trinity College Dublin); Yvette Graham (School of Computer Science and Statistics, Trinity College Dublin)

The ability to express contextually appropriate emotions remains a defining challenge in conversational AI. Meeting this challenge requires accurately annotated corpora, a need that inevitably collides with practical constraints. Manual annotation, while reliable, is often prohibitively expensive and time-consuming. Consequently, automatic emotion recognition has become a critical component in the fine-tuning pipeline, with Large Language Models (LLMs) emerging as a promising alternative to human annotators.Despite their considerable potential for generating training and evaluation data, the use of LLMs for this purpose introduces a subtle but significant risk: the creation of circular, potentially biased evaluation protocols that frequently go unacknowledged in empirical reporting. In this paper, we present a systematic evaluation of LLMs for conversational emotion recognition, using human annotation as a gold standard to quantify the degree to which reliance on LLMs may introduce error into both fine-tuning and evaluation workflows. We conduct a comprehensive assessment of multiple LLMs for emotion labeling across conversational contexts, and additionally examine a second critical factor: the impact of conversational context on LLM performance. Our

results indicate that although LLMs benefit from access to prior conversational context, their utilization of such context differs substantially from human conversational understanding, suggesting that LLMs may not rely on the temporal ordering of conversational context in the same way humans do.

P010 · **Sim.API: A Middleware to Simplify the Use of User Simulators for Shared Tasks in Conversational Search** [Resource]

Marcel Gohsen (Bauhaus-Universität Weimar); Nailia Mirzakhmedova (Bauhaus-Universität Weimar); Zahra Abbasiantaeb (University of Amsterdam); Johannes Kiesel (GESIS - Leibniz Institute for the Social Sciences); Simon Lupart (University of Amsterdam); Jeffrey Dalton (University of Edinburgh); Benno Stein (Bauhaus-Universität Weimar); Mohammad Aliannejadi (University of Amsterdam)

Shared tasks offer excellent opportunities for advancing scientific fields and standardizing evaluation methods through collaborative community efforts. With the recent paradigm shift in information retrieval from keyword-based to conversational search, shared tasks can serve as promising sources for standardized evaluation methods in this domain. In particular, user simulation could develop into such an evaluation standard, since it overcomes many of the shortcomings of the Cranfield-paradigm for conversational search. However, employing user simulators in a shared task poses technical and infrastructural challenges, in part because the participants' systems and the organizer-provided simulators must be "online" at the same time, and the interactions between them must be recorded in a uniform format (run submission) in order to carry out the evaluation of the submissions. In this paper, we present Sim.API, a middleware-based run submission system that exposes user simulators through simple API endpoints and records run information without having to rely on run files. This resource aims to simplify the run submission for participants and organizers of shared tasks in conversational search. The source code, resources, and documentation are available in a public repository.

P011 · **LUMI: Unsupervised Intent Clustering with Multiple Pseudo-Labels** [Short]

I-Fan Lin (Leiden University); Faegheh Hasibi (Radboud University); Suzan Verberne (Universiteit Leiden)

In this paper, we propose an intuitive, training-free and label-free method for intent clustering in conversational search. Current approaches to short text clustering use LLM-generated pseudo-labels to enrich text representations or to identify similar text pairs for pooling. The limitations are: (1) each text is assigned only a single label, and refining representations toward a single label can be unstable; (2) text-level similarity is treated as a binary selection, which fails to account for continuous degrees of similarity. Our method LUMI is designed to amplify similarities between texts by using shared pseudo-labels. We first generate pseudo-labels for each text and collect them into a pseudo-label set. Next, we compute the mean of the pseudo-label embeddings and pool it with the text embedding. Finally, we perform text-level pooling: Each text representation is pooled with its similar pairs, where similarity is determined by the degree of shared labels. Our evaluation on four benchmark sets shows that our approach achieves competitive results, better than recent state-of-the-art baselines, while avoiding the need to estimate the number of clusters during embedding refinement, as is required by most methods. Our findings indicate that LUMI can effectively be applied in unsupervised short-text clustering scenarios. Our source code is available at https://github.com/tom192180/pseudo_label

P012 · **Learning to Reason for Multi-Step Retrieval of Personal Context in Personalized Question Answering** [Short]

Maryam Amirizaniani (University of Washington); Alireza Salemi (University of Massachusetts Amherst); Hamed Zamani (University of Massachusetts Amherst)

Personalization in Question Answering (QA) requires answers that are both accurate and aligned with users' background, preferences, and historical context. Existing state-of-the-art methods primarily rely on retrieval-augmented generation (RAG) solutions that construct personal context by retrieving relevant items from the user's profile. Existing methods use the user's query directly to retrieve personal documents and such strategies often lead to surface-level personalization. We propose PR^2 Personalized Retrieval-Augmented Reasoning), a reinforcement learning framework that integrates reasoning and retrieval from personal context for personalization. PR^2 learns adaptive retrieval-reasoning policies, determining when to retrieve, what evidence to retrieve from user profiles, and how to incorporate it into intermediate reasoning steps. By optimizing multi-turn reasoning trajectories under a personalized reward function, the framework reinforces reasoning paths that better align with user-specific preferences and contextual signals reflected by the reward model. Extensive experiments on the LaMP-QA benchmark using three LLMs show that PR^2 consistently outperforms strong baselines, achieving an average relative improvement of 8.26\%-12.0\% in personalized QA.

P013 · **Efficient Memory Alignment for Long-term Conversational Information Seeking** [Short]

Qingyang Xu (Monash University); Xiao Liu (Zhejiang University); Zhouhua Fang (Ant Group); Yong Li (Ant Group); Zhiwei Liu (Ant Group); Vincent Lee (Monash University); Haishuai Wang (Zhejiang University)

Long-term conversational agents rely on personal memory to maintain coherence and personalization, yet practical systems must operate under context budgets and cope with evolving or contradictory user information. We frame persona memory as a retrieval problem over a growing memory store, and propose REMAP, a reflection-guided memory editing approach for online alignment of persona facts that selectively writes and revises memory entries based on the current dialogue evidence and retrieved related items. The method aims to preserve salient facts while reducing redundancy and resolving apparent conflicts, enabling more efficient context utilization over extended interaction horizons. Experiments on multi-session dialogue datasets show consistent gains in persona-consistent retrieval and response continuity over commonly used memory strategies, while achieving more selective memory updates under comparable operational overhead.

P014 · **PEARL: Profile-based Explainable Agent for Edge LLM Recommendation via Latent Decomposition** [Short]

Jinkwon Lee (Sungkyunkwan University); Hayoung Oh (Sungkyunkwan University)

Which on-device LLM best fits a given user? 2B-class edge models exhibit user-specific behavioral variation: one may better align with genre-sensitive recommendations, another with procedural instructions, and no single model dominates across all users or domains. We present PEARL (Profile-based Explainable Agent for edge-model Recommendation via Latent Decomposition), which selects the best-fitting 2B-class edge LLM from a structured user profile and generates a natural-language explanation grounded in interpretable latent alignment dimensions. Its core, Explainable Latent Decomposition (ELD), maps per-model profile-alignment fingerprints into a shared low-correlation latent subspace for direct profile-to-model matching. On PersonaLens (200 profiles, 8,521/1,974 TSD/TMD dialogues), PEARL achieves P = 2.23 \pm .02 on TSD, outperforming the best single-fixed edge model by +0.11 and closing 35.0% of the Oracle-UB gap; on TMD, PEARL achieves P = 2.14 \pm .03 (+0.11, 34.0% Oracle-UB gap closure). A perturbation analysis shows that masking the top-1 ELD dimension alters 60% of recommendations (Δ P = -0.43), with a modest 1.3× specificity ratio over a random-dimension control.

P015 · **AlpsBench: An LLM Personalization Benchmark for Real-Dialogue Memorization and Preference Alignment** [Resource]

Jianfei Xiao (University of Science and Technology of China); Xiang Yu (University of Science and Technology of China); Chengbing Wang (University of Science and Technology of China); Wuqiang Zheng (University of Science and Technology of China); Xinyu Lin (National University of Singapore); Kaining Liu (University of Science and Technology of China); Hongxun Ding (University of Science and Technology of China); Yang Zhang (National University of Singapore); Wenjie Wang (University of Science and Technology of China); Fuli Feng (University of Science and Technology of China); Xiangnan He (University of Science and Technology of China)

As Large Language Models (LLMs) evolve into lifelong AI assistants, LLM personalization has become a critical frontier. However, progress is currently bottlenecked by the absence of a gold-standard evaluation benchmark. Existing benchmarks either overlook personalized information management that is critical for personalization or rely heavily on synthetic dialogues, which exhibit an inherent distribution gap from real-world dialogue. To bridge this gap, we introduce \textbf{AlpsBench}, \textbf{A}n \textbf{L}LM \textbf{P}er\textbf{S}onalization benchmark derived from real-world human--LLM dialogues. AlpsBench comprises 2,500 long-term interaction sequences curated from WildChat, paired with human-verified structured memories that encapsulate both explicit and implicit personalization signals. We define four pivotal tasks---personalized information \textit{extraction}, \textit{updating}, \textit{retrieval}, and \textit{utilization}---and establish protocols to evaluate the entire lifecycle of memory management. Our benchmarking of frontier LLMs and memory-centric systems reveals that: (i) models struggle to reliably extract latent user traits; (ii) memory updating faces a performance ceiling even in the strongest models; (iii) retrieval accuracy declines sharply in the presence of large distractor pools; and (iv) while explicit memory mechanisms improve recall, they do not inherently guarantee more preference-aligned or emotionally resonant responses. AlpsBench aims to provide a comprehensive framework to accelerate research toward truly personalized AI assistants.

P016 · **Every Preference Has Its Strength: Injecting Ordinal Semantics into LLM-Based Recommenders** [Short]

Jiwon Jeong (Korea Advanced Institute of Science and Technology (KAIST)); Donghee Han (Korea Advanced Institute of Science and Technology (KAIST)); Sungrae Hong (Korea Advanced Institute of Science and Technology (KAIST)); Woosung Kang (Korea Advanced Institute of Science and Technology (KAIST)); Mun Yong Yi (Korea Advanced Institute of Science and Technology (KAIST))

Recent work has shown that large language models (LLMs) can enhance recommender systems by integrating collaborative filtering (CF) signals through hybrid prompting. However, most existing CF-LLM frameworks collapse explicit ratings into implicit or positive-only feedback, discarding the ordinal structure that conveys fine-grained preference strength. As a result, these models struggle to exploit graded semantics and nuanced preference distinctions. We propose Ordinal Semantic Anchoring (OSA), a hybrid CF-LLM framework that explicitly incorporates preference strength by modeling interaction-level user feedback. OSA represents ordinal preference levels as numeric textual tokens and uses their token embeddings as semantic anchors to align user-item interaction representations in the LLM latent space. Through strength-aware alignment across ordinal levels, OSA preserves preference semantics when integrating collaborative signals with LLMs. Experiments on multiple real-world datasets demonstrate that OSA consistently outperforms existing baselines, particularly in pairwise preference evaluation, highlighting its effectiveness in modeling fine-grained

user preferences over prior CF-LLM methods.

P017 · **An Analysis of Attribute Utilization in Side Information-integrated Sequential Recommendation** [Short]

Minje Kim (Gyeongsang National University); Wooseung Kang (Gyeongsang National University); Suwon Lee (Gyeongsang National University); Gun-Woo Kim (Gyeongsang National University); Sang-Min Choi (Gyeongsang National University)

Side Information-integrated Sequential Recommendation (SISR) incorporates item-dependent side information, such as category or genre, to complement item ID representations; however, despite numerous proposed fusion strategies, it remains unclear to what extent current SISR models actually exploit such information in practice. In this work, we present a shuffle-based analysis that examines the functional contribution of item attributes across datasets, models, and fusion strategies by randomly breaking item-attribute alignments and observing the resulting performance changes. Extensive experiments on five real-world datasets show that sensitivity to shuffling is highly heterogeneous and jointly determined by dataset characteristics, attribute types, and model design, with no fusion strategy consistently exhibiting stronger reliance on attributes. Overall, our findings indicate that the contribution of item attributes is conditional and dataset-specific rather than universal, highlighting the need for deeper theoretical and empirical analyses of the conditions under which attributes contribute to performance in SISR.

P018 · **LLM-based Semantic and ID Representations for Sequential Recommendation** [Short]

Donglin Zhou (Shenzhen University); Weike Pan (Shenzhen University); Zhong Ming (Shenzhen Technology University)

ID-based sequential recommendation (SR) methods learn user preferences based on the item-ID interaction sequences, which are often limited by the data sparsity problem. Large language model (LLM)-enhanced SR methods leverage LLM to improve the performance. Although prior works have made significant progress, there are still three key challenges: (i) how to leverage the open-world knowledge and reasoning ability of LLM to enhance item representations; (ii) how to adaptively capture semantic preferences and collaborative preferences from multi-modality sequences; and (iii) how to construct signals to optimize model training and guide the fusion of ID and semantic information. To address these challenges, we propose a novel model, i.e., LLM-based semantic and ID representations for sequential recommendation (SIDSRec). Firstly, we propose an LLM-driven knowledge enhancement module that generates textual features and transfers it to item-level semantic representations. Secondly, we design a dual-channel preference modeling module that captures collaborative preferences and semantic preferences independently, and then aggregates them via a late-fusion layer. Specifically, we introduce modern LLM architectures into recommender systems. The embedding layer introduces a mixture of experts (MoE)-based adapter to improve the discriminative ability of the semantic representations. The sequential encoder introduces a self-attention and a gating mechanism to facilitate user preference learning. Finally, we introduce the next-item prediction task and the user preference alignment task to jointly optimize model training and modality fusion. Experiments conducted on three datasets show that our SIDSRec outperforms the state-of-the-art sequential recommenders by 24.96% and 18.48% on NDCG@5 and NDCG@10 on average. The source codes are available at https://github.com/donglinzhou/SIDSRec.

P019 · **ACE: Anisotropy-Controllable Embedding for LLM-enhanced Sequential Recommendation** [Short]

Dongcheol Lee (Sungkyunkwan University); Hye-young Kim (Sungkyunkwan University); Jongwuk Lee (Sungkyunkwan University)

Recent advances in the LLM-as-Extractor paradigm leverage large language models (LLMs) to transfer semantically rich item embeddings into sequential recommendation (SR) backbones. However, LLM-generated embeddings often suffer from strong anisotropy. Most vectors are concentrated in similar directions, resulting in a geometric imbalance that makes it difficult to adapt to collaborative signals during fine-tuning. To address this challenge, we propose Anisotropy-Controllable Embedding (ACE), which explicitly controls the anisotropy of LLM-generated embeddings. Specifically, ACE utilizes a linear autoencoder (LAE) to reshape the embedding distribution while preserving its semantic structure. In this process, the L2-regularization term mitigates the anisotropy by controlling the dispersion of embedding dimensions, while the reconstruction loss maintains semantic relationships among items. That is, ACE balances geometric uniformity and semantic embedding preservation for more stable learning. Extensive experiments demonstrate that ACE consistently outperforms existing LLM-enhanced SR models, yielding improvements of up to 12.4\% and 11.8\% in Recall@20 and NDCG@20, respectively.

P020 · **WPGRec: Wavelet Packet Guided Graph Enhanced Sequential Recommendation** [Short]

Peilin Liu (Jilin University); Zhiquan Ji (Jilin University); Gang Yan (Jilin University)

Sequential recommendation aims to model users' evolving interests from noisy, non-stationary interaction streams, where long-term preferences and short-term intents alternate over time with occasional localized bursts. Existing frequency-domain methods mainly rely on either global spectral operations or filter-based wavelet processing. However, global spectral operations may mix local transients with long-range dependencies, while filter-based wavelet processing can suffer from misalignment and boundary artifacts during multi-scale reconstruction. In addition, collaborative signals from the user-item interaction graph are often injected through scale-inconsistent auxiliary modules, making cross-scale interference difficult to control. To address these issues, we propose WPGRec, a unified time-frequency and graph-enhanced framework that aligns multi-resolution temporal modeling with graph propagation at matching scales. WPGRec first applies a full-tree undecimated stationary wavelet packet transform to generate equal-length, shift-invariant subband sequences. It then performs interaction-graph propagation within each subband to inject high-order collaborative information while preserving temporal alignment. Finally, an energy- and spectral-flatness-aware gated fusion module adaptively aggregates subbands and suppresses noise-like components. Experiments on four public benchmarks show that WPGRec consistently outperforms strong sequential and graph-based baselines, highlighting the value of band-consistent structure injection and adaptive subband fusion.

P021 · **ESCOMIC: User Adaptive Explainable Search for Comic Books** [Demo]

Suraj Shashidhar (Otto-von-Guericke-University); Sayantan Polley (Otto-von-Guericke-University); Soumit Roy (Liverpool John Moores University); Andreas Nürnberger (Otto-von-Guericke-University)

Comics combine visual art, text, and narrative pacing, making them highly engaging but difficult to search using metadata or conventional content-based retrieval. Existing comic search platforms often ignore what comic fans often seek, such as a visual style, character type, or fast narrative flow. LLM or AI-based search performs well on text and images but is often opaque, leaving users unsure why results are relevant. We introduce ESCOMIC, an adaptive and explainable search system for multimodal comic content. In a query-by-example setting, ESCOMIC combines low-level visual and textual features with high-level comic facets (e.g., genre, character gender, story pace) in a facet-based re-ranking framework. Implicit feedback from simple interactions (e.g., hovering) dynamically adapts facet weights to evolving intent. ESCOMIC provides global explanations that show how facets shape ranking and local explanations that compare top-k results to the query. Due to the lack of comic retrieval benchmarks, we conduct a controlled user study with eye-tracking and appropriate baselines. Results show higher retrieval effectiveness, satisfaction, and trust with adaptive, content-aware explanations. Our findings highlight the value of transparent multimodal search and motivate future work using LLM and RAG based narrative explanations.

P022 · **FEDIN: Frequency-Enhanced Deep Interest Network for Click-Through Rate Prediction** [Short]

Zenan Dai (Tsinghua University); Jinpeng Wang (Harbin Institute of Technology, Shenzhen); Junwei Pan (Tencent); Dapeng Liu (Tencent); Lei Xiao (Tencent); Shu-Tao Xia (Tsinghua University)

Sequential recommendation models often struggle to capture latent periodic patterns in user interests, primarily due to the noise inherent in time-domain behavioral data. While frequency-domain analysis offers a global perspective to address this, existing approaches typically treat user sequences in isolation, overlooking the crucial context of the target item.In this work, we present a novel empirical observation: user attention scores exhibit distinct spectral entropy distributions when conditioned on positive versus negative target items. Specifically, true user interests manifest as highly concentrated spectral patterns with lower entropy in the frequency domain, whereas irrelevant behaviors appear as high-entropy noise.Leveraging this insight, we propose the Frequency-Enhanced Deep Interest Network (\textbf{FEDIN}). FEDIN introduces a frequency-domain branch that utilizes a target-aware spectrum filtering mechanism to isolate these periodic interest signals.Extensive experiments on three public datasets demonstrate that FEDIN consistently outperforms state-of-the-art sequential recommendation baselines, demonstrating superior robustness against noise.We have released our code at: \url{https://github.com/otokoneko/FEDIN}.

P023 · **HuffmanEmbed: Using Huffman Coding for Embedding Table Compression in Deep Learning Recommendation Models** [Short]

Chaoyi Jiang (University of Southern California); Abdulla Alshabanah (University of Southern California); Hossein Entezari Zarch (University of Southern California); Keshav Balasubramanian (University of Southern California); Murali Annavaram (University of Southern California)

Deep Learning Recommendation Models (DLRMs) have become widely popular for click-through rate (CTR) prediction tasks. These models often rely on embedding tables to enhance their predictive performance. Each row in these tables represents a trainable weight vector associated with a specific feature instance (embedding index) of a categorical feature. The embedding table sizes have grown into Terabytes over time as model accuracy improves with larger table sizes. Large models put a significant burden on GPU memory and compute resources. To solve this burden, prior work employed embedding table compression schemes, but at the cost of reduced model accuracy. This paper introduces HuffmanEmbed, a novel embedding table compression framework that improves accuracy at a given compression ratio. This work is based on the observation that not every categorical feature instance should be given the same learnable parameter width. Categorical feature instances have highly skewed access counts in any training dataset. HuffmanEmbed encodes embedding indices into varying-length unique codes based on their access frequencies. The embedding tables themselves are restructured to exploit the varying-length codes enabling significant model compression while preserving crucial information for accurate prediction. Experimental results demonstrate that HuffmanEmbed significantly outperforms existing embedding table compression methods. The source code of HuffmanEmbed is available at

https://github.com/chaoyij/HuffmanEmbed.

P024 · **KuaiLive: A Real-time Interactive Dataset for Live Streaming Recommendation** [Resource]

Changle Qu (Renmin University of China); Sunhao Dai (Renmin University of China); Ke Guo (Renmin University of China); Xiao Zhang (Renmin University of China); Liqin Zhao (Kuaishou Technology); Shijun Wang (Kuaishou Technology); Yanan Niu (Kuaishou Technology); Lantao Hu (Kuaishou Technology); Han Li (Kuaishou Technology); Jun Xu (Renmin University of China)

Live streaming platforms have become a dominant form of online content consumption, offering dynamically evolving content, real-time interactions, and highly engaging user experiences. These unique characteristics introduce new challenges that differentiate live streaming recommendation from traditional recommendation settings and have garnered increasing attention from industry in recent years. However, research progress in academia has been hindered by the lack of publicly available datasets that accurately reflect the dynamic nature of live streaming environments. To address this gap, we introduce KuaiLive, the first real-time, interactive dataset collected from Kuaishou, a leading live streaming platform in China with over 400 million daily active users. The dataset records the interaction logs of 23,772 users and 452,621 streamers over a 21-day period. Compared to existing datasets, KuaiLive offers several advantages: it includes precise live room start and end timestamps, multiple types of real-time user interactions (click, comment, like, gift), and rich side information features for both users and streamers. These features enable more realistic simulation of dynamic candidate items and better modeling of user and streamer behaviors. We conduct a thorough analysis of KuaiLive from multiple perspectives and evaluate several representative recommendation methods on it, establishing a strong benchmark for future research. KuaiLive can support a wide range of tasks in the live streaming domain, such as top-K recommendation, click-through rate prediction, watch time prediction, and gift price prediction. Moreover, its fine-grained behavioral data also enables research on multi-behavior modeling, multi-task learning, and fairness-aware recommendation. We believe that KuaiLive will serve as a valuable resource to advance the development of intelligent live streaming services. The dataset and related resources are publicly available at~\textcolor{blue}{\url{https://imgkkk574.github.io/KuaiLive}}.

P025 · **RQ-GMM: Residual Quantized Gaussian Mixture Model for Multimodal Semantic Discretization in CTR Prediction** [Short]

Ziye Tong (Tencent WeChat); Jiahao Liu (Fudan University); Weimin Zhang (Tencent WeChat); Hongji Ruan (Tencent WeChat); Derick Tang (Tencent WeChat); Zhanpeng Zeng (Tencent WeChat); Qinsong Zeng (Tencent WeChat); Peng Zhang (Fudan University); Tun Lu (Fudan University); Ning Gu (Fudan University)

Multimodal content is crucial for click-through rate (CTR) prediction. However, directly incorporating continuous embeddings from pre-trained models into CTR models yields suboptimal results due to misaligned optimization objectives and convergence speed inconsistency during joint training. Discretizing embeddings into semantic IDs before feeding them into CTR models offers a more effective solution, yet existing methods suffer from limited codebook utilization, reconstruction accuracy, and semantic discriminability. We propose RQ-GMM (Residual Quantized Gaussian Mixture Model), which introduces probabilistic modeling to better capture the statistical structure of multimodal embedding spaces. Through Gaussian Mixture Models combined with residual quantization, RQ-GMM achieves superior codebook utilization and reconstruction accuracy. Experiments on public datasets and online A/B tests on a large-scale short-video platform serving hundreds of millions of users demonstrate substantial improvements: RQ-GMM yields a 1.502% gain in Advertiser Value over strong baselines. The method has been fully deployed, serving daily recommendations for hundreds of millions of users.

P026 · **Structural and Disentangled Adaptation of Large Vision Language Models for Multimodal Recommendation** [Short]

Zhongtao Rao (The Hong Kong University of Science and Technology (Guangzhou)); Peilin Zhou (New York University Abu Dhabi); Dading Chong (Peking University); Zhiwei Chen (The Hong Kong University of Science and Technology (Guangzhou)); Shoujin Wang (University of Technology Sydney); Nan Tang (The Hong Kong University of Science and Technology (Guangzhou))

Multimodal recommendation enhances accuracy by leveraging visual and textual signals, and its success largely depends on learning high-quality cross-modal representations. Recent advances in Large Vision-Language Models (LVLMs) offer unified multimodal representation learning, making them a promising backbone. However, applying LVLMs to recommendation remains challenging due to (\textit{i}) \textit{representation misalignment}, where gaps between specific domain data and general pre-training lead to unaligned embedding spaces, and (\textit{ii}) \textit{gradient conflicts} during fine-tuning, where shared adapters cause interference and a lack of discriminative power.To address this, we propose \textbf{SDA}, a lightweight framework for \underline{\textbf{S}}tructural and \underline{\textbf{D}}isentangled \underline{\textbf{A}}daptation, which integrates two components: \emph{Cross-Modal Structural Alignment} (CMSA) and \emph{Modality-Disentangled Adaptation} (MoDA). CMSA aligns embeddings using intra-modal structures as a soft teacher, while MoDA mitigates gradient conflicts via expertized, gated low-rank paths to disentangle gradient flows. Experiments on three public Amazon datasets show \textbf{SDA} integrates seamlessly with existing multimodal and sequential recommenders, yielding average gains of \textbf{6.15\%} in Hit@10 and \textbf{8.64\%} in NDCG@10. It also achieves up to \textbf{12.83\%} and \textbf{18.70\%} gains on long-tail items. Our code and full experimental results are available at \url{https://github.com/RaoZhongtao/SDA}.

P027 · **From Raw Features to Effective Embeddings: A Three-Stage Approach for Multimodal Recipe Recommendation** [Short]

Jeeho Shin (KAIST); Kyungho Kim (KAIST); Kijung Shin (KAIST)

Recipe recommendation has become an essential task in web-based food platforms. A central challenge is effectively leveraging rich multimodal features beyond user–recipe interactions. Our analysis shows that even simple uses of multimodal signals yield competitive performance, suggesting that systematic enhancement of these signals is highly promising. We propose TESMR, a 3-stage framework for recipe recommendation that progressively refines raw multimodal features into effective embeddings through: (1) content-based enhancement using foundation models with multimodal comprehension, (2) relation-based enhancement via message propagation over user–recipe interactions, and (3) learning-based enhancement through contrastive learning with learnable embeddings. Experiments on two real-world datasets show that TESMR outperforms existing methods, achieving 7–15% higher Recall@10.

P028 · **Negotiating the Punchline: Contextual Meme Understanding via Discrete Semantic Energy Minimization** [Short]

Bingbing Wang (Harbin Institute of Technology); Zihan Wang (Harbin Institute of Technology); Zhengda Jin (Harbin Institute of Technology); Jing Li (The Hong Kong Polytechnic University); Ruifeng Xu (Harbin Institute of Technology); Min Zhang (Harbin Institute of Technology)

Contextual meme understanding decodes implicit meaning from ambiguous visual metaphors and social context, which is crucial for online communication analysis. However, existing linear reasoning paradigms reductively cast this process as deterministic decoding, overlooking the fundamental reality that meme interpretation necessitates the dynamic alignment of multimodal cues. Lacking mechanisms to measure and correct misalignment, these static models inevitably allow initial perceptual failures to cascade into irreversible hallucinations. To address this, we propose \textbf{S}emantic \textbf{E}nergy \textbf{E}ntropy \textbf{D}escent (\textbf{SEED}), a framework that reformulates contextual meme understanding as an energy minimization problem within a discrete semantic space. Specifically, SEED constructs a Convergent Heterogeneous Thought Tree (CHTT) as the optimization workspace, where an Evaluator Agent quantifies the semantic inconsistency of hypotheses via a Semantic Energy Mechanism. By calculating a Discrete Semantic Gradient as structured feedback, the framework activates a Supplier Agent to retrieve knowledge and context evidence and a Reasoner Agent to refine explanations, thereby orchestrating Semantic Descent Dynamics that iteratively drive the reasoning trajectory to converge on a stable, low-entropy interpretation. Experimental results show that SEED consistently outperforms strong baselines on both classification and generation, reducing logical hallucinations and improving cultural grounding.

P029 · **A Graph-Enhanced MLLM for Hierarchical Multimodal Emotion Understanding and Support in Conversations** [Short]

Geng Tu (Harbin Institute of Technology); Taiyu Niu (Harbin Institute of Technology); Xi Zeng (The 30th Research Institute of China Electronics Technology Group Corporation); Ruifeng Xu (Harbin Institute of Technology); Min Zhang (Harbin Institute of Technology)

Multimodal emotion dialogue system involves a hierarchical progression: Perception (Multimodal Emotion Recognition in Conversation, MERC), Reasoning (Multimodal Emotion-Cause Pair Extraction, MECPE), and Support (Multimodal Emotional Support Conversation, MESC). Although existing methods have shifted toward MLLM-based paradigms with stronger generalization ability, they still handle MERC, MECPE, and MESC as independent components. This decoupled design disrupts the hierarchical relationship among perception, reasoning, and support. Moreover, their multimodal fusion relies on coarse-grained concatenation of encoder features into the prompt, ignoring structured relational dependencies among speakers and modalities. To address these limitations, we propose GraphEMO, a unified multimodal emotion understanding and support framework based on a graph-enhanced MLLM. It models MERC, MECPE, and MESC as a residual chain of task-specific priors, enabling MESC to build on perceptual and causal representations from preceding stages. In addition, we leverages dual graph attention to capture cross- and intra-modal relational dynamics among speakers, enhancing structured relational modeling. Experiments on MERC, MECPE, and MESC benchmarks show that GraphEMO consistently improves performance on Qwen and LLaMA models, achieving state-of-the-art results.

P030 · **Visual RAG at Scale: Tile-Level Spatial Pooling for Efficient Multi-Vector Document Retrieval** [Demo]

Ara Yeroyan (ServiceTitan)

Multi-vector visual retrievers (e.g., ColPali-style late interaction models) deliver strong accuracy, but scale poorly because each page yields thousands of vectors, making indexing and search increasingly expensive. We present Visual RAG Toolkit, a practical system for scaling visual multi-vector retrieval with training-free, model-aware pooling and multi-stage retrieval. Motivated by Matryoshka Embeddings, our method performs static spatial pooling--including a lightweight sliding-window averaging

variant--over patch embeddings to produce compact tile-level and global representations for fast candidate generation, followed by exact MaxSim reranking using full multi-vector embeddings. Our design yields a quadratic reduction in vector-to-vector comparisons by reducing stored vectors per page from thousands to dozens, notably without requiring post-training, adapters, or distillation. Across experiments with interaction-style models such as ColPali and ColSmol-500M, we observe that over the limited (size) ViDoRe v2 benchmark corpus 2-stage retrieval typically preserves NDCG and Recall @5/10 with minimal degradation, while substantially improving throughput (~4x QPS); with sensitivity mainly at very large k. The toolkit additionally provides robust preprocessing--high resolution PDF to image conversion, optional margin/empty-region cropping and token hygiene (indexing only visual tokens)--and a reproducible evaluation pipeline, enabling rapid exploration of two-, three-, and cascaded retrieval variants. By emphasizing efficiency at common cutoffs (e.g., k≤10), the toolkit lowers hardware barriers and makes state-of-the-art visual retrieval more accessible in practice.

P031 · **WISE: A Multimodal Search Engine for Visual Scenes, Audio, Objects, Faces, Speech, and Metadata** [Demo]

Prasanna Sridhar (University of Oxford); Horace Lee (University of Oxford); David M. S. Pinto (University of Oxford); Andrew Zisserman (University of Oxford); Abhishek Dutta (University of Oxford)

In this paper, we present WISE, an open-source audiovisual search engine which integrates a range of multimodal retrieval capabilities into a single practical tool, accessible to users without machine learning expertise. WISE supports natural-language and reverse-image queries at both the scene level (e.g. empty street) and object level (e.g. horse) across images and videos; face-based search for specific individuals; audio retrieval of acoustic events using text (e.g. wood creak) or an audio file; search over automatically transcribed speech; and filtering by user-provided metadata. Rich insights can be obtained by combining queries across modalities -- for example, retrieving German trains from a historical archive by applying the object query "train" and the metadata query "Germany", or searching for a face in a place. By employing vector search techniques, WISE can scale to support efficient retrieval over millions of images or thousands of hours of video. Its modular architecture facilitates the integration of new audio or visual models. WISE can be deployed locally for private or sensitive collections, and has been applied to a number of disparate real-world use cases. Code is available at https://gitlab.com/vgg/wise/wise

P032 · **ADaFuSE: Adaptive Diffusion-generated Image and Text Fusion for Interactive Text-to-Image Retrieval** [Short]

Zhuocheng Zhang (Hunan University); Xingwu Zhang (Hunan University); Kangheng Liang (University of Glasgow); Guanxuan Li (Hunan University); Richard McCreadie (University of Glasgow); Zijun Long (Hunan University)

Recent advances in interactive text-to-image retrieval (I-TIR) use diffusion models to bridge the modality gap between the textual information need and the images to be searched. However, existing frameworks fuse multi-modal user feedback by simple embedding addition. In this work, we show that this basic fusion strategy indiscriminately incorporates generative noise produced by the diffusion model, leading to performance degradation for up to 55.62% of samples. We further propose ADaFuSE (Adaptive Diffusion-Text Fusion with Semantic-aware Experts), a lightweight fusion model designed to align and calibrate multi-modal views for diffusion-augmented I-TIR, which can be plugged into existing frameworks without modifying the backbone encoder. Specifically, we introduce a dual-branch fusion mechanism that employs an adaptive gating branch to dynamically balance modality reliability, alongside a semantic-aware mixture-of-experts branch to capture fine-grained cross-modal nuances. Via thorough evaluation over four standard I-TIR benchmarks, ADaFuSE achieves state-of-the-art performance, surpassing the DAR model by up to 3.49% in Hits@10 with only a 5.29% parameter increase, while exhibiting stronger robustness to noisy and longer interactive queries. These results show that generative augmentation coupled with principled fusion provides a simple, generalizable alternative to fine-tuning for tasks like product search.

P033 · **Improving Ad-hoc Search Effectiveness for Conversational Information Retrieval via Model Merging** [Short]

Ahmed Rayane Kebir (University of Toulouse, IRIT); Jose G. Moreno (University of Toulouse, IRIT); Lynda Tamine (University of Toulouse, IRIT)

Conversational information retrieval is challenging since it requires the consideration of the conversation history which potentially gives rise to topic shifts and coreference resolution across previous turns. To address these challenges, previous work mainly rely on traditional fine-tuning of ad-hoc retrievers on conversational datasets or extrapolates their generalizability through multi-tasking. However, this mainstream approach is costly—since it requires model re-training—and exhibits catastrophic forgetting, where the model loses its foundational ad-hoc retrieval performance. In this paper, we fill this gap by introducing model merging as a training-free strategy enabling the design of a single retrieval model that operates across both ad-hoc and conversational settings with no additional fine-tuning. We conduct experiments using linear and non-linear parameter-wise merging strategies—namely Model Soup and Slerp—on standard ad-hoc search and conversational retrieval datasets. Our results demonstrate that model merging significantly enhances the ad-hoc search capabilities of conversational retrievers while improving generalizability across task-specific datasets, achieving up to 15% higher NDCG@3 under zero-shot conditions.

P034 · **AgentSim: A Platform for Verifiable Agent-Trace Simulation** [Resource]

Saber Zerhoudi (University of Passau); Michael Granitzer (University of Passau); Jelena Mitrovic (University of Passau)

Training trustworthy agentic LLMs requires data that shows the grounded reasoning process, not just the final answer. Existing datasets fall short: question-answering data is outcome-only, chain-of-thought data is not tied to specific documents, and web-agent datasets track interface actions rather than the core retrieval and synthesis steps of a RAG workflow. We introduce AgentSim, an open-source platform for simulating RAG agents. It generates verifiable, stepwise traces of agent reasoning over any document collection. AgentSim uses a policy to ensure the agent widely explores the document set. It combines a multi-model validation pipeline with an active human-in-the-loop process. This approach focuses human effort on difficult steps where models disagree. Using AgentSim, we construct and release the Agent-Trace Corpus (ATC), a large collection of grounded reasoning trajectories spanning three established IR benchmarks. We make three contributions: (1) the AgentSim platform with two mechanisms, Corpus-Aware Seeding and Active Validation, that improve trace diversity and quality; (2) the Agent-Trace Corpus (ATC), over 103,000 verifiable reasoning steps spanning three IR benchmarks, with 100% grounding rate on substantive answers; and (3) a comparative behavioral analysis revealing systematic differences in how state-of-the-art models approach information seeking. Platform, toolkit, and corpus are publicly available.

P035 · **IIRSim Studio: A Dashboard for User Simulation** [Resource]

Saber Zerhoudi (University of Passau); Adam Roegiest (Zuva); Michael Granitzer (University of Passau)

User simulation is a valuable methodology for evaluation in Information Retrieval (IR), enabling low-cost experimentation and counterfactual analysis. However, existing simulation frameworks are primarily code-centric libraries that require substantial setup effort, which limits adoption and hinders reproducibility. The bottleneck is not the simulation engines themselves, but the lack of infrastructure connecting experiment design, execution, and sharing into a single verifiable workflow. This paper introduces IIRSim Studio, a web-based workbench that addresses this gap through four contributions: (1) a visual environment for composing simulation pipelines on top of simulation frameworks, serving both novices learning simulation concepts and experts piloting large-scale experiments; (2) a component lifecycle that supports authoring, versioning, and sharing custom simulation components through Git-backed storage and runtime injection; (3) a provenance model based on experiment bundles and environment templates that makes the scope of replication explicit; and (4) a shared-task workflow, demonstrated through the re-deployment of a Sim4IA micro-task. IIRSim Studio is available as a hosted service and as a portable containerized deployment.

P036 · **Query Performance Prediction under Corpus Growth in Dense Retrieval** [Short]

Kanishka Ghosh Dastidar (University of Passau); Michael Dinzinger (University of Passau); Laura Caspari (University of Passau); Jelena Mitrovi■ (University of Passau); Michael Granitzer (University of Passau)

LLM-based chatbots are increasingly augmented with retrieval mechanisms operating over web-scale corpora. Evaluating the effectiveness of these retrieval components is challenging, as explicit relevance judgments are often unavailable. Query performance prediction (QPP) addresses this limitation by providing unsupervised estimates of retrieval effectiveness. However, existing QPP methods assume a static corpus and do not account for the impact of corpus growth on query performance. In this work, we extend the QPP paradigm by studying query performance degradation under corpus inflation in dense retrieval systems. Using tiered corpora with fixed relevance judgments, we analyze how query effectiveness evolves as the corpus (index) size increases and evaluate the ability of established score-based and embedding-based post-retrieval QPP methods to predict such degradation. Our findings show that the reliability of these predictors is dependent on the dataset. We propose simple adaptations to established QPP measures, most notably a top-k vs background Wasserstein distance measure, which yield more consistent associations with degradation and outperform their original counterparts. These findings highlight limitations of several QPP approaches in large-scale, continuously expanding retrieval environments and motivate the development of corpus-growth-aware QPP measures.

P037 · **CLAX: Fast and Flexible Neural Click Models in JAX** [Resource]

Philipp Hager (Booking.com); Onno Zoeter (Booking.com); Maarten de Rijke (University of Amsterdam)

CLAX is a JAX-based library that implements classic click models using modern gradient-based optimization. While neural click models have emerged over the past decade, complex click models based on probabilistic graphical models (PGMs) have not systematically adopted gradient-based optimization, preventing practitioners from using modern deep learning frameworks while preserving the interpretability of classic models. CLAX addresses this gap by replacing EM-based optimization with direct gradient-based optimization in a numerically stable manner. The framework's modular design enables the integration of any component, from embeddings and deep networks to custom modules, into classic click models for end-to-end optimization. We demonstrate CLAX's efficiency by running experiments on the full Baidu-ULTR dataset~\cite{Zou2022Baidu} comprising over a billion user sessions in $\sim$ 2 hours on a single GPU, orders of magnitude faster than traditional EM approaches. CLAX implements ten classic click models, serving both industry practitioners seeking to

understand user behavior and improve ranking performance at scale and researchers developing new click models. CLAX is available at: \url{https://github.com/philipphager/clax}

P038 · **Orcheo: A Modular Full-Stack Platform for Conversational Search** [Resource]

Shaojie Jiang (University of Amsterdam); Svitlana Vakulenko (WU Vienna University of Economics and Business); Maarten de Rijke (University of Amsterdam)

Conversational search (CS) requires a complex software engineering pipeline that integrates query reformulation, ranking, and response generation.CS researchers currently face two barriers: the lack of a unified framework for efficiently sharing contributions with the community, and the difficulty of deploying end-to-end prototypes needed for user evaluation.We introduce Orcheo, an open-source platform designed to bridge this gap.Orcheo offers three key advantages:\begin{enumerate*}[label=(\roman*)]\item \textbf{A modular architecture} promotes component reuse through single-file node modules, facilitating sharing and reproducibility in CS research;\item \textbf{Production-ready infrastructure} bridges the prototype-to-system gap via dual execution modes, secure credential management, and execution telemetry, with built-in AI coding support that lowers the learning curve;\item \textbf{Starter-kit assets} include 45+ off-the-shelf components for query understanding, ranking, and response generation, enabling the rapid bootstrapping of complete CS pipelines.\end{enumerate*}%By treating CS pipelines as versionable, graph-structured artifacts, Orcheo greatly simplifies the software development lifecycle.We describe the framework architecture and validate Orcheo's utility through case studies that highlight modularity and ease of use.Orcheo is released as open source under the MIT License at \url{https://github.com/AI-Colleagues/orcheo}.

P039 · **Is a Busy Search Agent a Good One? Overthinking and Overretrieval at Scale** [Short]

Xin Liu (State Key Laboratory of AI Safety); Ruqing Zhang (State Key Laboratory of AI Safety); Yu-An Liu (State Key Laboratory of AI Safety); Lixin Su (Baidu Inc.); Jiafeng Guo (State Key Laboratory of AI Safety); Xueqi Cheng (State Key Laboratory of AI Safety)

Search agents enables large language models (LLMs) to iteratively interleave retrieval and reasoning, yielding strong performance on knowledge-intensive tasks. However, their multi-step autonomy also introduces substantial inefficiencies. In practice, search agents often exhibit overretrieval, where redundant or irrelevant documents are repeatedly fetched, and overthinking, where reasoning steps become excessive or unproductive. Both behaviors significantly inflate retrieval and inference cost, yet remain poorly understood, particularly under model scaling. In this work, we conduct a systematic study of overthinking and overretrieval in search agents from a scaling perspective. We formalize both phenomena at the trajectory level and propose fine-grained evaluation protocols that combine automatic statistics with LLM-based judgments. Through controlled experiments across search agents built on LLMs of varying sizes, we find that increasing model capacity generally alleviates both behaviors, but to markedly different extents. Building on these analysis results, we further propose a lightweight post-hoc reflection framework that converts the proposed evaluation signals into explicit feedback rewards to guide agents' reasoning trajectories. Our findings provide a principled foundation for diagnosing and controlling inefficiencies in search agents.

P040 · **T-RADAR: Simulating Trademark Examination as an Interactive Retrieval Interface for Conflict Risk Assessment** [Demo]

Yongdeuk Seo (Pukyong National University); Noah Lee (Pukyong National University); Hyun-seok Min (Tomocube Inc.); Sungchul Choi (Pukyong National University)

Ranked lists excel at finding candidates but provide no mechanism for judging them. In trademark clearance, where practitioners must assess likelihood of confusion across visual, phonetic, and conceptual dimensions under goods-and-services constraints, the core task is not retrieval but multi-criteria comparative judgment informed by examination practice. We propose agentic simulation as an IR interface and demonstrate it in T-RADAR, a trademark clearance system. Beyond retrieval, T-RADAR introduces simulation-driven adjudication: candidates retrieved via multimodal dual-encoder search are examined through an adversarial Examiner-versus-Applicant exchange that may cover appearance, pronunciation, concept, and goods/services factors, producing Conflict risk and Registrability scores that serve as review prioritization signals. When multiple candidates are simulated, a final aggregation step highlights high-risk pairs and produces an overall summary. The interaction proceeds in three steps. Query: users input a candidate mark and its goods designation to retrieve potentially conflicting registrations. Select and Simulate: users choose candidate pairs and launch simulated examination. Refine and Re-simulate: users adjust inputs such as the mark name or goods designation and re-run the simulation to compare how outcomes change.

P041 · **EvoMem: Timeline-Grounded Validity Adjudication for Non-Monotonic Multi-Agent Memory** [Short]

Zhenhua Wang (Huazhong University of Science and Technology); Chunlei Wang (North China Institute of Computer Systems Engineering); Hanchen Luo (North China Institute of Computer Systems Engineering); Yike Gao (North China Institute of Computer Systems Engineering); Bang Wang (Huazhong University of Science and Technology)

Multi-agent LLM systems accumulate long, non-monotonic memories where claims are revised, retracted, and later resurfaced, making final aggregation a validity-adjudication problem rather than simple summarization. However, existing agent systems provide little support for final-stage validity adjudication, leaving the final aggregator to conflate conflicting versions and resurface outdated claims. We present EvoMem, a lightweight plug-in for post-retrieval evidence structuring that converts retrieved interaction traces into an evolutionary timeline: it pools backend evidence, selects salient timepoints under a fixed budget, and summarizes each timepoint to trace updates and suppress outdated answers. We also introduce EvoMemBench, a timeline-style benchmark with controlled revisions, retractions, and cross-topic noise. Across long-context agents, RAG base agents, and agentic memory frameworks, EvoMem consistently improves answer accuracy and reduces outdated errors with modest overhead.

P042 · **iAgentBench: Benchmarking Sensemaking Capabilities of Information-Seeking Agents on High-Traffic Topics** [Resource]

Preetam Prabhu Srikar Dammu (University of Washington); Arnav Palkhiwala (University of Washington); Tanya Roosta (UC Berkeley); Chirag Shah (University of Washington)

With the emergence of search-enabled generative QA systems, users are increasingly turning to tools that browse, aggregate, and reconcile evidence across multiple sources on their behalf. Yet many widely used QA benchmarks remain answerable by retrieving a single relevant passage, making them poorly suited for measuring cross-source sensemaking, such as integrating evidence, tracking causal links, and resolving dependencies across facets of a topic. We present iAgentBench, a dynamic ODQA benchmark that targets these higher-level information needs while keeping questions natural and grounded in realistic information-seeking behavior. iAgentBench draws seed topics from real-world attention signals and uses common user intent patterns to construct user-like questions whose answers require combining evidence from multiple sources, not just extracting a single snippet. Each instance is released with traceable evidence and auditable intermediate artifacts that support contamination checks and enable fine-grained diagnosis of failures in retrieval versus synthesis. Experiments across multiple LLMs show that retrieval improves accuracy, but retrieval alone does not reliably resolve these questions, underscoring the need to evaluate evidence use, not just evidence access.

P043 · **DeepResearch-9K: A Challenging Benchmark Dataset of Deep-Research Agent** [Resource]

Tongzhou Wu (Shenzhen Technology University); Yuhao Wang (City University of Hong Kong); Xinyu Ma (China); Xiuqiang He (Shenzhen Technology University); Shuaiqiang Wang (China); Dawei Yin (China); Xiangyu Zhao (City University of Hong Kong)

Deep-research agents are capable of executing multi-step web exploration, targeted retrieval, and sophisticated question answering. Despite their powerful capabilities, deep-research agents face two critical bottlenecks: (1) the lack of large-scale, challenging datasets with real-world difficulty, and (2) the absence of accessible, open-source frameworks for data synthesis and agent training. To bridge these gaps, we first construct DeepResearch-9K, a large-scale challenging dataset specifically designed for deep-research scenarios built from open-source multi-hop question-answering (QA) datasets via a low-cost autonomous pipeline. Notably, it consists of (1) 9000 questions spanning three difficulty levels from L1 to L3 (2) high-quality search trajectories with reasoning chains from Tongyi-DeepResearch-30B-A3B, a state-of-the-art deep-research agent, and (3) verifiable answers. Furthermore, we develop an open-source training framework DeepResearch-R1 that supports (1) multi-turn web interactions, (2) different reinforcement learning (RL) approaches, and (3) different reward models such as rule-based outcome reward and LLM-as-Judge feedback. Finally, empirical results demonstrate that agents trained on DeepResearch-9K under our DeepResearch-R1 achieve state-of-the-art results on challenging deep-research benchmarks. We release the DeepResearch-9K dataset on https://huggingface.co/datasets/artillerywu/DeepResearch-9K and the code of DeepResearch-R1 on https://github.com/Applied-Machine-Learning-Lab/SIGIR2026_DeepResearch-R1.

P044 · **The Wikidata Query Logs Dataset** [Resource]

Sebastian Walter (University of Freiburg); Hannah Bast (University of Freiburg)

We present the Wikidata Query Logs (WDQL) dataset, a dataset consisting of 335k question-query pairs over the Wikidata knowledge graph. It is over 11x larger than the largest existing Wikidata datasets of similar format without relying on template-generated queries. Instead, we construct it using real-world SPARQL queries sent to the Wikidata Query Service and generate questions for them. Since these log-based queries are anonymized and therefore often do not produce results, a significant amount of effort is needed to convert them back into meaningful SPARQL queries. To achieve this, we present an agent-based method that iteratively de-anonymizes, cleans, and verifies queries against Wikidata while also generating corresponding natural-language questions. We demonstrate the benefit of this dataset for training question-answering methods. All WDQL assets, as well as the agent code, are publicly available via https://github.com/ad-freiburg/wikidata-query-logs under a permissive license.

P045 · **GraphSynthQA: Knowledge-Graph-Guided Query Synthesis and Step-Level Preference Optimization for Web**

Agents [Short]

Chiwei Zhu (University of Science and Technology of China); Mingxuan Du (University of Science and Technology of China); Benfeng Xu (University of Science and Technology of China); Shengzhuo Zhang (Beijing Yuanshi Technology Co., Ltd.); Xiaorui Wang (Beijing Yuanshi Technology Co., Ltd.); Zhendong Mao (University of Science and Technology of China)

Web browsing—widely used for information retrieval and fact verification—has become a fundamental capability of recently emerged large language model (LLM) agents, which is often elicited by training on complex questions requiring web search. However, this task faces challenges with respect to data and training: existing QA datasets are mostly 1-3 hop over closed corpora (e.g., Wikipedia); meanwhile, outcome-based on-policy RL that used by recent works is inefficient and brittle in long-horizon, tool-heavy browsing environments. To address these challenges, we introduce GraphSynthQA, a knowledge-graph (KG)—guided synthesis framework in an open-web setting. Starting from Wikidata seed entities, GraphSynthQA iteratively retrieves and verifies evidence from the internet to expand a KG, then synthesizes complex, answer-verifiable queries grounded in multi-evidence dependencies. Building on the synthesized data, we train web-browsing agents with a compute-efficient two-stage recipe: (i) cold-start supervised fine-tuning on ReAct-style trajectories, and (ii) step-level Direct Preference Optimization (DPO), where preferences are constructed offline via single-step branched rollouts that contrast candidate actions by their downstream success rates, providing dense process supervision without expensive on-policy exploration. Experiments show that our approach consistently improves performance on challenging web-browsing benchmarks and remains competitive among models of similar size.

P046 · **FAST-MEL: A Fast, Accurate, and Storage Efficient Solution for Multimodal Entity Linking** [Short]

Thomas Derrien (Univ. Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074); Laurent Amsaleg (Univ. Rennes, CNRS, Inria, IRISA - UMR 6074); Pascale Sébillot (Univ. Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074)

Multimodal entity linking (MEL) is the task that consists of matching textual and visual mentions of entities in unstructured data to their corresponding entities in a knowledge base (KB). To be effective in large-scale practical settings, MEL systems must meet three objectives: high linking accuracy, computational efficiency, and storage efficiency, i.e., a compact yet efficient index of the KB. In this paper, we highlight that state-of-the-art systems fail to simultaneously satisfy these 3 requirements. To meet this three-fold objective, we propose FAST-MEL, a lightweight encoder-based MEL solution that relies on a novel and compact fixed-size vectorized representation of both the textual and visual information of each entity or mention. It matches the accuracy of the best systems but performs three orders of magnitude faster. It also consumes one order of magnitude less storage than the fastest systems.

P047 · **BeLink: Biomedical Entity Linking Meets Generative Re-Ranking** [Short]

Darya Shlyk (University of Milan); Stefano Montanelli (University of Milan); Lawrence Hunter (University of Chicago)

Despite recent progress, Biomedical Entity Linking (BEL) with large language models (LLMs) remains computationally inefficient and challenging to deploy in practical settings. In this work, we demonstrate that instruction-tuning of open-source generative models can offer an effective solution when applied at the re-ranking stage of the BEL pipeline. We propose a set-wise instruction-tuning formulation that enables fast and accurate candidate selection. Our method demonstrates strong performance on multiple BEL benchmarks, yielding significant improvements in linking accuracy (3%–24%) while reducing inference time compared to the state-of-the-art. We integrate our generative re-ranker into BeLink, a modular, end-to-end system designed for practical real-world BEL applications.

P048 · **Diagnosing LLM Reranker Behavior Under Fixed Evidence Pools** [Short]

Baris Arat (Ozyegin University); Emre Sefer (Ozyegin University)

Standard reranking evaluations study how a reranker orders candidates returned by an upstream retriever. This setup couples ranking behavior with retrieval quality, so differences in output cannot be attributed to the ranking policy alone. We introduce a controlled diagnostic for reranking that uses Multi-News clusters as fixed evidence pools. We limit each pool to eight documents and pass identical inputs to all rankers. Within this setup, BM25 and MMR serve as interpretable reference points for lexical matching and diversity optimization. Across 345 clusters, we find that redundancy patterns vary by model: one LLM implicitly diversifies at larger selection budgets, while another increases redundancy. In contrast, LLMs underperform on lexical coverage at small selection budgets. As a result, LLM rankings diverge substantially from both baselines rather than consistently approximating either strategy. By reducing retrieval variance through fixed pools, we interpret these differences more directly as differences in ranking policy. This diagnostic is model agnostic and can be applied to any ranker, including open source systems and proprietary APIs. Our code and processed data for both the Multi-News diagnostic and the complementary TREC-DL evaluation are publicly available at https://github.com/aratbaris/llm_reranker_multinews.git.

P049 · **On the Robustness of LLM Re-Rankings** [Short]

Reyhaneh Goli (The University of Melbourne); Alistair Moffat (The University of Melbourne)

Large language models (LLMs) have become an indispensable part of document ranking systems.In this study we employ two datasets and two LLM models to explore the implications of relevant-item density in the input candidate list, and resilience to the ordering cues provided in the prompt. We found that rankings are consistent across density modes, but that re-ranking effectiveness is vulnerable to the ordering instructions provided in the prompt. In particular, in our experiments LLM re-rankers were much better at placing documents into "most relevant first" order than into "most relevant last" order. This difference has critical implications for practitioners.

P050 · **MathMex-PDF: Towards Accessible Visual Mathematics** [Demo]

Nathaniel Serrano (University of Southern Maine); Lucas Matheson (University of Southern Maine); Clayton Durepos (University of Southern Maine); Behrooz Mansouri (University of Southern Maine)

Mathematical information retrieval remains fragmented as readers commonly view math-heavy PDFs in standard viewers while relying on separate specialized formula search tools. This separation restricts exploration of documents where meaning arises from the combination of symbolic notation and natural language. MathMex-PDF is an interactive PDF reader that unifies visual mathematics extraction and multimodal retrieval within a single workflow. The system processes raw PDF pages to detect and recognize mathematical expressions, grounds those expressions in the document layout, and provides unified search over both text and formulas. Users may submit natural-language, LaTeX, or hybrid queries and perform query-by-example interactions such as click-to-search Reverse Formula Lookup directly from rendered pages. Retrieval uses a late-fusion pipeline that combines structure-aware formula matching with dense text embeddings and applies reciprocal rank fusion to adopt modality-specific rankings. Lightweight client-side rendering is paired with cached server-side processing to support responsive interaction on standard hardware. The demo provides interactive formula grounding, hybrid math–text querying, visual navigation of mathematical content, and spoken-math annotations to improve accessibility.

P051 · **EasyRAG: A Beginner-Friendly and Interactive Framework for Retrieval-Augmented Generation** [Demo]

Xuanchen Zhou (University of Tsukuba); Haitao Yu (University of Tsukuba); Kaipeng Li (University of Tsukuba); Yubo Fang (University of Tsukuba)

Retrieval-Augmented Generation (RAG) has emerged as an effective paradigm for enhancing large language models (LLMs) with external knowledge, delivering substantial performance gains without costly parameter updates. However, for beginners, RAG remains difficult to approach due to its conceptual complexity, computational requirements, and non-trivial system design choices. Motivated by this observation, we propose EasyRAG, a beginner-friendly and interactive framework designed to lower the entry barrier to understanding and experimenting with RAG systems. EasyRAG provides multi-level demonstrations of RAG, progressively supporting different depths of conceptual and practical understanding. At the first level, EasyRAG adopts a Search-API-based RAG setup, emphasizing ease of deployment. This level avoids large-scale dataset preprocessing and complex retriever configurations, enabling users to quickly grasp the fundamental differences between LLM-only generation and naive RAG. At the second level, EasyRAG implements three representative RAG methods (i.e., FiD, FiD-Light, and Stochastic RAG), which are strongly related, with each method extending the previous one. Through these implementations, users can gain deeper insights into key RAG properties, including computational efficiency trade-offs, the impact of end-to-end optimization, and source attribution. At the third level, EasyRAG supports systematic comparison of different RAG approaches on standard benchmark datasets. In addition, we provide an interactive interface that exposes fine-grained details of both training and inference, allowing users to closely examine and understand each step of the RAG pipeline. Overall, EasyRAG provides a unified platform for learning, experimenting with, and analyzing RAG systems, making it particularly suitable for beginners in RAG. The source code and a video demonstration are available at https://github.com/ii-research/EasyRAG.

P052 · **Query-Aware Context Selection for Retrieval-Augmented Generation** [Short]

Maya Iratni (University of Toulouse, IRIT); Mohand Boughanem (University of Toulouse, IRIT); Taoufiq Dkaki (University of Toulouse, IRIT)

Retrieval augmented generation (RAG) combines language models with external corpora to support knowledge-intensive tasks, such as open-domain question answering. Standard RAG systems typically employ a fixed top-k retrieval strategy, retrieving the same number of passages regardless of query needs. This can lead to either insufficient evidence, or the inclusion of irrelevant contexts that lead to a degradation of generation performance. In this work, we conduct an empirical study of how irrelevant retrieved passages affect downstream generation, analyzing their impact across multiple standard generator models. Building on these insights, we propose a lightweight, context-size classification module that dynamically predicts how much context is required based on query-specific needs. We integrate this approach into a full RAG pipeline and demonstrate improved performance over several baselines.

P053 · **Calibrating Uncertainty with Cross-Model Consistency for LLM Hallucination Mitigation** [Short]

Shu Zhou (Nanjing University); Rui Ling (Nanjing University); Junan Chen (Nanjing University); Tao Fan (Nanjing University of Finance and Economics); Hao Wang (Nanjing University)

Large Language Models (LLMs) are known to hallucinate, generating non-factual outputs that undermine user trust. Recent ensemble-based

approaches leverage uncertainty estimation to select among multiple LLM responses, achieving promising results in hallucination mitigation. However, these methods treat each model's uncertainty independently, overlooking a crucial signal: cross-model consistency. In this work, we observe that answers agreed upon by multiple models are significantly more likely to be correct---a manifestation of the ``wisdom of crowds'' principle. Leveraging this insight, we propose Consistency-Calibrated Uncertainty Fusion (CCUF), a framework that calibrates individual model uncertainties using cross-model consistency scores. When multiple models converge on the same answer, CCUF reduces the associated uncertainty estimate; when answers diverge, uncertainty remains elevated. This calibration mechanism enables more reliable answer selection for factoid question answering. Extensive experiments on TruthfulQA, TriviaQA, and FACTOR-news benchmarks demonstrate that CCUF consistently outperforms state-of-the-art hallucination mitigation methods, surpassing the previous best ensemble method UAF by 3.4% in accuracy while exceeding GPT-4 performance on TruthfulQA by 5.2%.

P054 · **Latent Retrieval Augmented Generation** [Short]

Shu Zhou (Nanjing University); Junan Chen (Nanjing University); Rui Ling (Nanjing University); Tao Fan (Nanjing University of Finance and Economics); Hao Wang (Nanjing University)

Retrieval-augmented generation (RAG) has emerged as a promising solution to enhance the reliability of large language models (LLMs) with external knowledge. Existing RAG methods operate in explicit representation spaces: in-context methods inject knowledge through text tokens in the input, while parametric methods like Parametric RAG encode documents into model parameters. Although effective, these approaches face inherent limitations. In-context injection suffers from quadratic computational complexity with context length and degraded performance in complex reasoning tasks. Parametric injection, while reducing inference costs, requires substantial storage overhead and computationally expensive offline preprocessing. More fundamentally, both paradigms rely on explicit discrete representations tokens or parameters that may introduce information bottlenecks and hinder seamless knowledge integration. To address these challenges, we introduce Latent RAG, a novel paradigm that performs knowledge injection entirely within the continuous latent space. Our approach encodes documents into ultra-compact latent representations through an offline compression phase, and directly fuses them with the LLM's hidden states via a learned injection mechanism during inference. By operating in the semantic latent space rather than explicit token or parameter spaces, Latent RAG enables more natural knowledge integration while achieving 9,200X storage reduction compared to Parametric RAG. Experimental results on multiple RAG benchmarks demonstrate that Latent RAG substantially enhances both effectiveness and efficiency. Furthermore, it can be seamlessly combined with existing in-context and parametric methods to achieve even better performance.

P055 · **Numerical Hallucinations in Retrieval-Augmented Generation: Detection and Analysis** [Short]

Sera Singha Roy (Independent Researcher)

With the rise in the usage of Retrieval-Augmented Generation (RAG) systems to improve the factual accuracy of the large language models (LLM), there still exists a concern regarding these systems producing hallucinating outputs not grounded in the retrieved documents. Although prior work has studied general hallucination detection, the specific challenge of numerical fabrication remains unquantified. This research study analyzes 500 RAG outputs using GPT-3.5-turbo on MS MARCO queries to address this specific challenge and found that 38.2% of failures involve numerical fabrication. The evaluation consists of four detection methods that span different paradigms: embedding-based (Semantic Similarity), metric-based (BERTScore), LLM-based (GPT-4o-mini), and a number-aware heuristic. The results show that all standard methods struggle with numerical hallucinations, notably GPT-4o-mini achieves only 25.7% recall on numerical failures despite being a state-of-the-art LLM judge. In contrast, the simple number-aware heuristic of this research study achieves a 100% recall on numerical failures with F1=0.616, significantly outperforming all baselines (McNemar's test, p<0.001). These findings highlight numerical fabrication as a critical gap in current hallucination detection approaches and recommend the need for specialized, number-aware methods in RAG systems.

P056 · **Attend to Fragments: How Key Information Affects Large Language Models for Factual Inconsistency Detection** [Short]

Xindi Guo (State University of New York at Binghamton); Zhen Xie (State University of New York at Binghamton); Patrick H. Chen (State University of New York at Binghamton)

\begin{abstract}As large language models (LLMs) continue to advance, a key challenge remains their tendency to hallucinate, generating fluent yet inconsistent content that lacks factual grounding. Natural language inference (NLI)-based methods, which determine whether one statement can be logically inferred from another, are widely considered the most effective for detecting input-output inconsistencies in LLMs. However, several fundamental questions, such as whether LLMs can identify relevant information to make correct factual inconsistency detections and how different arrangements of the source document affect reasoning, are not discussed in prior studies. To bridge this research gap, we design a new benchmark, \textit{\ourmethod{}}, which comprises 1032 carefully selected instances from the TRUE and ScreenEval datasets, with key information annotated. Using \ourmethod{}, we show that LLMs frequently fail to use the appropriate information to make correct decisions. In addition, we find that LLMs tend to make predictions by overemphasizing certain keywords or fragments, a new phenomenon we term \textit{``Attend to Fragments"}. We further introduce a novel token-based permutation method to identify untrustworthy inconsistencies. Experiments show that filtering out these instances improves the overall correlation by 1.3\% on the standard TRUE benchmark. The project is available at https://github.com/VibeHPC/attend-to-fragments\end{abstract}

P057 · **Mine over Yours: How Authorship Biases Evaluation in Generative Information Retrieval** [Short]

Jeongwoo Ryu (Seoul National University); Kyusik Kim (Seoul National University); Soomin Kim (Taejae University); Jinsu Eun (Seoul National University); Changhoon Oh (Yonsei University); Bongwon Suh (Seoul National University)

Generative information retrieval (GenIR) enables users to obtain synthesized information through iterative interaction with LLMs, fundamentally reshaping how AI-generated content is produced and consumed. Within this shift, users may encounter AI-generated informational content through two primary pathways: actively creating it themselves or consuming content generated by others. We examine whether authorship biases evaluation---whether users judge AI output from their own interactions more favorably than equivalent output from others. In a mixed-methods experiment (N=28, 2×2 within-subjects), participants interacted with an AI system to retrieve and craft information, then evaluated both their own result and equivalent output generated through the same process but framed as someone else's. Results reveal a selective authorship bias: participants rated self-obtained information significantly higher in quality, but showed no corresponding difference in trust. This pattern suggests that hallucination-aware skepticism constrained trust judgments, but could not prevent quality-driven selection behavior, even in the presence of information conflicts. Given that iterative interactions are inherent to GenIR, diverse interventions seem needed to support users' critical evaluation.

P058 · **Auto-Judge: A Cross-Task Benchmark for Comparing LLM Judges for Citation-Grounded RAG Systems** [Resource]

Naghmeh Farzi (University of New Hampshire); Tim Hagen (University of Kassel); Eugene Yang (Johns Hopkins University); Maik Fröbe (Friedrich-Schiller-Universität Jena); Ronak Pradeep (University of Waterloo); Hossein A. Rahmani (University College London); Xi Wang (University of Sheffield); Oleg Zendel (RMIT University); Martin Potthast (University of Kassel); Laura Dietz (University of New Hampshire)

We present the Auto-Judge resource for the meta-evaluation of automated LLM judges, especially judges that evaluate Retrieval-Augmented Generation (RAG) systems that ground their response with citations. The resource couples (i) a data release of topics, pooled RAG responses, and human judgments, with (ii) a standardized protocol and software infrastructure for implementing "LLM-as-a-judge" methods in a reproducible and extensible way, including support for parameter sweeps and variant tracking. Auto-Judge is designed to support rigorous meta-evaluation by comparing LLM judges against human assessments and against each other across different evaluation tasks, and by examining known vulnerabilities of LLM judging such as circularity, overfitting, self-preference, and content manipulation. We describe the dataset, the submission and execution workflow (including TIRA-based execution), and we provide baseline judge implementations and correlation-based results to illustrate how the benchmark can be used to develop and diagnose new judge methods.

P059 · **Automating Generation of Long-Form Queries** [Short]

Shivani Upadhyay (University of Waterloo); Daniel Campos (Zipf AI); Nandan Thakur (University of Waterloo); Ronak Pradeep (University of Waterloo); Nick Craswell (Microsoft); Jimmy Lin (University of Waterloo)

Traditional short keyword queries are increasingly being replaced by longer, more detailed queries that reflect complex and nuanced user information needs, especially in conversational assistants equipped with web search capabilities. In this work, we present a methodology for automatically generating such human-style long-form queries (narratives) by clustering raw short queries to form synthetic search sessions, designed to reflect a real user's search behavior. Based on our human interpretation study of a 50-narrative set (comprising both human-written and automated narratives), 44% of the automated narratives are misidentified as human-written, underscoring not only the realism and complexity of the generated content but also its indistinguishability from authentic human narratives. Furthermore, we share a collection of automated narratives as a testbed for evaluating LLMs on long-form question answering (QA), which was used in the TREC 2025 RAG track. Our code is available at https://github.com/castorini/narrative-generation.

P060 · **Revisiting BM25 Feedback Models using HyDE** [Short]

Nour Jedidi (University of Waterloo); Jimmy Lin (University of Waterloo)

Recent approaches that leverage large language models (LLMs) for pseudo-relevance feedback (PRF) have generally not utilized well-established feedback models like Rocchio and RM3 when expanding queries for BM25. Instead, they often opt for a simple string concatenation of the query and LLM-generated expansion content. In this paper, we revisit and systematically evaluate traditional BM25 feedback models in the context of HyDE, a popular method that enriches query representations using LLM-generated hypothetical answer documents. Experiments demonstrate a mutually beneficial relationship between BM25 feedback models and HyDE. Feedback models are more effective when provided LLM-generated documents versus top-ranked documents retrieved by BM25, while HyDE benefits from feedback algorithms that select and weight expansion terms. In fact, by incorporating BM25 feedback models within the HyDE setup, we can

further narrow the gap between BM25 and a strong "single-shot" dense retriever, without incurring the costs associated with such embedding-based retrieval methods. Ultimately, our results demonstrate that traditional BM25 feedback models can play an important role within modern LLM PRF methods and provide a simple approach to further enhance the accuracy of HyDE. Our code is available at https://github.com/nourj98/hyde-feedback.

P061 · **SOMIN: Agentic AI for Automating Professional Visibility and Countering Content Homogenization** [Demo]

Aleksandr Farseev (SoMin.ai Research); Kirill Lepikhin (SoMin.ai Research); Kevin Manuel (National University of Singapore); Maksim Gorodilov (SoMin.ai Research); Zhao Kui (ITMO University); Ilya Makarov (AXXX); Jaime Francisco Maldonado (Qubit Research); Ian Cassidy (DEFY Research); Ronan Byrne (SoMin.ai Research)

In the evolving digital economy, professional visibility on networks such as LinkedIn has become a critical determinant of career trajectory, influence, and the emerging metric of "LLM discoverability". For the general public and the research community alike, the ability to translate technical expertise into engaging market narratives is essential for professional survival. However, the vast majority of individuals lack the marketing acumen required to maintain a high-frequency, high-quality public presence. While the democratization of LLMs appears to offer a solution, these systems inherently suffer from probability maximization mechanisms that drive outputs toward the statistical mean. Consequently, standard automated systems frequently result in "lexical homogenization", where generated content becomes repetitive, generic, and devoid of the unique perspective necessary for effective personal branding. This technological limitation leads to poor content performance and fails to distinguish individual professionals from the noise of AI-generated text. To address this deficit, we introduce SOMIN, an Agentic AI-ready marketing ideation and generation platform designed to bridge the gap between individual professional expertise and effective public communication. Moving beyond general-purpose AI, SOMIN employs Dynamic Retrieval Augmented Generation (DRAG). This process anchors outputs in a curated corpus of high-impact insights, captured in near-real-time from the LinkedIn timelines of leading Information Retrieval researchers. This domain-specific alignment ensures that the system does not merely generate text, but internalizes the best practices of thought leadership, allowing users to adopt a communicative style that is both professional and optimized for engagement. By deploying this system, individuals can overcome the "cold start" problem of content creation, leveraging a tool that helps them learn from leading voices while automating the production of distinct, high-value content. During the demonstration, we will exhibit SOMIN's capability to facilitate both real-time user interaction and fully autonomous agentic workflows. Conference participants will interact with the system to ideate marketing content on the spot, utilizing the platform to analyze SIGIR themes and generate personalized LinkedIn narratives that reflect their own perspectives. Furthermore, we will demonstrate the system's "agentic readiness" by embedding SOMIN into a Zapier-orchestrated workflow. This integration highlights a fully autonomous pipeline where the platform serves as the source of technical narrative, identifying key market tensions and automatically triggering downstream agents to schedule and publish content. This approach illustrates how the system can automate 100% of the content production process, enabling professionals to build sophisticated communication strategies on the fly and secure their visibility in an increasingly AI-mediated information ecosystem.

P062 · **Clustering-Based Methods for Vector-Based Pseudo-Relevance Feedback** [Short]

Xavier Velez (Johns Hopkins University); Andrew Yates (Johns Hopkins University); Eugene Yang (Johns Hopkins University); Trevor Adriaanse (Johns Hopkins University); Sanjeev Khudanpur (Johns Hopkins University)

Prior work has shown that vector-based pseudo relevance feedback (PRF) is an effective technique for query expansion for improving retrieval results in dense information retrieval. In dense retrieval, ColBERT-PRF has emerged as a novel mechanism, using cluster centroids built from feedback documents as PRF expansion tokens and leveraging statistical information from the closest neighboring token ids to dictate how useful these expansion tokens are. While this approach has been shown to work well in the monolingual retrieval setting for English using the original ColBERT infrastructure, such systems have since evolved to improve inference speed, reduce storage and memory usage, and support cross-language (CLIR) and multilingual (MLIR) retrieval. As a result, many of these advancements have reduced the ability to utilize token-level statistics. In this work, we aim to explore how well this type of approach can adapt to dense retrieval models when it is not feasible to use surface-form information to pick discriminating expansion tokens. Furthermore, we explore alternative clustering mechanisms, such as HDBScan, to compare how different clustering methods perform at building clusters that can be useful for PRF. Experiments on MLIR, CLIR, and Report Generation tasks, such as those in the TREC 2024 NeuCLIR Report Generation Pilot Task, show that even without access to these token statistics, the use of cluster centroids for PRF can still improve nDCG and α-nDCG by up to 12%.

P063 · **Spectral Tempering for Embedding Compression in Dense Passage Retrieval** [Short]

Yongkang Li (University of Amsterdam); Panagiotis Eustratiadis (University of Amsterdam); Evangelos Kanoulas (University of Amsterdam)

Dimensionality reduction is critical for deploying dense retrieval systems at scale, yet mainstream post-hoc methods face a fundamental trade-off: principal component analysis (PCA) preserves dominant variance but underutilizes representational capacity, while whitening enforces isotropy at the cost of amplifying noise in the heavy-tailed eigenspectrum of retrieval embeddings. Intermediate spectral scaling methods unify these extremes by reweighting dimensions with a power coefficient γ, but treat γ as a fixed hyperparameter that requires task-specific tuning. We show that the optimal scaling strength γ is not a global constant: it varies systematically with target dimensionality k and is governed by the signal-to-noise ratio (SNR) of the retained subspace. Based on this insight, we propose Spectral Tempering (SpecTemp), a learning-free method that derives an adaptive $\gamma(k)$ directly from the corpus eigenspectrum using local SNR analysis and knee-point normalization, requiring no labeled data or validation-based search. Extensive experiments demonstrate that Spectral Tempering consistently achieves near-oracle performance relative to grid-searched $\gamma^*(k)$ while remaining fully learning-free and model-agnostic. Our code is publicly available at https://github.com/liyongkang123/SpecTemp.

P064 · **ColBERTSaR: Sparsified ColBERT Index via Product Quantization** [Short]

Eugene Yang (Johns Hopkins University); Andrew Yates (Johns Hopkins University); Dawn Lawrie (Johns Hopkins University); James Mayfield (Johns Hopkins University); Saron Samuel (Johns Hopkins University); Rohan Jha (Johns Hopkins University)

While ColBERT is an effective neural retrieval architecture, it requires a heavy index structure to support candidate set retrieval based on approximated token embeddings, gathering and decompressing document token embeddings, and applying the MaxSim operation. Indexes in PLAID and similar ColBERT implementations require five to ten times the disk storage of the original raw text, which limits their scalability. Furthermore, prior work has identified that the gathering and decompression stages are the primary inefficiencies at query time. Limiting the number of document tokens that must be gathered by thresholding and score approximation does not eliminate the need for the entire index to support ad hoc queries. In this work, we propose an embedding quantization approach that turns a ColBERT index into a true inverted index. We show that, theoretically, ColBERT with embedding quantization is equivalent to learned-sparse retrieval except for the scoring mechanism. Empirically, we demonstrate that our index is 50-70% smaller than a one-bit PLAID index while retaining retrieval effectiveness.

P065 · **Sparton: Fast and Memory-Efficient Triton Kernel for Learned Sparse Retrieval** [Short]

Thong Nguyen (University of Amsterdam); Cosimo Rulli (ISTI-CNR); Franco Maria Nardini (ISTI-CNR); Rossano Venturini (University of Pisa); Andrew Yates (Johns Hopkins University)

State-of-the-art Learned Sparse Retrieval (LSR) models, such as \splade, typically employ a Language Modeling (LM) head to project latent hidden states into a lexically-anchored logit matrix. This intermediate matrix is subsequently transformed into a sparse lexical representation through element-wise operations (ReLU, ■ogp) and max-pooling over the sequence dimension. Despite its effectiveness, the LM head creates a massive memory bottleneck due to the sheer size of the vocabulary (\mathcalV), which can range from 30,000 to over 250,000 tokens in recent models. Materializing this matrix creates a significant memory bottleneck, limiting model scaling. The resulting I/O overhead between operators further throttles throughput and runtime performance. In this paper, we propose SPARTON, a fast---memory-efficient---Triton kernel tailored for the LM head in LSR models. SPARTON utilizes a fused approach that integrates the tiled matrix multiplication, ReLU, Log1P, and max-reduction into a single GPU kernel. By performing an early online reduction directly on raw logit tiles, SPARTON avoids materializing the full logit matrix in memory. Our experiments demonstrate that the SPARTON kernel, in isolation, achieves up to a 4.8× speedup and an order-of-magnitude reduction in peak memory usage compared to PyTorch baselines. Integrated into SPLADE (|~\mathcalV | \approx 30\textk), SPARTON enables a 33% larger batch size and 14% faster training with no effectiveness loss. On a multilingual backbone (|\mathcalV | \approx 250\textk), these gains jump to a 26× larger batch size and 2.5× faster training.

P066 · **Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning** [Short]

Shengyao Zhuang (The University of Queensland); Xueguang Ma (University of Waterloo); Zheng Yao (The University of Queensland); Shuai Wang (The University of Queensland); Bevan Koopman (CSIRO); Jimmy Lin (University of Waterloo); Guido Zuccon (The University of Queensland)

We introduce Rank-R1, an LLM-based reranker that reasons over queries and candidate documents before ranking. Existing rerankers based on LLMs rely on prompting or fine-tuning to order documents by relevance without explicit reasoning. Rank-R1 uses reinforcement learning with only relevance labels (no reasoning supervision) to learn an intermediate reasoning step that improves ranking. Experiments show that Rank-R1 is competitive with supervised fine-tuning on in-domain TREC DL queries. On the out-of-domain BRIGHT benchmark, which requires complex reasoning, Rank-R1 outperforms both zero-shot and supervised baselines; the 14B model surpasses RankGPT4. With a scaled configuration combining a 32B backbone, improved first-stage retrieval, and harder training data, Rank-R1 achieves state-of-the-art performance on BRIGHT.

P067 · **LTRR: Learning To Rank Retrievers for LLMs** [Short]

To Eun Kim (Carnegie Mellon University); Fernando Diaz (Carnegie Mellon University)

Retrieval-Augmented Generation (RAG) systems typically rely on a single fixed retriever, despite growing evidence that no single retriever performs optimally across all query types. In this paper, we explore a query routing

approach that dynamically selects from a pool of retrievers based on the query, using both train-free heuristics and learned routing models. We frame routing as a learning-to-rank problem and introduce LTRR, a framework that Learns To Rank Retrievers according to their expected contribution to downstream RAG performance. Through experiments on diverse question-answering benchmarks with controlled variations in query types, we demonstrate that routing-based RAG consistently surpasses the strongest single-retriever baselines. The gains are particularly substantial when training with the Answer Correctness (AC) objective and when using pairwise ranking methods, with XGBoost yielding the best results. Additionally, our approach exhibits stronger generalization to out-of-distribution queries. Overall, our results underscore the critical role of both training strategy and optimization metric choice in effective query routing for RAG systems.

P068 · **Layer-wise Token Compression for Efficient Document Reranking** [Short]

Shengyao Zhuang (Amazon AGI); Zhichao Xu (Amazon AWS); Ivano Lauriola (Amazon AGI)

Transformer-based document cross-encoder rerankers are a central component of modern information retrieval systems. Despite their success, these models suffer from high computational costs due to processing long query-document sequences at inference time. A known approach to improve efficiency is token compression, which consists of aggregating groups of tokens together in the initial embedding layer, reducing the effective number of tokens and making the computation faster. While token compression has proven to be successful for bi-encoder retrievers, we empirically observed that this approach may be ineffective for cross-encoder rerankers. In this paper, we propose Layer-wise Token Compression (LTC), which applies adaptive token pooling at intermediate transformer layers. Through extensive ablation studies on MS MARCO passage and document ranking tasks, we demonstrate that compression at middle layers preserves ranking quality while increasing inference QPS by up to 25% for passage ranking and up to 116% for document ranking. We also extend LTC to listwise LLM rerankers and show that the same approach can be easily applied to long-context listwise reranking, where the QPS improvements are even greater. More surprisingly, when applying rerankers trained on short passages to long-document ranking tasks, models trained with compression outperform their uncompressed counterparts, suggesting that compression may act as a beneficial regularizer that encourages length-invariant representations.

P069 · **Learning Evidence of Depression Symptoms via Prompt Induction** [Short]

Eliseo Bao (IRLab, CITIC, Universidade da Coruña); Anxo Perez (IRLab, CITIC, Universidade da Coruña); David Otero (IRLab, CITIC, Universidade da Coruña); Javier Parapar (IRLab, CITIC, Universidade da Coruña)

Depression places substantial pressure on mental health services, and many people describe their experiences outside clinical settings in high-volume user-generated text (e.g., online forums and social media). Automatically identifying clinical symptom evidence in such text can therefore complement limited clinical capacity and scale to large populations. We address this need through sentence-level classification of 21 depression symptoms from the BDI-II questionnaire, using BDI-Sen, a dataset annotated for symptom relevance. This task is fine-grained and highly imbalanced, and we find that common LLM approaches (zero-shot, in-context learning, and fine-tuning) struggle to apply consistent relevance criteria for most symptoms. We propose Symptom Induction (SI), a novel approach which compresses labeled examples into short, interpretable guidelines that specify what counts as evidence for each symptom and uses these guidelines to condition classification. Across four LLM families and eight models, SI achieves the best overall weighted F1 on BDI-Sen, with especially large gains for infrequent symptoms. Cross-domain evaluation on an external dataset further shows that induced guidelines generalize across other diseases shared symptomatology (bipolar and eating disorders).

P070 · **HALO: Hyperbolic Adaptation via LoRA Overlay for Hierarchy-Aware Cross-Modal Retrieval** [Short]

Teng Long (University of Amsterdam); Andrew Yates (Johns Hopkins University)

The same image can be described at multiple levels of detail (e.g., \emph{``a dog"} vs.\ \emph{``a golden retriever sleeping on a couch"}), and humans naturally organize these descriptions into a hierarchy from coarse to fine. A practical retrieval system would better align with human recognition if respecting this semantic hierarchy during search. Hyperbolic geometry is well-suited to represent such hierarchical relations, but due to the geometry discrepancy, existing hyperbolic VLMs often demand costly retraining. We propose HALO (\textbf{H}yperbolic \textbf{A}daptation via \textbf{L}ifted \textbf{O}verlay), a lightweight hyperbolic retrieval framework that equips Euclidean-pretrained VLMs with hierarchy-aware search at low computational overhead, without compromising retrieval performance. We shift Lorentzian modeling to the embedding space where cross-modal similarity is computed, while keeping parameter updates fully Euclidean and LoRA-friendly. This design enables plug-and-play fine-tuning without entailment-style auxiliary losses used by prior hyperbolic VLMs. Empirically, HALO matches strong Euclidean pretrained baselines on retrieval performance.\,Across CLIP ViT-B/32, ViT-B/16, and ViT-L/14, our setting improves zero-shot recall by $\sim$ 70--90\% over pretrained hyperbolic (hierarchy-aware) baselines and by $\sim$ 2\%--6\% over standard pre-trained CLIP baselines, while equipping euclidean CLIP with hierarchy-aware search.

P071 · **A Sensitivity-Aware Test Collection for Search Among Personal Information** [Resource]

Jack McKechnie (University of Glasgow); Graham McDonald (University of Glasgow); Craig Macdonald (University of Glasgow)

Traditional search tasks aim to satisfy user information needs by returning a subset of a collection of documents, ranked by the documents' relevance to a user query. However, some collections that contain useful information also contain \emph{sensitive} personal information. Recently, there has been increasing interest in the development of \emph{Sensitivity-Aware Search} (SAS) retrieval models to provide users with effective retrieval results without revealing such sensitive information. To develop such systems, test collections containing both sensitive and non-sensitive information, a set of queries, and query-document relevance assessments are required. The Enron email corpus contains real business-related emails, where some emails also contain sensitive personal information. However, the original Enron collection does not contain queries or query-relevance assessments. To this end, we crowdsource 150 query formulations for 50 different topics and 11,471 query-relevance assessments for a subset of the Enron documents that have been manually labelled for sensitivity. \jm{We follow best practices for using large language models (LLMs) in Information Retrieval evaluation to extend the collection further with additional LLM judged query-relevance assessments and sensitivity labels. We present baseline performances for relevance, sensitivity classification, and sensitivity-aware search on the collection.} We make the collection available, including through the popular ir_datasets package, and provide pre-built sparse and dense indices on Huggingface to facilitate easy experimentation.

P072 · **Simulating the Lateral Reader for News Trustworthiness Reports with an Iterative Multi-Agent RAG System** [Short]

Dake Zhang (University of Waterloo); Mark D. Smucker (University of Waterloo)

Readers of online news often lack the time and domain expertise required to verify unfamiliar claims and sources. Professional fact-checkers address this gap through lateral reading, an iterative workflow of asking investigative questions, searching for external evidence, and synthesizing findings with attribution. We present an iterative multi-agent Retrieval-Augmented Generation (RAG) system that operationalizes this workflow for the TREC 2025 DRAGUN Track. Given a news article, specialized agents (1) generate investigative queries, (2) retrieve and filter evidence from the MS MARCO V2.1 Segmented Corpus using a three-stage retriever (BM25+RM3, cross-encoder reranking, and LLM-based selection), and (3) apply an information-sufficiency evaluator that decides whether additional searching is required before writing. The final report generator produces a 250-word trustworthiness report grounded in retrieved segments, guided by automatically generated critical investigative questions. On the official DRAGUN rubric-based evaluation with 30 news articles, our system using GPT-4.1 ranked first on report generation quality, achieving the highest mean supportive score (0.230) with low contradiction (0.013).

P073 · **APR: Adaptive Personalised Reranking For Conversational Search** [Short]

Shen Dong (University of Glasgow); Iadh Ounis (University of Glasgow); Debasis Ganguly (University of Glasgow)

Conversational search systems help users satisfy complex information needs through natural language interactions, yet incorporating user preferences into ranking remains challenging. Existing rewrite-then-rerank pipelines capture topical relevance but struggle with fine-grained constraints such as negative preferences or formatting requirements. Instruction-following retrieval approaches are promising for enforcing such constraints, yet their use in personalised conversational search remains underexplored, since ``instructions'' within this context are implicit and situated within user history and profiles, rather than being explicitly stated. We show that instruction-following models can assist with complex queries but introduce noise and latency on simpler keyword queries. To address this issue, we propose Adaptive Personalised Reranking (APR), a framework that routes queries based on intent. APR uses efficient similarity-based reranking for simple queries and dynamically generates tailored instructions to guide an instruction-following reranker for constraint-heavy contexts. Oracle analysis on TREC iKAT 2023 and 2024 shows that instruction-following provides a ``rescue'' potential for hard queries. We also show that APR trained with synthetic data performs competitively against strong baselines such as MonoT5 while offering promising new research avenues.

P074 · **One Pass, Any Order: Position-Invariant Listwise Reranking for LLM-Based Recommendation** [Short]

Ethan Bito (RMIT University); Yongli Ren (RMIT University); Estrid He (RMIT University)

Large language models (LLMs) are increasingly used for recommendation reranking, but their listwise predictions can depend on the order in which candidates are presented. This creates a mismatch between the set-based nature of recommendation and the sequence-based computation of decoder-only LLMs, where permuting an otherwise identical candidate set can change item scores and final rankings. Such order sensitivity makes LLM-based rerankers difficult to rely on, since rankings may reflect prompt serialization rather than user preference. We propose InvariRank, a permutation-invariant listwise reranking framework that addresses this dependence at the architectural level. InvariRank blocks cross-candidate attention with a structured attention mask, and negates position-induced scoring changes through shared positional framing under Rotary Positional Embeddings (RoPE). Combined with a listwise learning-to-rank objective, the model scores all candidates in a single forward pass, avoiding

permutation-based invariance training objectives which require multiple permutations of a candidate set. Experiments on recommendation benchmarks show that InvariRank maintains competitive ranking effectiveness while producing stable rankings across candidate permutations. The results suggest that architectural invariance is a practical route to reliable and efficient LLM-based recommendation reranking. The source code is at https://github.com/ejbito/InvariRank.

P075 · **L2Rec: Towards Dual-View Understanding of LLMs for Personalized Recommendation** [Short]

Pingjun Pan (NetEase Cloud Music); Tingting Zhou (NetEase Cloud Music); Peiyao Lu (NetEase Cloud Music); Tingting Fei (NetEase Cloud Music); Hongxiang Chen (NetEase Cloud Music); Chuanjiang Luo (NetEase Cloud Music)

Adapting large language models (LLMs) for personalized recommendation requires aligning their general-purpose capabilities with user-specific preferences while effectively leveraging both behavioral and semantic signals. Existing approaches typically integrate these signals at either the input level (e.g., injecting behavioral embeddings into the token space) or the output level (e.g., contrastive alignment of separate encoders), suffering from distribution gaps or lack of end-to-end task supervision. In this work, we introduce \textbf{L2Rec}, which unifies behavioral and semantic understanding at the \emph{parameter level} of LLMs. Our key insight is that the same set of Transformer parameters can serve as a shared medium for both views: by applying view-specific, personalized low-rank perturbations via a Dual-view Personalized Mixture-of-Experts (DPMoE) mechanism, L2Rec enables a single LLM backbone to produce complementary behavioral and semantic adaptations for each user with minimal representation-level misalignment. An adaptive cross-view fusion module further integrates the dual-view outputs into a unified user preference. Experiments on four datasets show that L2Rec consistently outperforms state-of-the-art baselines, and online A/B testing on a large-scale industrial platform validates significant improvements in key engagement metrics.

P076 · **iTIMO: An LLM-empowered Synthesis Dataset for Travel Itinerary Modification** [Resource]

Zhuoxuan Huang (Central South University); Yunshan Ma (Singapore Management University); Hongyu Zhang (Central South University); Hua Ma (Hunan Normal University); Zhu Sun (Singapore University of Technology and Design)

Addressing itinerary modification is crucial for enhancing the travel experience as it is a frequent requirement during traveling. However, existing research mainly focuses on fixed itinerary planning, leaving modification underexplored due to the scarcity of shape need-to-modify itinerary data. To bridge this gap, we formally define the itinerary modification task and propose a general pipeline to construct the corresponding dataset, namely iTIMO. This pipeline frames the generation of shape need-to-modify itinerary data as an intent-driven perturbation task. It instructs large language models to perturb real-world itineraries using three operations: REPLACE, ADD, and DELETE. Each perturbation is grounded in three intents: disruptions of popularity, spatial distance, and category diversity. Furthermore, hybrid metrics are proposed to ensure perturbation effectiveness. We conduct comprehensive benchmarking on iTIMO to analyze the capabilities and limitations of state-of-the-art LLMs. Overall, iTIMO provides a comprehensive testbed for the modification task, and empowers the evolution of traditional travel recommender systems into adaptive frameworks capable of handling dynamic travel needs. Dataset, code and supplementary materials are available at https://github.com/zelo2/iTIMO.

P077 · **ZIPBid: Hierarchical Zero-shot Incremental Spend Planning for Auto-bidding** [Short]

Yunke Bai (Kuaishou Technology); Wenzheng Shu (Kuaishou Technology); Jinan Pang (Kuaishou Technology); Wentao Bai (Kuaishou Technology); Yunshan Peng (Kuaishou Technology); Ji Wu (Kuaishou Technology); Yanxiang Zeng (Kuaishou Technology); Kaiyuan Li (Kuaishou Technology); Xialong Liu (Kuaishou Technology)

The Auto-bidding plays a critical role in automating ad delivery. However, current bidding systems face two critical challenges: (i) sparse and uncertain conversions, especially the long conversion attribution latency caused by privacy policies, deep conversion paths, or marketing campaigns, collectively degrading traditional target cost-per-acquisition(CPA)-based closed-loop control, and (ii) long-term performance patterns and semantic information are valuable for bidding tasks, yet they are challenging to effectively leverage in feedback control systems.To address these limitations, we propose a hierarchical \textbf{z}ero-shot \textbf{i}ncremental spend \textbf{p}lanning for auto-\textbf{bid}ding (\M). Specifically, for high-level planning, an LLM-based analysis of the AD content and historical trends to generate robust daily spend allocations with explicit cost-tolerance bounds. As for low-level control, we utilize the Model Predictive Controller (MPC) to track allocations via real-time bid adjustments while strictly adhering to advertiser objectives. This design enables low-frequency LLM inference, making it suitable for industrial deployment. Deployed in production systems, \M~ achieves +1.8\% improvement in conversion and significantly enhances campaign stability under conditions of sparse conversions while managing delayed feedback. This framework has been practically applied on the Online advertising platform.

P078 · **RCLRec: Reverse Curriculum Learning for Modeling Sparse Conversions in Generative Recommendation** [Short]

Yulei Huang (Alibaba International Digital Commerce Group); Hao Deng (Alibaba International Digital Commerce Group); Haibo Xing (Alibaba International Digital Commerce Group); Jinxin Hu (Alibaba International Digital Commerce Group); Chuanfei Xu (Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)); Zulong Chen (Alibaba Group); Yu Zhang (Alibaba International Digital Commerce Group); Xiaoyi Zeng (Alibaba International Digital Commerce Group)

Conversion objectives in large-scale recommender systems are sparse, making them difficult to optimize. Generative recommendation (GR) partially alleviates data sparsity by organizing multi-type behaviors into a unified token sequence with shared representations, but conversion signals remain insufficiently modeled. While recent behavior-aware GR models encode behavior types and employ behavior-aware attention to highlight decision-related intermediate behaviors, they still rely on standard attention over the full history and provide no additional supervision for conversions, leaving conversion sparsity largely unresolved. To address these challenges, we propose RCLRec, a reverse curriculum learning–based GR framework for sparse conversion supervision. For each conversion target, RCLRec constructs a short curriculum by selecting a subsequence of conversion-related items from the history in reverse. Their semantic tokens are fed to the decoder as a prefix, together with the target conversion tokens, under a joint generation objective. This design provides additional instance-specific intermediate supervision, alleviating conversion sparsity and focusing the model on the user's critical decision process. We further introduce a curriculum quality-aware loss to ensure that the selected curricula are informative for conversion prediction. Experiments on offline datasets and an online A/B test show that RCLRec achieves superior performance, with +2.09% advertising revenue and +1.86% orders in online deployment.

P079 · **Beyond Unimodal Perspectives: Generative Retrieval with Multimodal Semantics** [Short]

Jing Zhu (University of Michigan, Ann Arbor); Mingxuan Ju (Snap Inc.); Yozen Liu (Snap Inc.); Shubham Vij (Snap Inc.); Danai Koutra (University of Michigan, Ann Arbor); Neil Shah (Snap Inc.); Tong Zhao (Snap Inc.)

Generative retrieval (GR) has revolutionized recommendation systems by integrating the content of the items into semantic identifiers. However, existing frameworks predominantly isolate modalities (e.g., relying solely on text), overlooking the inherently multimodal nature of real-world items. This work addresses the underexplored challenge of Multimodal Generative Retrieval (MGR). Through a systematic analysis of Early and Late Fusion strategies, we reveal that naive integration fails due to two critical limitations: modality sensitivity, where one modality dominates the representation, and modality correspondence, where the model fails to align distinct semantic IDs across modalities. To overcome these challenges, we introduce MGR-LF++, an enhanced late fusion framework. MGR-LF++ employs contrastive modality alignment to enforce cross-modal consistency and incorporates special tokens to preserve autoregressive integrity. Extensive experiments demonstrate that MGR-LF++ achieves performance improvements of over 20% compared to unimodal and naive multimodal alternatives.

P080 · **ATVG: Agentic System for Factually Grounded Travel Advertisement Video Generation** [Demo]

Byung Eun Jeon (Snap Inc.); Xiao Bai (Snap Inc.); Wen Zhang (Snap Inc.); Jinchao Li (Snap Inc.)

Consumer content platforms and advertisers stand to benefit from generating high-quality video ads at scale, but current video generation systems often produce structurally unreliable or factually incorrect outputs that can degrade the viewer experience. We propose ATVG (Agentic system for factually grounded Travel advertisement Video Generation), an agentic system that generates travel video ads from only a city name. The system plans clips across multiple themes; grounds clips via an LLM-guided iterative image retrieval process that selects an anchor image; expands lightweight concepts into detailed video prompts via meta-prompts; and verifies generation outputs to detect visual artifacts before assembling the final video. Compared to a single-prompt baseline, our agentic approach improves multi-scene composition and factuality. We demonstrate the system through a web app that exposes intermediate artifacts and full agent logs for 20 cities, covering popular and long-tail cities around the world. The video demonstrating the web app can be found at: https://www.youtube.com/watch?v=-yh_jBDmhGo.

P081 · **A Demonstration of WikiRAG: An Evidence-based Link Prediction for Wikidata with Retrieval Augmented Generation** [Demo]

Rohan Sabu (New York University Abu Dhabi); Ola El Khatib (New York University Abu Dhabi); Djellel Difallah (New York University Abu Dhabi)

Knowledge graphs (KGs) such as Wikidata store large collections of structured facts but remain inherently incomplete. Link prediction aims to identify missing relations between entities and plays a key role in keeping KGs up to date. However, predicted links are not always reliable and must be supported by clear evidence and efficiently validated by humans before they can be added to the KG. We present an interactive demonstration of our WikiRAG (a framework that combines automatic link prediction with retrieval augmented generation) designed for Wikidata link prediction and integrating evidence-based human-in-the-loop validation. Given a head entity and a relation, the system generates candidate tail entities using knowledge graph embeddings, retrieves relevant Wikipedia passages, and applies a large language model to assess each candidate based on the retrieved evidence. The interface enables users to inspect supporting passages, validate suggested links, and export confirmed triples in batch for upload to Wikidata. Our demonstration integrates automated link prediction with evidence-driven reasoning and human-in-the-loop validation, providing a practical workflow for reliable knowledge graph completion. The system can be seen in action at https://youtu.be/UlnaxCRxlp8, with a live version available at

https://wikirag.com.

P082 · **Not All Imputations are Trustworthy: An Uncertainty-aware Multi-modal Entity Alignment Framework** [Short]

Weijie Wang (Tongji University); Shijie Luo (Tongji University); Xinyuan Lu (Tongji University); Qinpei Zhao (Tongji University); Weixiong Rao (Tongji University)

Multi-modal Entity Alignment (MMEA) aims to identify equivalent entities across diverse knowledge graphs by leveraging structural, attribute, and visual information. However, real-world datasets frequently suffer from missing modalities, necessitating feature imputation. A critical yet underexplored issue is that not all imputed modalities are inherently trustworthy. Ignoring the aleatoric uncertainty of such modalities introduces severe noise which propagates through the fusion process and degrades alignment performance. To address this challenge, we propose a novel MMEA framework, namely SURE, to Suppress Uncertainty for tRustworthy Entity alignment. Specifically, SURE introduces an uncertainty-aware variational imputation module to estimate the aleatoric uncertainty of generated features. Crucially, rather than using these imputation features blindly, SURE leverages the estimated uncertainty to suppress noise propagation via a confidence-gated multi-modal fusion and an adaptive contrastive learning objective. Extensive experiments on DBP15K datasets demonstrate that SURE significantly outperforms state-of-the-art baselines, exhibiting exceptional robustness particularly in scenarios with high modality missing rates.

P083 · **Retrieval-Augmented Contrastive Learning for Dynamic Graph Anomaly Detection** [Short]

Kamal Berahmand (RMIT University); Saman Forouzandeh (RMIT University); Mehrnoush Mohammadi (The University of Queensland); Mahdi Jalili (RMIT University)

Detecting anomalous nodes in continuously evolving graphs without labeled supervision requires representations that capture both local temporal context and globally consistent normal behavior—a combination that current methods do not jointly address. Existing dynamic anomaly detectors rely on localized temporal neighborhoods and cannot leverage globally similar normal patterns elsewhere in the graph, while existing retrieval-augmented graph methods either require labels or do not enforce strict temporal causality during retrieval. We propose DGRA-CL (Dynamic Graph Retrieval-Augmented Contrastive Learning), an unsupervised framework that learns discriminative temporal node representations for anomaly detection without labeled data. DGRA-CL transforms dynamic graphs into temporal sequences, employs time- and context-aware contrastive learning to learn normal node behavior patterns, retrieves similar normal exemplars from a training pool under a strict causality constraint, and fuses them via similarity-weighted aggregation to construct baseline representations. Anomalies are detected via deviation-based scoring measuring distance from these normal baselines. On four real-world dynamic graphs, DGRA-CL achieves statistically significant AUC gains of 1–2 points over the strongest baselines on three of four benchmarks (UCI Messages, Bitcoin-Alpha, Digg) and competitive performance on Reddit, while operating without anomaly labels and generalizing to unseen nodes.

P084 · **OpeNTF2: Fairness-aware Graph Neural Team Formation** [Resource]

Hamed Loghmani (University of Windsor); Md Jamil Ahmed (University of Windsor); Kap Thang (University of Windsor); Gabriel Rueda (University of Windsor); Hossein Fani (University of Toronto)

We contribute OpeNTF2, a unified framework for neural team formation, which achieves state-of-the-art efficacy while improving efficiency in recommending teams of experts who are almost surely successful at completing complex tasks. OpeNTF2 extends its initial release in four prime directions: it (1) integrates fairness-aware debiasing rerankers to mitigate popularity and gender disparities; (2) models team formation as link prediction on expert–skill collaboration graphs to capture multi-hop cross-team relational dependencies via graph neural networks; (3) formulates team formation as a sequence prediction task to leverage encoder-decoder recurrent models and transformers; and (4) introduces a temporal training strategy to account for the evolution of experts' skills and collaboration ties. While prior libraries support reproducible pipelines, none incorporate fairness, sequential and graph-based modeling, and temporal dynamics within a single framework. The codebase, along with the installation and execution instructions as well as case studies of extension components, can be obtained under cc-by-nc-sa-4.0 license at https://github.com/fani-lab/OpeNTF/tree/v0.2.0.0.

P085 · **Adaptive Autoguidance for Item-Side Fairness in Diffusion Recommender Systems** [Short]

Zihan Li (Johannes Kepler University Linz); Gustavo Escobedo (Johannes Kepler University Linz); Marta Moscati (Johannes Kepler University Linz); Oleg Lesota (Johannes Kepler University Linz); Markus Schedl (Johannes Kepler University Linz)

Diffusion recommender systems achieve strong recommendation accuracy but often suffer from popularity bias, resulting in unequal item exposure. To address this shortcoming, we introduce A2G-DiffRec, a diffusion recommender that incorporates adaptive autoguidance, where the main model is guided by a less-trained version of itself. Instead of using a fixed guidance weight, A2G-DiffRec learns to adaptively weigh the outputs of the main and weak models during training, supervised by a fairness-aware regularization that promotes balanced exposure across items with different popularity levels. Experimental results on three public datasets show that A2G-DiffRec is effective in enhancing item-side fairness at a marginal cost of accuracy reduction compared to existing guided diffusion recommenders and other non-diffusion baselines.

P086 · **Temporal User-Agnostic Ranking: Detecting Preference Evolution while Preserving Ethical Principles** [Short]

Guilherme Ramos (Instituto Superior Técnico); Ludovico Boratto (University of Cagliari); Mirko Marras (University of Cagliari)

Online ranking systems face the challenge of incorporating temporal information while maintaining ethical principles that avoid user profiling. Traditional approaches either ignore temporal dynamics entirely or rely on user reputation schemes that raise privacy and bias concerns. In this work, we propose T-UARS (Temporal User-Agnostic Ranking System), a methodological framework that integrates temporal awareness into ranking systems without assigning scores to users. Our approach uses exponential decay weighting and adaptive change detection to distinguish between anomalous ratings and genuine temporal patterns in item evaluations. We evaluate T-UARS on four datasets with varying temporal characteristics, showing that our framework maintains robust performance across domains, adapts to temporal signals in 10–44\% of items, and remains resistant to manipulation while being computationally efficient. Our results show that temporal awareness and ethical ranking principles can be successfully combined, providing a foundation for privacy-preserving ranking systems that adapt to changing information landscapes. \textbf{Source code}: \href{https://tinyurl.com/bp87nk8f}{https://tinyurl.com/bp87nk8f}.

P087 · **Beyond Maintenance: A Benchmark and Multi-Agent Framework for Repository-Usage Code Generation** [Resource]

Kaitao Lin (South China University of Technology); Songwen Gong (South China University of Technology); Adam Jatowt (University of Innsbruck); Jiexin Wang (South China University of Technology); Yi Cai (South China University of Technology)

Repository-level code generation has attracted growing interest, yet most benchmarks and methods remain maintainer-centric, emphasizing bug fixing and feature implementation. In contrast, a common yet underexplored scenario is repository usage: external users want to build applications by correctly invoking repository-internal APIs, composing them into runnable end-to-end workflows rather than modifying the codebase. To support this setting, we introduce RUCCE, a benchmark for repository-usage code generation built from real-world Python repositories. Each instance pairs a natural-language usage instruction with grounded target APIs and a verified reference script, enabling evaluation of both API retrieval and repository-usage code generation. Building on RUCCE, we propose RUCACoder, a closed-loop multi-agent framework with a Retriever for hierarchical repository exploration, a Verifier for reranking and validation, and a Coder for feedback-driven script synthesis. Experiments across multiple backbone LLMs show that RUCACoder consistently outperforms strong retrieval and generation baselines.

P088 · **Refairmulate: A Large-Scale Dataset for Gender-Fair Query Reformulations** [Resource]

Hai Son Le (Toronto Metropolitan University); Shirin Seyedsalehi (Toronto Metropolitan University); Morteza Zihayat (Toronto Metropolitan University); Ebrahim Bagheri (University of Toronto)

Information Retrieval systems can amplify societal inequalities when training data and ranking algorithms encode biases toward certain gender identities. Query reformulation methods are known to improve retrieval effectiveness, yet they seldom treat retrieval fairness as a first-class criterion. We introduce Refairmulate, an open resource that supports training and benchmarking gender-fair query reformulation models through a multi-objective procedure that jointly balances retrieval effectiveness and gender bias. The resource contains three dataset subsets. The Optimal subset consists of 112, 261 query pairs, where each pair includes an original query and a reformulated variant that attains perfect RR@10 (=1) alongside a complete reduction of measured gender bias. The Effective subset includes 209, 343 query pairs, where the reformulated variant is guaranteed to achieve both higher retrieval effectiveness and lower gender bias than the original query, enabling direct supervision for joint optimization. The Fair subset contains 321, 604 query pairs and guarantees reduced gender bias regardless of changes in retrieval effectiveness, supporting fairness-focused training and analysis. We benchmark Refairmulate with BM25 and several dense retrievers, including SPLADE, SBERT, TCT-ColBERT, and ANCE, all of which show consistent gains, with up to 76.0% relative improvement in MRR@10 and up to 48.5% reduction in gender bias. To our knowledge, Refairmulate is the first dataset explicitly designed for fairness-aware query reformulation, providing large-scale training and benchmarking data for the community.

P089 · **LearnDCG: End-to-End Joint Optimization of Ranker and Loss in Neural Ranking** [Short]

Mohammad Hossein Saliminabi (Toronto Metropolitan University); Dimitrios Androutsos (Toronto Metropolitan University); Ebrahim Bagheri (University of Toronto)

Neural rankers are typically trained with surrogate objectives because target evaluation metrics such as Normalized Discounted Cumulative Gain (NDCG) are non-differentiable. Existing approaches rely on hand-designed surrogate losses or fixed smooth relaxations, which can be misaligned with the target metric or impose rigid inductive biases. More recent efforts to learn the loss itself (e.g., NeuralLoss) require multi-stage pretraining on synthetic data and may not match real-world relevance distributions. We introduce LearnDCG, a differentiable and learnable approximation of NDCG that unifies ranker and loss optimization within a single end-to-end pipeline. LearnDCG

parameterizes the gain, discount, and temperature components of NDCG and learns them jointly with the ranker, enabling adaptation to dataset-specific relevance patterns while preserving the theoretical guarantees of differentiable ranking formulations. Experiments on MQ2007, MQ2008, WEB10K, and YLTR show that LearnDCG consistently outperforms classical surrogate losses, differentiable relaxations, and learnable alternatives. Beyond improved effectiveness, LearnDCG eliminates synthetic pretraining and reduces training complexity, achieving state-of-the-art results with improved efficiency. Gains are statistically significant across diverse architectures, from MLPs to transformer-based rankers, demonstrating the generality of the approach.

P090 · Peerispect: Claim Verification in Scientific Peer Reviews [Demo]

Ali Ghorbanpour (Reviewerly); Soroush Sadeghian (Reviewerly); Alireza Daghighfarsoodeh (Reviewerly); Sajad Ebrahimi (University of Toronto); Negar Arabzadeh (UC Berkeley); Seyed Mohammad Hosseini (Reviewerly); Ebrahim Bagheri (University of Toronto)

Peer review is central to scientific publishing, yet reviewers frequently include claims that are subjective, rhetorical, or misaligned with the submitted work. Assessing whether review statements are factual and verifiable is crucial for fairness and accountability. At the scale of modern conferences and journals, manually inspecting the grounding of such claims is infeasible. We present Peerispect, an interactive system that operationalizes claim-level verification in peer reviews by extracting check-worthy claims from peer reviews, retrieving relevant evidence from the manuscript, and verifying the claims through natural language inference. Results are presented through a visual interface that highlights evidence directly in the paper, enabling rapid inspection and interpretation. Peerispect is designed as a modular Information Retrieval (IR) pipeline, supporting alternative retrievers, rerankers, and verifiers, and is intended for use by reviewers, authors, and program committees. We demonstrate Peerispect through a live, publicly available demo at https://app.reviewer.ly/app/peerispect and API services at https://github.com/Reviewerly-Inc/Peerispect, accompanied by a video tutorial https://www.youtube.com/watch?v=pc9RkvkUh14.

P091 · SLeDoC: System for Legal Document Comparison [Demo]

Elisei Rykov (Applied AI Institute); Nikolay Ivanov (Applied AI Institute); Kseniia Petrushina (Applied AI Institute); Maria Bandulevich (HSE); Valentin Malykh (MWS AI); Vasily Konovalov (AXXX); Alexander Panchenko (Applied AI Institute); Ilseyar Alimova (Applied AI Institute)

Frequent revisions of complex, lengthy regulatory documents in large organizations often create inconsistencies and contradictions that lawyers and auditors find difficult to detect manually. Existing tools rely on character-level diffs that fail to capture paraphrases and semantic shifts. We present SLeDoC, a system for pairwise, span-aware semantic document comparison that moves beyond token- or character-level matching to semantic judgments. SLeDoC combines information-retrieval (IR) techniques with state-of-the-art large language models (LLMs). We evaluated SLeDoC on the new RuLeSM benchmark, which contains manually annotated legal paragraph pairs. The demo video and code for SLeDoC are publicly available to facilitate reproducibility and further research~. https://github.com/s-nlp/SLeDoC.

P092 · Pretraining Exposure Explains Popularity Judgments in Large Language Models [Short]

Jamshid Mozafari (University of Innsbruck); Bhawna Piryani (University of Innsbruck); Adam Jatowt (University of Innsbruck)

Large language models (LLMs) exhibit systematic preferences for well-known entities, a phenomenon often attributed to popularity bias. However, the extent to which these preferences reflect real-world popularity versus statistical exposure during pretraining remains unclear, largely due to the inaccessibility of most training corpora. We provide the first direct, large-scale analysis of popularity bias grounded in fully observable pretraining data. Leveraging the open OLMo models and their complete pretraining corpus, Dolma, we compute precise entity-level exposure statistics across 7.4 trillion tokens. We analyze 2,000 entities spanning five types (Person, Location, Organization, Art, Product) and compare pretraining exposure against Wikipedia pageviews and two elicited LLM popularity signals: direct scalar estimation and pairwise comparison. Our results show that pretraining exposure strongly correlates with Wikipedia popularity, validating exposure as a meaningful proxy for real-world salience during the training period. More importantly, we find that LLM popularity judgments align more closely with exposure than with Wikipedia, especially when elicited via pairwise comparisons. This alignment is strongest for larger models and persists in the long tail, where Wikipedia popularity becomes unreliable. Overall, our findings demonstrate that popularity priors in LLMs are primarily shaped by pretraining statistics rather than external popularity signals, offering concrete evidence that data exposure plays a central role in driving popularity bias.

P093 · Teaching Small Models When Not to Call Functions: Structured Reasoning for Tool Refusal in Low-Resource Languages [Short]

Dung Pham Tuan Vo (UNEY); Thai Trung Tran (UNEY); Tushar Semwal (UNEY)

Tool invocation refusal - correctly refusing calls when no appropriate tool exists or tools cannot solve the task - is critical for agentic IR systems. This challenge becomes acute in low-resource languages deployed on small models, where weak multilingual representations compound the problem. We demonstrate that conventional fine-tuning shows inconsistent behavior: while improving tool-calling capability, it can degrade refusal metrics (Gemma-Thai: irrelevance -4pp, live hallucination -21pp). We propose structured reasoning forms: a training-time approach that teaches models to generate explicit reasoning about tool-invocation decisions through lightweight key-value schemas. Rather than simply adding more refusal examples, our method trains models to decompose decisions into interpretable stages (task understanding, tool assessment, risk evaluation), internalizing the reasoning process for "when NOT to call" rather than just "how to call." Moreover, standard format-compliance evaluation masks refusal failures, artificially inflating baseline scores by approximately 30 percentage points because it incorrectly counts malformed tool-call attempts as correct refusals. On two multilingual 1B-parameter models evaluated on newly created Vietnamese and Thai benchmarks (BFCL-VN and BFCL-TH), models trained to generate reasoning forms improve irrelevance detection and live hallucination by roughly +10-30pp (reaching 50-60% and 43-61% respectively) while maintaining tool-calling capability. Compared with free-form CoT reasoning, structured reasoning achieves stronger refusal performance with 1.77-1.91× fewer reasoning tokens. Perplexity analysis demonstrates reasoning fundamentally changes internal model behavior, reducing hallucination probability from 67-76% to 1.2-1.3%.

P094 · TM-Bench: Benchmarking Large Language Models on Low-Resource Traditional Mongolian [Resource]

Zhenjie Gao (Inner Mongolia University); Feilong Bao (Inner Mongolia University); Aruukhan Bai (Inner Mongolia University); Ruichen Hou (Inner Mongolia University); Xieqi Ji (Inner Mongolia University); Dabalgan Wang (Inner Mongolia University); Hugjil Ming (Inner Mongolia University); Yuan Li (Inner Mongolia University)

Large language models (LLMs) have achieved remarkable success in high-resource languages, yet their performance on Traditional Mongolian remains highly limited. A primary bottleneck is the absence of a systematic evaluation framework, which precludes quantitative comparison and obscures directions for model optimization. In this paper, we introduce TM-Bench, the first comprehensive benchmark for LLMs on Traditional Mongolian. TM-Bench adopts a hybrid construction strategy consisting of human-verified Translation-based Adaptation, Expert-Original Authoring, and Semi-automated Synthesis. It comprises 18,357 instances spanning five tasks across both natural language understanding and generation to evaluate models' reasoning, knowledge application, and linguistic proficiency. We conduct systematic evaluations across representative model families. The results show that on understanding tasks, model performance lags significantly behind high-resource languages, with only a few models performing slightly above the random baseline. For generation tasks, both automatic metrics and double-blind human evaluations reveal severe semantic collapse, failing to generate coherent text and often producing unreadable gibberish. These findings underscore the critical role of TM-Bench as a foundational infrastructure for evaluating LLMs in Traditional Mongolian and catalyzing future model optimization. Our benchmark and code are available at https://github.com/gao1948083886/TM-Bench.

P095 · The Alignment Gap: A Benchmark Demonstrating the Lack of Cross-Lingual Mapping in Dialect-Specialized Language Models - The Case of Ehugbo [Resource]

Ukachi Agnes Eze-Mbey (Carnegie Mellon University Africa); Victor Tolulope Olufemi (Carnegie Mellon University Africa); Athanase Biluge Bahizire (Carnegie Mellon University Africa); Mikel K. Ngueajio (Carnegie Mellon University Africa); Prasenjit Mitra (Carnegie Mellon University Africa)

Cross-lingual information retrieval (CLIR) for low-resource dialects remains underexplored, despite millions of speakers worldwide. This work addresses a critical gap by introducing the first information retrieval (IR) benchmark resource for Ehugbo (the Afikpo dialect of Igbo with \textasciitilde 150,000 speakers in Nigeria), constructed from a high-quality parallel multimodal corpus: 1 hour of transcribed Ehugbo Bible audio aligned with standard English translations. This parallel design enables rigorous evaluation of retrieval models across language pairs. Our benchmark reveals a surprising and counterintuitive finding: the "Alignment Gap", where African-centric foundation models (Serengeti, Afro-XLMR, AfriBERTA) that excel at linguistic familiarity with Igbo achieve <5% retrieval accuracy, while global models like LaBSE achieve 85% despite less exposure to the language family. Through diagnostic analysis (t-SNE visualizations, tokenization studies, error patterns), we show that regional models lack cross-lingual alignment bridges despite deep language understanding, while global models achieve language invariance through explicit translation supervision. This finding has immediate implications for the design of multilingual systems: pre-training diversity alone is insufficient for dialectal IR. We release our Ehugbo corpus, results, and evaluation splits on GitHub to enable future work on dialect-specific fine-tuning and alignment strategies for African languages.

P096 · CTCL: A Cross-Language Benchmark for Matching Patients to Clinical Trials [Resource]

Maciej Rybinski (Universidad de Málaga); Wojciech Kusa (NASK National Research Institute); Necva Bölücü (CSIRO); Georgios Peikos (University of Milano-Bicocca); Aditya Joshi (University of New South Wales); Sarvnaz Karimi (CSIRO); Aitziber Atutxa (University of the Basque Country (UPV/EHU)); Javier Del Ser (TECNALIA); Ahmet Bölücü (Ministry of Health of Turkey); Monica Chierichetti (Independent Researcher); Pritam Dasgupta (St John of God Health Care); Nicolás Jiménez García (Hospital Universitario Costa del Sol); Borja Pedruzo (Spain); Angelika Roma■ska (Independent Researcher); Ioulia Symeonidou (University of Nicosia Medical School)

The success of clinical trials depends on the recruitment of patients who match strict inclusion criteria. The development of effective patient to clinical trial matching systems depends on benchmarking datasets that support systematic evaluation. Apart from resources that are created in English by the TREC Clinical trials tracks (2021-23), very limited corpora exist for other languages and cross-language settings, despite the need of automatic support for clinical trial recruitment being global. To address this gap, we combine machine translation with medical expert annotation to construct CTCL (Clinical Trials Cross Lingual retrieval), a cross-lingual evaluation benchmark for patient-clinical trial retrieval in seven languages. We benchmark the cross-lingual retrieval task using 14 large language (embedding) models. We showcase how our dataset can be used to evaluate the cross-lingual capability of the models for languages with varying resource availability.

P097 · **M-DaQ: Retrieving Multilingual High Quality and Diverse Samples for Instruction Fine-Tuning Datasets** [Short]

Chunguang Zhao (Huawei Technologies Ltd.); Yilun Liu (Huawei Technologies Ltd.); Pufan Zeng (University of Science and Technology of China); Yuanchang Luo (Huawei Technologies Ltd.); Shimin Tao (Huawei Technologies Ltd.); Minggui He (Huawei Technologies Ltd.); Weibin Meng (Huawei Technologies Ltd.); Song Xu (University of Science and Technology of China); Chen Liu (Huawei Technologies Ltd.); Mahong Xia (Huawei Technologies Ltd.); Zhang Li (Huawei Technologies Ltd.); Boxing Chen (Huawei Technologies Ltd.); Daimeng Wei (Huawei Technologies Ltd.)

Nultilingual instruction fine-tuning (IFT) empowers large language models to generalize across diverse linguistic and cultural contexts; however, high-quality, systematically curated multilingual IFT datasets remain scarce. To address this gap, we propose M-DaQ (Multilingual Diversity and Quality), a diversity-aware sampling framework that jointly optimizes instruction-response quality and cross-lingual semantic diversity. M-DaQ leverages a fine-tuned Quality Scoring Model alongside a maximal marginal relevance-inspired selection strategy to construct balanced, high-fidelity training data. Furthermore, we present the first systematic investigation of the Superficial Alignment Hypothesis in multilingual settings. Extensive evaluations across 18 languages demonstrate that models trained on M-DaQ-curated data achieve average win rates exceeding 60% against strong baselines on Alpaca-Eval and MT-Bench. Complementary human evaluations corroborate these gains, highlighting significant improvements in cultural relevance, contextual appropriateness, and instruction-following capability. The code are publicly released to facilitate reproducibility and future research.. https://github.com/zhaocorey/M-DaQ

P098 · **EVADE-Bench: Multimodal Benchmark for Evaluating and Enhancing Evasive Content Detection** [Resource]

Ancheng Xu (Chinese Academy of Sciences); Zhihao Yang (University of Chinese Academy of Sciences); Jingpeng Li (Alibaba Group); Guanghu Yuan (Chinese Academy of Sciences); Longze Chen (Chinese Academy of Sciences); Liang Yan (Alibaba Group); Jiehui Zhou (Alibaba Group); Zhen Qin (Alibaba Group); Hengyu Chang (Alibaba Group); Yukun Chen (Chinese Academy of Sciences); Hamid Alinejad-Rokny (University of New South Wales); Min Yang (Chinese Academy of Sciences)

E-commerce platforms increasingly rely on Large Language Models (LLMs) and Vision Language Models (VLMs) to detect illicit or misleading product content. However, these models remain vulnerable to evasive content, which refers to inputs that have been deliberately modified through techniques such as word splitting, euphemistic language, or image cropping to conceal policy violations while still conveying prohibited claims. Crucially, detecting such content requires a model to simultaneously master two capabilities: accurately comprehending complex rules, and correctly inferring the true intent behind deliberately obfuscated multimodal inputs. While prior work has separately explored LLM reasoning over complex rules and LLM-based detection of evasive content, no existing benchmark combines both within a unified evaluation framework. This gap is particularly consequential in e-commerce, where accurate moderation demands that both capabilities operate in concert. To address this gap, we introduce EVADE-Bench, the first expert-curated Chinese multimodal benchmark specifically designed to evaluate LLMs and VLMs on evasive content detection in real-world e-commerce scenarios. The dataset contains 2,833 annotated text samples and 13,961 annotated images spanning six violation categories. Our comprehensive evaluation of 26 open- and closed-source LLMs and VLMs reveals that even state-of-the-art models frequently misclassify evasive samples. We further demonstrate that clearer rule categorization significantly improves model prediction consistency and reduces false predictions, highlighting the critical role of benchmark design in enabling reliable evaluation. We analyze common error patterns across these models and identify systematic limitations in their ability to reason over metaphorical expressions and complex regulatory rules. To explore paths for performance improvement, we investigate the feasibility of multi-agent decomposition for multimodal reasoning, wherein visual description and logical inference are decoupled into separate agents, and find that this strategy yields notable accuracy gains. By releasing EVADE-Bench, we provide the first rigorous standard for evaluating evasive content detection and aim to support the development of safer and more trustworthy content moderation systems. The dataset is publicly available at https://huggingface.co/datasets/koenshen/EVADE-Bench.

P099 · **Equip Pre-ranking with Target Attention by Residual Quantization** [Short]

Yutong Li (Taobao & Tmall Group of Alibaba); Yu Zhu (Taobao & Tmall Group of Alibaba); Yichen Qiao (Shanghai Jiao Tong University); Ziyu Guan (Xidian University); Lv Shao (Taobao & Tmall Group of Alibaba); Tong Liu (Taobao & Tmall Group of Alibaba); Bo Zheng (Taobao & Tmall Group of Alibaba)

The pre-ranking stage in industrial recommendation systems faces a fundamental conflict between efficiency and effectiveness. While powerful models like Target Attention (TA) excel at capturing complex feature interactions in the ranking stage, their high computational cost makes them infeasible for pre-ranking, which often relies on simplistic vector-product models. This disparity creates a significant performance bottleneck for the entire system. To bridge this gap, we propose TARQ, a novel pre-ranking framework. Inspired by generative models, TARQ's key innovation is to equip pre-ranking with an architecture approximate to TA by Residual Quantization. This allows us to bring the modeling power of TA into the latency-critical pre-ranking stage for the first time, establishing a new state-of-the-art trade-off between accuracy and efficiency. Extensive offline experiments and large-scale online A/B tests at Taobao demonstrate TARQ's significant improvements in ranking performance. Consequently, our model has been fully deployed in production, serving tens of millions of daily active users and yielding substantial business improvements. The code and data are available at https://github.com/zyody/tarq_sigir2026.

P100 · **MuseKG: An Interactive Knowledge Graph Over Museum Collections** [Demo]

Jinhao Li (The University of Melbourne); Jianzhong Qi (The University of Melbourne); Soyeon Caren Han (The University of Melbourne); Eun-Jung Holden (The University of Melbourne)

Digitisation in the cultural heritage sector has produced large but fragmented repositories of museum collection data, spanning structured catalogue records, images, and unstructured descriptions. Existing museum information systems often make it difficult to integrate these sources into a unified, queryable representation that supports relation-aware exploration. We present MuseKG, an interactive knowledge graph system that organises heterogeneous museum data into a typed graph that links objects, people, organisations, images, image-derived labels, and extracted semantic entities within a coherent schema. MuseKG supports natural-language queries by grounding user questions to graph entities and retrieving a compact neighbourhood of evidence for answer generation. Through an interactive demonstration on real museum collections, we show that MuseKG supports common exploration tasks such as attribute lookup, relation exploration, and relation-aware retrieval, with answers that remain inspectable via explicit graph structures.

P101 · **CoCo: Conformal Confidence Suppression to Optimize Search Results** [Short]

Zhou Qin (Amazon); Shuang Wu (Amazon); Yoon Kim (Amazon); Yi Liu (Amazon); Wenyang Liu (Amazon)

Modern e-commerce recommendation systems aim to improve customer experience by ranking content on search results page (SRP). However, displaying content is not always beneficial for customers across all contexts; even top-ranked content can be irrelevant, misleading, or redundant in certain scenarios. In this work, we propose a robust content suppression mechanism to selectively suppress content when necessary. Our approach leverages causal effect learning to measure the incremental value of showing versus suppressing content. Prior work uses Conditional Average Treatment Effect (CATE) to estimate treatment effect without considering the inherent uncertainty in uplift predictions, resulting in over-suppression and degraded customer experience. Our work introduces a novel Conformal Confidence (CoCo) suppressor that explicitly accounts for prediction uncertainty in uplift-based modeling. We first evaluate it on synthetic datasets with controlled noise, bias, and contexts. Results demonstrate superior performance compared to baseline approaches. Subsequently, our online traffic tests show statistically significant improvements in revenue and profit compared to existing methods.

P102 · **WebMall - A Multi-Shop Benchmark for Evaluating Web Agents** [Resource]

Ralph Peeters (University of Mannheim); Aaron Steiner (University of Mannheim); Luca Schwarz (University of Mannheim); Julian Yuya Caspary (University of Mannheim); Christian Bizer (University of Mannheim)

LLM-based web agents have the potential to automate long-running web tasks, such as searching for products in multiple e-shops and subsequently ordering the cheapest products that meet the user's needs. Benchmarks for evaluating web agents either require agents to perform tasks online using the live Web or offline using simulated environments, the latter allowing for the exact reproduction of the experimental setup. While DeepShop and ShoppingComp provide online benchmarks that require agents to perform challenging shopping tasks, existing offline benchmarks such as WebShop, WebArena, and Mind2Web cover only comparatively simple e-commerce tasks performed against a single shop containing product data from a single source. What is missing is an e-commerce benchmark that simulates multiple shops containing heterogeneous product data and requires agents to perform complex retrieval tasks. We fill this gap by introducing WebMall, the first offline multi-shop benchmark for evaluating web agents on challenging comparison shopping tasks. WebMall consists of four simulated shops populated with product data extracted from the Common Crawl. The WebMall tasks range from specific product searches and price comparisons to advanced searches for complementary or substitute products, as well as checkout processes. We validate WebMall using eight agents that differ in observation space, availability of short-term memory, and the employed LLM. The validation highlights the difficulty of the benchmark, with the best-performing agents achieving task completion rates below 65% in the

task categories cheapest product search and vague product search.

P103 · **ZoRRO: A Zero-Weight Personalized Recommender System for Scalable News Recommendation** [Short]

Johannes Kruse (JP/Politikens Media Group); Ryotaro Shimizu (ZOZO Research); Kasper Lindskow (JP/Politikens Media Group); Jon Tofteskov (JP/Politikens Media Group); Michael Riis Andersen (Technical University of Denmark); Julian McAuley (University of California San Diego); Jes Frellsen (Technical University of Denmark)

In this paper, we present ZoRRO (Zero-Weight Personalized Recommender System), a zero-weight and training-free framework for personalized news recommendation designed for scalable real-world deployment. We show that ZoRRO outperforms strong neural baselines in offline ranking evaluations and delivers click-through rate performance in online A/B testing that is nearly on par with a state-of-the-art deep learning model, while operating more than \(600\times\) faster. Our experiments reveal gaps between offline and online performance, and show that models with similar click-through rate (CTR) outcomes can produce markedly different recommendation distributions, influencing the overall news flow. These findings position ZoRRO as a practical and efficient solution for large-scale news recommendation and highlight the importance of evaluating recommender systems using metrics beyond accuracy alone. Our code is available at \url{https://github.com/johanneskruse/zorro}.

P104 · **Reinforcing User Interest Evolution in Multi-Scenario Learning for recommender systems** [Short]

Zhijian Feng (ByteDance); Wenhao Zheng (Microsoft Corp); Xuanji Xiao (Independent Researcher)

In real-world recommendation systems, users engage in a variety of scenarios, such as homepages, search pages, and related-item recommendation pages. Each of these scenarios is designed to capture distinct facets of user intent. However, user interests are often inconsistent across different scenarios, attributable to variations in their decision-making processes and modes of preference expression. This inherent heterogeneity poses a substantial challenge to unified modeling, rendering multi-scenario recommendation a non-trivial task. To address these challenges, we propose RUIE, a novel reinforcement learning-enhanced framework that models dynamic user preference evolution as sequential decision-making problems, and leverages cross-scenario behavior sequences to detect interest shifts and dynamically adjust sample utilization for accurate interest capture. Experiments demonstrate that our proposed approach significantly outperforms state-of-the-art methods in multi-scenario recommendation tasks and is widely applicable. This work offers a fresh perspective on multi-scenario modeling and highlights promising directions for future research. The source code is available at https://github.com/rl4rec/RUIE.

P105 · **RRE-GTC: A Geo-Temporal Cluster Dataset for Real-Estate Price Estimation and Market-Aware Search** [Resource]

Irina Govorova (CIAN); Aleksandr Alekseitsev (CIAN); Irina Podlipnova (Innopolis University); Meruza Kubentayeva (MIRAI); Yuriy Dorn (Lomonosov Moscow State University)

Real-estate platforms represent a complex Information Retrieval (IR) environment where user relevance depends on high-dimensional intersections of location, time, and physical attributes. However, research into Geo-spatial Information Retrieval (GIR) and location-based recommendations is currently impeded by a lack of open, high-quality datasets that capture both the temporal evolution of markets and granular accessibility signals. To bridge this gap, we introduce RRE-GTC (Ru-Real-Estate Geo-Temporal Clusters), an open resource designed to benchmark ranking, pricing, and recommendation algorithms in dynamic spatial contexts. Derived from a massive collection of apartment listings in major Russian cities, RRE-GTC aggregates 456,182 geo-temporal clusters, offering a privacy-preserving alternative to releasing raw listing-level data. The dataset provides rich and distinct features: (i) geo-temporal price dynamics (stratified percentiles over time), (ii) detailed item metadata (building stock, supply volume), and (iii) transport accessibility signals (e.g., proximity, transport density). Uniquely, the resource includes a "transport-isolated" price signal derived from a linear attribution model, enabling novel research into attribute-based ranking and explainable search (e.g., separating "location value" from "intrinsic quality"). Released under the Apache-2.0 license and hosted on Hugging Face, RRE-GTC lowers the barrier for experimenting with content-based recommendation and retrieval, spatiotemporal ranking, and market-aware search filtration.

P106 · **Additive Control Variates Dominate Self-Normalisation in Off-Policy Evaluation** [Short]

Olivier Jeunen (aampe); Shashank Gupta (Microsoft AI)

Off-policy evaluation (OPE) is essential for assessing ranking and recommendation systems without costly online interventions.Self-Normalised Inverse Propensity Scoring (SNIPS) is a standard tool for variance reduction in OPE, leveraging a multiplicative control variate.Recent advances in off-policy learning suggest that additive control variates (baseline corrections) may offer superior performance, yet theoretical guarantees for evaluation are lacking.This paper provides a definitive answer: we prove that $\beta^\star$-IPS, an estimator with an optimal additive baseline, asymptotically dominates SNIPS in Mean Squared Error.By analytically decomposing the variance gap, we show that SNIPS is asymptotically equivalent to using a specific---but generally sub-optimal---additive baseline.Our results theoretically justify shifting from self-normalisation to optimal baseline corrections for both ranking and recommendation.

P107 · **KCC: Korean Civil Case Dataset for Legal Information Retrieval** [Resource]

Minhan Cho (Sungkyunkwan University); Soyoung Park (Sungkyunkwan University); S. Shyam Sundar (The Pennsylvania State University); Daejin Choi (Ewha Womans University); Jinyoung Han (Sungkyunkwan University)

Reliable relevance modeling and ranking are pivotal to legal information retrieval (IR), where graded relevance judgments are indispensable yet prohibitively expensive to acquire. Despite their importance, expert-validated graded relevance benchmarks remain scarce, especially for non-English statutory-law jurisdictions where citation practices differ fundamentally from common-law paradigms. We introduce KCC, a publicly available and reusable benchmark resource for legal IR in Korean civil law. KCC is constructed from 38,372 publicly accessible court decisions and comprises 2,942 query–candidate case pairs annotated with expert-validated four-level relevance labels. These labels jointly capture factual similarity and legal reasoning, explicitly addressing the absence of systematic citation links in statutory-law systems. To facilitate reproducibility, we provide a detailed description of the dataset construction pipeline, annotation protocol, relevance criteria, and inter-annotator agreement statistics validated by legal professionals. The annotation guidelines are designed to be adaptable to other statutory-law and non-English legal IR settings, enabling future extensions beyond the current dataset. We further present a set of reference benchmark experiments using representative retrieval approaches, including traditional IR models, neural rankers, and LLM-based prompting methods, to indicate recommended usage and evaluation practices with KCC. We believe that KCC provides a long-term resource for the research community to study relevance modeling, ranking behavior, and evaluation methodology in domain-specific legal IR scenarios beyond English and common-law traditions.

P108 · **StAR: Adaptive Structure-Aware Reranking for Semantic–Structural Alignment in GraphRAG** [Short]

Junghyun Oh (Chungnam National University); Sungsu Lim (Chungnam National University)

Retrieval-Augmented Generation (RAG) often struggles with multi-hop reasoning because Euclidean embeddings collapse hierarchical structure and let semantic similarity dominate ranking, leading to semantic–structural misalignment. As a result, structurally irrelevant yet semantically similar candidates can outrank the evidence needed for multi-hop QA. We introduce StAR (Structure-Aware Reranking), a plug-and-play reranking module that injects hyperbolic structural similarity into GraphRAG ranking. StAR constructs tree-like representations of the query and candidate subgraphs, computes structure-aware similarity in hyperbolic space, and adaptively modulates structural contributions using a query-level alignment signal based on Spearman's ?. Experiments on four QA benchmarks show consistent gains, with the largest improvements on datasets with stronger hierarchical reasoning demands, while remaining robust on flatter retrieval settings. These results suggest that restoring structural faithfulness through hyperbolic structural scoring improves ranking consistency for multi-hop QA.

P109 · **CalmSet: A Domain-Specific Test Collection for Affective Music Retrieval for Children with ASD** [Resource]

Abhishek Karwankar (University of Delaware); Liam Stapley (University of Delaware); Daniel Stevens (University of Delaware); Matthew Louis Mauriello (University of Delaware)

Information Retrieval (IR) increasingly relies on subjective, graded, and natural-language notions of relevance, motivating the development of reproducible test collections for ranking and recommendation. In affect-sensitive music domains, however, such resources with human-validated relevance signals remain scarce. We introduce CalmSet, a test collection for emotion-tagged music retrieval and recommendation in a therapeutic context for children with Autism Spectrum Disorder (ASD). CalmSet contains 432 modular music tracks instantiated from four purposefully composed base songs with controlled provenance, each formed by distinct combinations of seven active musical layers. Each track is annotated with ranked top-3 therapeutic intent labels and natural-language descriptions. Annotations are produced via a hybrid human-in-the-loop pipeline: CLAP proposes candidate intent labels, a large language model generates auxiliary semantic descriptions, and crowd workers provide ranked judgments without exposure to model outputs; final labels are aggregated using a Borda-based procedure. As initial baselines, we evaluate five one-vs-rest multi-label classifiers over CLAP audio embeddings, observing moderate micro-F1 scores (up to 0.60) but low exact-match accuracy (<0.10), while top-3 label overlap is substantially higher (Jaccard@3 up to 0.48), motivating graded-relevance evaluation. CalmSet supports both sparse (e.g., BM25) and dense audio–text retrieval models using therapeutic labels or natural-language descriptions as queries.

P110 · **BioCLEAR Benchmark for Biomedical Text Simplification** [Resource]

Jan Bakker (University of Amsterdam); Liana Ermakova (University of Brest); Jaap Kamps (University of Amsterdam)

Access to objective, reliable scientific information is crucial in a world of misinformation and disinformation, yet the general public often avoids scientific literature due to its perceived complexity. Modern generative information access models hold the promise of removing some of these barriers by developing appropriate approaches to scientific text simplification. We constructed BioCLEAR, a benchmark for Biomedical Comprehensible Lay Explanations of Academic Research used in the CLEF 2025 SimpleText track, comprising a large set of aligned abstracts and plain-language summaries from Cochrane systematic reviews. First, it provides a new

resource for scientific document-level text simplification, including discourse-level aspects that are absent from traditional sentence-level data. Second, we include existing resources created through human sentence-level simplification to run experiments across different datasets. Third, it includes a large set of sentence-level and document-level submissions to the CLEF track, providing a sample of current models and enabling detailed analysis of their effectiveness and remaining issues.

P111 · **Delay-Aware Sequential Recommendation with Dynamic Graphs and Variate-Temporal Decomposition** [Short]

Yoonhyuk Choi (Sookmyung Women's University)

Sequential recommendation systems may operate on partially observed histories when interaction logs become available only after non-trivial ingestion latency. Most existing methods assume immediately observed feedback, which can degrade when the model-visible interaction stream is temporally misaligned with the underlying action sequence. We propose DVGD, a delay-aware sequential recommendation framework that learns lag-conditioned message passing over censored histories induced by delayed log availability. Our method represents observed user-item interactions as a dynamic graph, decouples feature-wise and temporal dependencies via variate-temporal decomposition, and learns a time-lag adjacency tensor that routes information across multiple temporal lags. Experiments on benchmark datasets with controlled simulated delays show that the proposed method consistently outperforms strong baselines, with larger gains under longer and more heterogeneous delays. These results support DGVD as an effective approach for delayed-observation robustness in sequential recommendation, while further validation on real delayed logs remains future work.

P112 · **Follow the TRACE: Exploiting Post-Click Trajectories for Online Delayed Conversion Rate Prediction** [Short]

Xinyue Zhang (Institute of Computing Technology, Chinese Academy of Sciences); Yuanhao Ding (Institute of Computing Technology, Chinese Academy of Sciences); Xiang Ao (Institute of Computing Technology, Chinese Academy of Sciences)

Delayed feedback poses a core challenge for online CVR prediction, forcing a trade-off between label accuracy and data freshness. Existing methods address this through delay modeling or sample reweighting, yet neglect how post-click behaviors evolve over the observation period. To overcome this limitation, we formalize this evolution as feedback trajectory and propose TRACE. Instead of forcing hard labels on unrevealed samples, our method evaluates how well the accumulated feedback status aligns with conversion versus non-conversion, dynamically refining posteriors without waiting for final outcomes. To counteract early-stage trajectory sparsity, we further design a reliability-gated retrospective completer that leverages full-lifecycle data to provide adaptive posterior guidance for unrevealed samples. Extensive experiments validate TRACE's superiority over state-of-the-art baselines and confirm the retrospective completion module as a model-agnostic enhancer for existing systems. Our code is available at \hrefhttps://github.com/LunaZhangxy/TRACE https://github.com/LunaZhangxy/TRACE.

P113 · **HE-DeepFM: An FHE Inference System for CTR Prediction with Efficient FM Interactions** [Short]

Qiyue Su (University of Science and Technology of China); Hang Gu (University of Science and Technology of China); Zhiguang Wang (National Key Laboratory of Modeling and Simulation for Complex Systems); Teng Wang (University of Science and Technology of China); Zhendong Zheng (University of Science and Technology of China); Qianyu Cheng (University of Science and Technology of China); Lei Gong (University of Science and Technology of China); Chao Wang (University of Science and Technology of China)

Scoring models such as click-through rate (CTR) prediction underpin recommendation and advertising systems, but their features are highly sensitive, making plaintext cloud inference risky. Fully homomorphic encryption (FHE) enables inference directly on ciphertexts, yet homomorphic computation is expensive and bootstrapping often dominates end-to-end latency. We present HE-DeepFM, a FHE inference system for CTR prediction. We first design HE-FM, a homomorphic-friendly Factorization Machine (FM) operator that exploits CKKS SIMD packing to compute second-order interactions efficiently, thereby avoiding the naive O(F2) cost of pairwise feature interactions. Building on HE-FM, HE-DeepFM reduces bootstrapping under the same FHE budget, and can eliminate it for small model configurations. We implement HE-DeepFM with Orion and Lattigo and evaluate it on real-world datasets. On Criteo, HE-DeepFM reduces bootstrapping from 12 to 4 and achieves up to 2.85× end-to-end speedup while maintaining prediction quality comparable to the baseline.

P114 · **When Context Bites: Detecting RAG Poisoning via Document-Level Attention Collapse** [Short]

Yingtao Ren (University of Technology Sydney); Ziyi Zhao (University of Technology Sydney); Yiwei Fu (Peking University); Xiao Luo (University of Wisconsin - Madison); Yu-Cheng Chang (University of Technology Sydney); Chin-Teng Lin (University of Technology Sydney)

Retrieval-augmented generation (RAG) is indispensable for enhancing large language models. However, RAGs are increasingly susceptible to poisoning attacks, in which adversarial documents are injected to manipulate generator outputs. Previous methods rely on output-side signals such as perplexity and consistency checks to detect such attacks. Nevertheless, our analysis reveals that deliberate attacks often induce false confidence, where poisoned outputs exhibit even lower perplexity than benign ones, rendering uncertainty-based detection ineffective. To address this challenge, we explore the internal dynamics of the generator and identify a distinctive signature termed Attention Collapse. Unlike the dispersed attention in benign generations, attacked generations exhibit a decrease in entropy as attention concentrates on poisoned documents. Building on these findings, we propose D-SCAN (Document-level Signal Collapse Analysis), a lightweight detection framework that monitors attention dynamics to identify attacked generations. Extensive experiments on multiple attack benchmarks demonstrate the effectiveness of our method. Moreover, D-SCAN can detect attacks even when they fail to alter the final answer. Code is available at https://github.com/yingtaoren/D-Scan.git.

P115 · **Separating Wheat from Chaff: Fine-Grained Defenses Against Poisoning in Multi-Perspective RAG** [Short]

Yanxiao Zhao (Beijing University of Posts and Telecommunications); Longzhu He (Beijing University of Posts and Telecommunications); Li Sun (Beijing University of Post and Telecommunications); Sen Su (Beijing University of Posts and Telecommunications)

Retrieval-Augmented Generation (RAG) systems are vulnerable to perspective manipulation attacks in multi-perspective long-form generation scenarios: adversaries inject documents with implicit stance biases to induce the system to generate content favoring specific perspectives. While existing RAG defense methods perform well in closed-domain factoid QA, they face critical challenges in multi-perspective scenarios: excessive contradiction filtering leads to perspective homogenization, and insufficient capability to identify implicit stance biases. To address these limitations, we propose Perspective-Shielded RAG (PS-RAG), a defense framework based on hypergraph modeling. The core innovation lies in preserving perspective diversity through cluster-level conflict modeling, while designing a fine-grained contradiction classification mechanism to precisely distinguish malicious attacks from legitimate perspective disagreements. Experiments on four benchmark datasets demonstrate that PS-RAG achieves state-of-the-art defense performance, significantly outperforms existing methods in preserving perspective diversity in long-form generation scenarios.

P116 · **BANANA: Bounded Adaptive Navigation Architecture for Nested Archives** [Short]

Anup Roy (Inception AI); Rishabh Gyanendra Upadhyay (Inception AI); Animesh Rameshbhai Panara (Inception AI); Aidan Philip Millar (Mubadala); Larry Murray (Inception AI); Robin Mills (Inception AI)

Retrieval-augmented question answering (QA) over enterprise PDF archives frequently produces confident near-miss answers because dense embeddings discard entity identity, numerical units, and document provenance. Plain-text indices preserve the entity, unit, and provenance signals that dense embeddings discard. Banana (Bounded Adaptive Navigation Architecture for Nested Archives) replaces opaque vectors with plain-text Markdown indices built via one-time VLM page transcription, eliminating both a separate Optical Character Recognition (OCR) engine and a vector store. At query time it applies Progressive Retrieval with Overlap (PRO), an iterative selector with a deterministic contraction bound, followed by a K-adaptive extraction gate that escalates retrieval depth only upon evidence insufficiency. Across three benchmarks, Banana consistently outperforms the strongest baseline: 71.4~EM on TAT-DQA (+15.2 over PageIndex; p 0.001), within 5.2 of oracle TAT-LLM; 79.3% accuracy on FinanceBench-150 across 53K~pages (+17.3 over PageIndex; p 0.01); and 86.2% Completeness@10 on UniDoc-Bench across 8~domains (+20.8 over the strongest fusion baseline). All gains hold across multiple backbones, including a local model at zero API cost.

P117 · **An Asymmetric-difference based Document Exact Similarity Search Algorithm** [Short]

Lianyin Jia (Kunming University of Science and Technology); Yongxue Zhao (Kunming University of Science and Technology); Mengjuan Li (Yunnan Normal University); Xiuxing Li (Beijing Institute of Technology); Ying Jiang (Kunming University of Science and Technology); Jiaman Ding (Kunming University of Science and Technology)

Retrieving similar documents serves as a fundamental building block for numerous information retrieval applications. Existing symmetric-difference based exact similarity search algorithm generates excessively large key bounds, resulting in reduced efficiency. To address this issue, we introduce a novel asymmetric-difference, which decomposes the symmetric-difference into an upper-bound difference and a lower-bound difference, thereby achieving tighter key bounds. Based on the asymmetric-difference, we design a novel key inverted index, KIL, and the corresponding exact similarity search algorithm, AD-Search. KIL maps a document and its length to an integer key, enabling direct pruning of documents that fail to meet length constraints. AD-Search employs an asymmetric-difference based strategy for both key bound generation and key filtering. This approach efficiently prunes non-qualifying keys, and thus greatly reduces the number of candidates. Experimental results on real-world datasets demonstrate that AD-Search outperforms MultiTree by up to 58.71× and LeBQ+ by up to 3.93×.

P118 · **Partial Label Learning-Inspired Denoising Implicit Feedback for Recommendation** [Short]

Huilin Chen (University of Technology Sydney); Jie Lu (University of Technology Sydney); Kezhi Lu (University of Technology Sydney); Zhen Fang (University of Technology Sydney); Guangquan Zhang (University of Technology Sydney)

Implicit feedback, such as clicks and browsing behaviors, is ubiquitous in recommender systems as a proxy for user preferences. However, these signals are inherently noisy; interactions such as misclicks, unintended views, or unsatisfactory purchases often introduce false positive patterns that

mislead model learning. Existing denoising strategies mainly rely on heuristic "small-loss" principles to suppress the influence of high-loss samples. However, this approach creates a fundamental trade-off: by indiscriminately penalizing large-value losses, these methods inadvertently weaken the model's ability to learn "hard" true positive interactions, thereby compromising robustness and personalization. By conceptualizing this ambiguity as a "candidate label set" that encompasses both true and noisy feedback, we are the first to reformulate the recommendation denoising problem into a Partial Label Learning (PLL) task. This novel perspective allows us to address the fundamental challenge of unreliable pseudo-labels by transforming traditional heuristic-based filtering into a principled label disambiguation process. Specifically, we propose PLLD, a Partial Label Learning-inspired Denoising method. Unlike existing methods that rely on indirect signal filtering, PLLD innovatively leverages PLL paradigms to directly resolve ambiguous implicit feedback, effectively recovering clean signals from noisy candidate sets. Experiments on multiple real-world benchmark datasets demonstrate that PLLD consistently improves ranking performance and robustness under substantial noise.

Poster Session 2 · Wednesday 14:30–16:00 · Exhibition Hall 21B/22B · 117 boards

P001 · **How Do Large Language Models Understand Relevance? A Mechanistic Interpretability Perspective** [TOIS]

Qi Liu (Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China); Haozhe Duan (Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China); Jiaxin Mao (Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China); Ji-Rong Wen (Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China)

Recent studies have shown that large language models (LLMs) can assess relevance and support information retrieval (IR) tasks such as document ranking and relevance judgment generation. However, the internal mechanisms by which off-the-shelf LLMs understand and operationalize relevance remain largely unexplored. In this article, we systematically investigate how different LLM modules contribute to relevance judgment through the lens of mechanistic interpretability. Using activation patching techniques, we analyze the roles of various model components and identify a multi-stage, progressive process in generating either pointwise or pairwise relevance judgment. Specifically, LLMs first extract query and document information in the early layers, then process relevance information according to instructions in the middle layers, and finally utilize specific attention heads in the later layers to generate relevance judgments in the required format. Our findings provide insights into the mechanisms underlying relevance assessment in LLMs, offering valuable implications for future research on leveraging LLMs for IR tasks.

P002 · **Individual Turing Test: A Case Study of LLM-based Simulation Using Longitudinal Personal Data** [Short]

Minghao Guo (Rutgers University); Ziyi Ye (Fudan University); Wujiang Xu (Rutgers University); Xi Zhu (Rutgers University); Wenyue Hua (Microsoft); Dimitris N. Metaxas (Rutgers University)

Large Language Models (LLMs) have demonstrated remarkable human-like capabilities, yet their ability to replicate a specific individual remains underexplored. This paper presents a case study investigating LLM-based individual simulation using a volunteer-contributed archive of private messaging history spanning over ten years. Based on this dataset, we propose the "Individual Turing Test" to evaluate whether acquaintances of the volunteer can correctly identify which response in a multi-candidate pool most plausibly originates from the volunteer. We investigate prevalent approaches to LLM-based individual simulation, including fine-tuning, retrieval-augmented generation (RAG), memory-based methods, and hybrid approaches that integrate fine-tuning with RAG or memory. Empirical results show that current methods do not pass the Individual Turing Test, but perform substantially better when the same test is conducted on strangers to the target individual. Additionally, while fine-tuning improves performance in daily chats that reflect the individual's language style, retrieval-augmented and memory-based approaches demonstrate stronger performance on questions involving personal opinions and preferences. These findings reveal a fundamental trade-off between parametric and non-parametric approaches to individual simulation with LLMs under longitudinal context.

P003 · **TEC: A Collection of Human Trial-and-error Trajectories for Problem Solving** [Resource]

Xinkai Zhang (College of AI, Tsinghua University, Quancheng Laboratory, Gaoling School of Artificial Intelligence, Renmin University of China); Jingtao Zhan (Institute of Data and Information, Tsinghua Shenzhen International Graduate School); Yiqun Liu (Department of Computer Science and Technology, Tsinghua University); Qingyao Ai (Quancheng Laboratory, Department of Computer Science and Technology, Tsinghua University)

Trial-and-error is a fundamental strategy for humans to solve complex problems and a necessary capability for Artificial Intelligence (AI) systems operating in real-world environments. Although several trial-and-error AI techniques have recently been proposed, most of them rely on simple heuristics designed by researchers and achieve limited performance gains. The core issue is the absence of appropriate data: current models cannot learn from detailed records of how humans actually conduct trial-and-error in practice. To address this gap, we introduce a data annotation platform and a corresponding dataset, termed Trial-and-Error Collection (TEC). The platform records users' complete trajectories across multiple trials and collects their reflections after receiving error feedback. Using this platform, we record the problem-solving processes of 46 participants on 58 tasks, resulting in 5,370 trial trajectories along with error reflections across 41,229 webpages. With this dataset, we observe that humans achieve substantially higher accuracy compared to LLMs, which demonstrates that humans are more effective in trial-and-error than LLMs. We believe that the TEC platform and dataset provide a valuable foundation for understanding human trial-and-error behavior and for developing more capable AI systems. Platform and dataset are publicly available.. https://github.com/Serendipity0429/TEC.

P004 · **Enhancing Judgment Document Generation via Agentic Legal Information Collection and Rubric-Guided Optimization** [Short]

Weihang Su (Tsinghua University); Xuanyi Chen (Tsinghua University); Yueyue Wu (Quan Cheng Laboratory); Qingyao Ai (Quan Cheng Laboratory); Yiqun Liu (Tsinghua University)

Automating the drafting of judgment documents is pivotal to judicial efficiency, yet it remains challenging due to the dual requirements of comprehensive retrieval of legal information and rigorous logical reasoning. Existing approaches, typically relying on standard Retrieval-Augmented Generation and Supervised Fine-Tuning, often suffer from insufficient evidence recall, hallucinated statutory references, and logically flawed legal reasoning. To bridge this gap, we propose Judge-R1, a unified framework designed to enhance LLM-based judgment document generation by jointly improving legal information collection and judgment document generation. First, we introduce Agentic Legal Information Collection, which employs a dynamic planning agent to retrieve precise statutes and precedents from multiple sources. Second, we implement Rubric-Guided Optimization, a reinforcement learning phase utilizing Group Relative Policy Optimization (GRPO) with a comprehensive legal reward function to enforce adherence to judicial standards and reasoning logic. Extensive experiments on the JuDGE benchmark demonstrate that Judge-R1 significantly outperforms state-of-the-art baselines in both legal accuracy and generation quality.

P005 · **Personalized Deep Research: A User?Centric Framework, Dataset, and Hybrid Evaluation for Knowledge Discovery** [Resource]

Xiaopeng Li (City University of Hong Kong); Wenlin Zhang (City University of Hong Kong); Yingyi Zhang (City University of Hong Kong); Pengyue Jia (City University of Hong Kong); Yejing Wang (City University of Hong Kong); Yichao Wang (Huawei Technologies Co Ltd); Yong Liu (Huawei Technologies Co Ltd); Huifeng Guo (Huawei Technologies Co Ltd); Xiangyu Zhao (City University of Hong Kong)

Deep Research agents driven by LLMs have automated the scholarly discovery pipeline, from planning and query formulation to iterative web exploration. Yet they remain constrained by a static, "one-size-fits-all" retrieval paradigm. Current systems fail to adaptively adjust the depth and breadth of exploration based on the user's existing expertise or latent interests, frequently resulting in reports that are either redundant for experts or overly dense for novices. To address this, we introduce Personalized Deep Research (PDR), a framework that integrates dynamic user context into the core retrieval-reasoning loop. Rather than treating personalization as a post-hoc formatting step, PDR unifies user profile modeling with iterative query development, dual-stage (private/public) retrieval, and context-aware synthesis. This allows the system to autonomously align research sub-goals with user intent and optimize the stopping criteria for evidence collection. To facilitate benchmarking, we release the PDR Dataset, covering four realistic user tasks, and propose a hybrid evaluation framework combining lexical metrics with LLM-based judgments to assess factual accuracy and personalization alignment. Experimental results against commercial baselines demonstrate that PDR significantly improves retrieval utility and report relevance, effectively bridging the gap between generic information retrieval and personalized knowledge acquisition. The resource is available to the public at~ https://github.com/Applied-Machine-Learning-Lab/SIGIR2026_PDR.

P006 · **TREC iKAT 2025: A Test Collection for the Offline and Interactive Evaluation of Conversational Search** [Resource]

Zahra Abbasiantaeb (University of Amsterdam); Simon Lupart (University of Amsterdam); Marcel Gohsen (Bauhaus-Universität Weimar); Nailia Mirzakhmedova (Bauhaus-Universität Weimar); Johannes Kiesel (GESIS - Leibniz Institute for the Social Sciences); Jeff Dalton (University of Edinburgh); Mohammad Aliannejadi (University of Amsterdam)

Conversational search agents, especially with the advent of large language models, have developed into useful tools to satisfy complex information needs of their users. Former research has shown that personalization (i.e., adaptation of agent responses to the preferences and traits of the user) can increase the relevance and perceived answer quality of these systems even further. However, developing accurate personalization methods typically requires rich datasets, both in terms of user profiles and complex conversations, for which only a few resources are publicly available. Over the past three years, the goal of the TREC Interactive Knowledge Assistance Track (iKAT) has been to bridge this gap. In organizing this shared task, we have developed a collection of complex information needs and associated conversations, made to challenge today's conversational agents and thus highlight aspects in need of further research. In this paper, we present the resources made for iKAT 2025, focusing on multi-session conversations (i.e., multiple dialogues per user), dynamically evolving user models, mixed-initiative dialogues, and large-scale human and automatic assessments. In addition to manually designed user profiles and conversations, the test collection for 2025 also contains dialogues between

participating systems and our user simulators. All the resources are publicly available in our repository, including the system evaluation both as a result of the offline (i.e., test collection-based) and interactive tasks (i.e., user simulation-based), as well as their source code and model weights, to foster future research in this direction.

P007 · **Exploring and Improving Cross- and Multi-Domain Personalized Question Answering via a Dual Retrieval Augmentation Approach** [Short]

Ozel Yilmazel (University of Massachusetts Amherst); Hamed Zamani (University of Massachusetts Amherst); James Allan (University of Massachusetts Amherst)

Personalized long-form question answering aims to tailor responses based on a user's historical data, which often spans diverse information domains. While prior work typically assumes user profiles are domain-aligned with the users' questions, real-world profiles are often diverse, containing information across multiple domains. In this work, we use the LaMP-QA benchmark, where user profiles span diverse Stack Exchange domains (Arts & Entertainment, Lifestyle & Personal Development, and Society & Culture), to systematically investigate how domain composition of the user profile affects personalization performance. We examine three profile composition settings: Cross-Domain, where profiles contain only out-of-domain information; Multi-Domain, where profiles span multiple domains including the question's domain; and In-Domain, where profiles match the question domain. Our analysis reveals that state-of-the-art personalization methods struggle in Cross-Domain settings, often failing to outperform non-personalized baselines. To address this challenge, we introduce Dual Retrieval Augmentation, a framework that jointly retrieves from both the user profile and a public corpus. We apply our framework to existing personalization approaches and demonstrate that it consistently outperforms state-of-the-art methods across all three profile settings, validating its robustness for personalized question answering.

P008 · **Total Recall QA: A Verifiable Evaluation Suite for Deep Research Agents** [Resource]

Mahta Rafiee (University of Massachusetts Amherst); Heydar Soudani (Radboud University); Zahra Abbasiantaeb (University of Amsterdam); Mohammad Aliannejadi (University of Amsterdam); Faegheh Hasibi (Radboud University); Hamed Zamani (University of Massachusetts Amherst)

Deep research agents have emerged as LLM-based systems designed to perform multi-step information seeking and reasoning over large, open-domain sources to answer complex questions by synthesizing information from multiple information sources. Given the complexity of the task and despite various recent efforts, evaluation of deep research agents remains fundamentally challenging. This paper identifies a list of requirements and optional properties for evaluating deep research agents. We observe that existing benchmarks do not satisfy all identified requirements. Inspired by prior research on TREC Total Recall Tracks, we introduce the task of Total Recall Question Answering and develop a framework for deep research agents evaluation that satisfies the identified criteria. Our framework constructs single-answer, total recall queries with precise evaluation and relevance judgments derived from a structured knowledge base paired with a text corpus, enabling large-scale data construction. Using this framework, we build TRQA, a deep research benchmark constructed from Wikidata-Wikipedia as a real-world source and a synthetically generated e-commerce knowledge base and corpus to mitigate the effects of data contamination. We benchmark the collection with representative retriever and deep research models and establish baseline retrieval and end-to-end results for future comparative evaluation.

P009 · **How Fine-Grained Should a RAG Benchmark Be? A Hierarchical Framework for Synthetic Question Generation** [Short]

Chase M. Fensore (Emory University); Kaustubh Dhole (Emory University); Jason Fan (Emory University); Eugene Agichtein (Emory University); Joyce C. Ho (Emory University)

Evaluating retrieval-augmented generation (RAG) systems requires benchmarks that capture diverse question characteristics, yet practitioners lack empirical guidance on which dimensions to vary and at what granularity. We present HieraRAG, a hierarchical framework for studying granularity in RAG benchmark construction, defining optimal granularity as the level that maximizes discriminative power (the standard deviation of generation quality across categories) within a given RAG configuration. As a case study, we generate 5,872 synthetic question--answer (QA) pairs from FineWeb-10BT across 3 dimensions (Question Complexity, Answer Type, Linguistic Variation) at 3 granularity levels (2, 4, and 8 categories). With a BM25+Falcon-3-10B pipeline, optimal granularity varies by dimension: complexity benefits from fine-grained distinctions (discriminative power: 0.053) while answer type and linguistic variation peak at medium granularity. We introduce a Coherence Ratio metric to quantify whether fine-grained splits cleanly subdivide parent categories, revealing structural differences across dimensions (Question Complexity: 0.40 vs. Answer Type: 1.44). Human evaluation of 110 stratified QA pairs confirms synthetic quality. While these specific findings reflect a single configuration, HieraRAG provides a portable procedure and validation metric for practitioners to determine evaluation granularity within their own RAG settings.

P010 · **The LLM Effect on IR Benchmarks: A Meta-Analysis of Effectiveness, Baselines, and Contamination** [Short]

Moritz Staudinger (TU Wien); Wojciech Kusa (NASK - National Research Institute); Allan Hanbury (TU Wien)

Benchmark collections have long enabled controlled comparison and cumulative progress in Information Retrieval (IR). However, prior meta-analyses show that reported effectiveness gains often fail to accumulate, in part due to weak or outdated baselines. Large language models (LLMs) are increasingly used in retrieval pipelines, yet their impact on established IR benchmarks has not been systematically analyzed. We analyze 179 publications reporting on the TREC Robust04 collection and the TREC Deep Learning 2020 (DL20) Passage Retrieval benchmark, using ACM Digital Library keyword search supplemented by citation-graph backtracking for Robust04. We observe what we term an LLM effect: recent systems incorporating LLM components achieve 8.8% higher nDCG@10 on DL20 than the best TREC 2020 result and 11.9% higher on Robust04 than the strongest pre-2024 result. However, evaluation practice has shifted from MAP to nDCG@10 over the same window, and our adaptation of the Data Contamination Quiz reveals 12-41% contamination across two widely-used LLM rerankers. Filtering contaminated topics shows no statistically significant effectiveness difference, but small samples and the uncertainty of adapting contamination detection to reranking prevent us from ruling out memorization as a contributing factor. We read the LLM effect as real but unverified: visible in the aggregate numbers, but not cleanly separable from metric drift or pretraining overlap.

P011 · **Humans, LLMs, and Measures Do Not Align in Attributed Information Retrieval** [Short]

Lukas Gienapp (University of Kassel); Jenny Lang (University of Tübingen); Martin Potthast (University of Kassel); Harrisen Scells (University of Tübingen)

Evaluating attributed information retrieval (AIR) systems requires assessing both informativeness and attributability. To enable scalable evaluation, LLM-sourced ground truth data is frequently used, yet the validity of this practice remains unclear. We replicate the evaluation framework of Djeddal et al. [1] which relies on LLM-written ground truth answers, and additionally crowdsource human-written answers and pairwise preference judgments. This allows us to investigate (1) how robust reference-based evaluation measures are to gold reference variation; (2) to what extent do LLM judges agree with human annotators; and (3) which automatic measures best predict human and LLM preferences? Our findings reveal substantial sensitivity of reference-based measures to gold reference choice, and human and LLM judges exhibiting low agreement on preference judgments, despite similar aggregate tendencies. Furthermore, no automatic evaluation measure strongly predicts human preferences, suggesting a fundamental methodological shortcoming in current AIR evaluation practices.

P012 · **When LLM Judges Inflate Scores: Exploring Overrating in Relevance Assessment** [Short]

Chuting Yu (The University of Queensland); Hang Li (The University of Queensland); Guido Zuccon (The University of Queensland); Joel Mackenzie (The University of Queensland); Teerapong Leelanupab (The University of Queensland)

Human relevance assessment is time-consuming and cognitively intensive, limiting the scalability of Information Retrieval evaluation. This has led to growing interest in using large language models (LLMs) as proxies for human judges. However, it remains an open question whether LLM-based relevance judgments are reliable, stable, and rigorous enough to match humans for relevance assessment. In this work, we conduct a study of overrating behavior in LLM-based relevance judgments across model backbones, evaluation paradigms (pointwise and pairwise), and passage modification strategies. We show that models consistently assign inflated relevance scores-often with high confidence-to passages that do not genuinely satisfy the underlying information need, revealing a system-wide bias rather than random fluctuations in judgment. Furthermore, controlled experiments show that LLM-based relevance judgments can be highly sensitive to passage length and surface-level lexical cues. These results raise concerns about the usage of LLMs as drop-in replacements for human relevance assessors, and highlight the urgent need for careful diagnostic evaluation frameworks when applying LLMs for relevance assessments. Our code and results are publicly available. https://github.com/chutingyu/Exploring-Overrating

P013 · **Beyond Top-e: Simulation-Based Interactive Evaluation for Query Suggestions** [Resource]

Jorge Gabín (Universidade da Coruña); Javier Parapar (University of A Coruña); Xi Wang (University of Sheffield)

Evaluating query suggestion systems in a manner that reflects real-world query formulation remains a persistent challenge. Most offline methodologies adopt static assumptions, such as users accepting all or the top-e suggestions, ignoring the inherently selective and intent-driven nature of interactive search. While online experiments provide realistic behavioural signals, they are costly, difficult to scale, and often irreproducible.To bridge this gap, we introduce SIQSE (Simulation-based Interactive Query Suggestion Evaluation), a framework that models query reformulation as an interactive selection task performed by a simulated user. In SIQSE, a Large Language Model (LLM) acts as a surrogate user that progressively selects suggestions according to contextual relevance and explicit search intent. Unlike static offline protocols, this simulation captures the iterative and selective dynamics of real query formulation. Our contributions are twofold. First, we develop and validate an LLM-based selection model, systematically analysing how varying levels of intent information and selection strategies affect its ability to approximate human selection behaviour. Second, we employ this selector to benchmark multiple query suggestion systems across

diverse datasets under interactive conditions. Importantly, while the selector is LLM-based, the final evaluation is computed exclusively through ranking-based effectiveness metrics over the rankings produced by selected expansions, ensuring that system performance reflects retrieval quality rather than alignment with the surrogate user model.By modelling round-based interaction while maintaining metric independence, SIQSE offers a scalable, reproducible evaluation paradigm that brings offline assessment closer to the complexity of real-world search behaviour. To facilitate adoption and reproducibility, we release SIQSE as an open-source Python library\footnote{\faGithub\hspace{0.7mm}\url{https://github.com/IRLab-UDC/SIQSE}}.

P014 · **Multilingual and Domain-Agnostic Tip-of-the-Tongue Query Generation for Simulated Evaluation** [Resource]

Xuhong He (Carnegie Mellon University); To Eun Kim (Carnegie Mellon University); Maik Fröbe (Friedrich-Schiller-Universität Jena); Jaime Arguello (UNC Chapel Hill); Bhaskar Mitra (Independent Researcher); Fernando Diaz (Carnegie Mellon University)

Tip-of-the-Tongue (ToT) retrieval benchmarks have largely focused on English, limiting their applicability to multilingual information access. In this work, we construct multilingual ToT test collections for Chinese, Japanese, Korean, and English, using an LLM-based query simulation framework. We systematically study how prompt language and source document language affect the fidelity of simulated ToT queries, validating synthetic queries through system rank correlation against real user queries. Our results show that effective ToT simulation requires language-aware design choices: non-English language sources are generally important, while English Wikipedia can be beneficial when non-English sources provide insufficient information for query generation. Based on these findings, we release four ToT test collections with 5,000 queries per language across multiple domains. This work provides the first large-scale multilingual ToT benchmark and offers practical guidance for constructing realistic ToT datasets beyond English.

P015 · **Efficient Multivector Retrieval with Token-Aware Clustering and Hierarchical Indexing** [Short]

Silvio Martinico (University of Pisa); Franco Maria Nardini (ISTI-CNR); Cosimo Rulli (ISTI-CNR); Rossano Venturini (University of Pisa)

Multivector retrieval models achieve state-of-the-art effectiveness through fine-grained token-level representations, but their deployment incurs substantial computational and memory costs. Current solutions---based on the well-known κ-means clustering algorithm---group similar vectors together to enable both effective compression and efficient retrieval. However, standard κ-means scales poorly with the number of clusters and dataset size, and favours frequent tokens during training while underrepresenting rare, discriminative ones. In this work, we introduce Tachiom, a multivector retrieval system that exploits token-level structure to significantly accelerate both clustering and retrieval. By accounting for tokens' distribution during centroid allocation, Tachiom easily scales to millions of centroids, enabling highly accurate document scoring using only centroids, avoiding expensive token-level computation. Tachiom combines a graph-based index over centroids with an optimized Product Quantization layout for efficient final scoring. Experiments on Ms Marco-v1 and LoTTE show that Tachiom achieves up to 247× faster clustering than κ-means and up to 9.8× retrieval speedup over state-of-the-art systems while maintaining comparable or superior effectiveness.

P016 · **Does LLM Relevance Labelling Work for Arabic?** [Short]

Marwah Alaofi (Taibah University); Fatima Haouari (The University of Sheffield)

Large Language Models (LLMs) are increasingly used in Information Retrieval, both within retrieval pipelines and for constructing evaluation resources. Existing studies on using LLMs for IR evaluation, however, focus almost exclusively on English, leaving their applicability to other languages, where evaluation resources are often limited and highly needed, unexplored. We examine the use of LLMs to generate relevance labels for an Arabic test collection (ArTest). Using about 10K relevance labels, generated by three LLMs, and used to order eight automated systems and nine simulated manual systems, we show that agreement on binary labels and system ordering is generally high, with no cases of significant opposite conclusions; however, there are cases of false or missed improvements and noticeable limitations when labelling manual, highly performing systems. These findings align with results reported for English and indicate that LLMs could, with some caveats, offer a viable approach to supporting IR evaluation in languages with limited evaluation resources.

P017 · **NERBench-Chhattisgarh: A Multi-Family NER Dataset for Low-Resource Indic Languages** [Resource]

Rajesh Kumar Mundotiya (IIT Bhilai)

We present NERBench-Chhattisgarh, a gold-standard Named Entity Recognition (NER) dataset covering seven under-resourced languages spoken in Central India: Baigani, Chhattisgarhi, Surgujia, Sadri, Kudukh, Halbi, and Gondi. Addressing the digital divide for tribal languages, our corpus comprises 166,444 annotated tokens across 8,391 sentences, spanning both Indo-Aryan and Dravidian language families. The dataset features high lexical sparsity and a ``nature-centric'' ontology of 22 entity types based on the CLIA Phase-II schema, capturing culturally specific entities often absent in standard benchmarks. To establish a benchmark for language variety-aware information access, we evaluate three multilingual encoders: mBERT, XLM-RoBERTa, and IndicBERT, using two adaptation strategies: direct parameter-efficient fine-tuning (LoRA) and a Chhattisgarhi-Pivot Adaptive Pre-training (CPAP) approach. Our results show a morphological barrier: while pivot-based adaptation facilitates transfer for distant languages through script alignment, it induces negative transfer in morphologically complex agglutinative languages like Gondi. We publicly release the dataset, code, and adapted model checkpoints\footnote{\url{https://github.com/Rajesh-NLP/NER-Chhattisgarh}} to support future research in inclusive Information Retrieval.

P018 · **Named Entity-Driven Graph Smoothing to Enhance Pretrained Document Embeddings in Clustering Tasks** [Short]

Imed Keraghel (Centre Borelli UMR9010, Université Paris Cité); Mohamed Nadif (Centre Borelli UMR9010, Université Paris Cité)

Named Entity Recognition (NER) extracts structured semantic elements, such as persons, organizations, or biomedical concepts, that enhance document understanding. Rather than replacing pretrained document embeddings, we refine them by injecting entity-level relational information. By constructing a document graph based on named-entity similarity, we apply spectral filtering to smooth the embedding matrix, yielding entity-aware representations compatible with any clustering algorithm. This framework is model-agnostic and lightweight. Furthermore, it can distill a set of high-confidence Must-Link (ML) constraints, pairs of documents that should be assigned to the same cluster, to guide constrained clustering. Experimental results on benchmark datasets demonstrate that our approach significantly improves clustering performance, highlighting the value of entity-aware smoothing over standard pretrained representations.

P019 · **Selective Distillation for Continual Named Entity Recognition with Memory Replay** [Short]

Weihua Wang (Inner Mongolia University); Yue Cao (Inner Mongolia University); Feilong Bao (Inner Mongolia University)

Continual Named Entity Recognition (CNER) aims to enable models to incrementally learn emerging entity types from new data while avoiding forgetting previously learned ones. However, existing methods tend to ignore class imbalance problem that resulted in consistently performance degradation when encounter rare and new class of entity. To address this problem, we propose a unified framework to defy forgetting through selective knowledge distillation by neurons and memory replay. Specifically, we identify and preserve neurons which is critical to old tasks with gradient-based neuron importance score, and these neurons are further constrained by distillation loss in the new tasks. We also construct a memory buffer that covering multi entity types and their features, and design a hybrid loss tailored to the constructed buffer. By jointly optimizing selective distillation and memory replay in a unified training objective, our framework achieves a balance between stability and plasticity while continual learning new entities. Experiments on multiple standard CNER benchmark datasets demonstrate that our method outperforms existing continual learning baselines, particularly in macro-F1 scores. Our code is available at https://github.com/kzcymos/sel-reply.

P020 · **NanoKnow: How to Know What Your Language Model Knows** [Resource]

Lingwei Gu (University of Waterloo); Nour Jedidi (University of Waterloo); Jimmy Lin (University of Waterloo)

How do large language models (LLMs) know what they know? Answering this question has been difficult because pre-training data is often a "black box" – unknown or inaccessible. The recent release of nanochat – a family of small LLMs with fully open pre-training data – addresses this as it provides a transparent view into where a model's parametric knowledge comes from. Towards the goal of understanding how knowledge is encoded by LLMs, we release NanoKnow, a benchmark dataset that partitions questions from Natural Questions and SQuAD into splits based on whether their answers are present in nanochat's pre-training corpus. Using these splits, we can now properly disentangle the sources of knowledge that LLMs rely on when producing an output. To demonstrate NanoKnow's utility, we conduct experiments using eight nanochat checkpoints. Our findings show: (1) closed-book accuracy is strongly influenced by answer frequency in the pre-training data, (2) providing external evidence can mitigate this frequency dependence, (3) even with external evidence, models are more accurate when answers were seen during pre-training, demonstrating that parametric and external knowledge are complementary, and (4) non-relevant information is harmful, with accuracy decreasing based on both the position and the number of non-relevant contexts. We release all NanoKnow artifacts at https://github.com/castorini/NanoKnow.

P021 · **MCP Servers for Pyserini and RankLLM: Enabling Agentic Retrieval-Augmented Generation** [Demo]

Yijun Ge (University of Waterloo); Zibo Guo (University of Waterloo); Sahel Sharifymoghaddam (University of Waterloo); Jimmy Lin (University of Waterloo)

Large language models (LLMs) increasingly rely on retrieval-augmented generation (RAG), where search results are provided as external context to improve grounding and factuality. The Model Context Protocol (MCP) is an emerging industry standard for connecting LLMs to external tools. We present MCP servers for Pyserini and RankLLM, two widely used information retrieval research toolkits, enabling their retrieval and reranking capabilities to be easily accessed by LLM agents. Our system exposes sparse and dense retrieval over prebuilt indexes, multimodal search, LLM-based reranking, and evaluation based on relevance judgments, all as tools callable by LLMs. This integration supports rapid experimentation with agentic RAG workflows and facilitates comparisons across models and clients. We demonstrate end-to-end retrieval, reranking, RAG, and multimodal search using both open-source and proprietary LLMs, bridging IR research infrastructure with LLM tool ecosystems.

P022 · **LACONIC: Dense-Level Effectiveness for Scalable Sparse Retrieval via a Two-Phase Training Curriculum** [Short]

Zhichao Xu (University of Utah); Shengyao Zhuang (The University of Queensland); Crystina Zhang (University of Waterloo); Xueguang Ma (University of Waterloo); Yijun Tian (University of Notre Dame); Maitrey Mehta (University of Utah); Jimmy Lin (University of Waterloo); Vivek Srikumar (University of Utah)

While dense retrieval models have been the standard for state-of-the-art information retrieval, their deployment is often constrained by high memory requirements and reliance on GPU accelerators for vector similarity search at scale. Learned sparse retrieval offers a compelling alternative by enabling efficient search via inverted indices, yet it has historically received less attention than dense approaches. In this paper, we introduce LACONIC, a family of learned sparse retrievers based on the Llama3 architecture (1B, 3B, and 8B). We propose a streamlined two-phase training curriculum consisting of (1) weakly supervised pre-finetuning to adapt causal LLMs for bidirectional contextualization and (2) high-signal finetuning using curated hard negatives. Our results demonstrate that LACONIC effectively bridges the performance gap with dense models: the 8B variant achieves a state-of-the-art 60.2 nDCG@10 on the MTEB Retrieval benchmark, ranking 15th on the leaderboard as of February 5th, 2026, while utilizing 74% less index memory than an equivalent dense model. By delivering high retrieval effectiveness on commodity CPU hardware with a fraction of the compute budget required by competing models, LACONIC provides a scalable and efficient solution for real-world search applications. We fully open source our code implementation and trained checkpoints to facilitate reproducibility.

P023 · **Beyond Hard Negatives: The Importance of Score Distribution in Knowledge Distillation** [Short]

Youngjoon Jang (Korea University); Seongtae Hong (Korea University); Hyeonseok Moon (Korea University); Heuiseok Lim (Korea University)

Transferring knowledge from a cross-encoder teacher via Knowledge Distillation (KD) has become a standard paradigm for training retrieval models. While existing studies have largely focused on mining hard negatives to improve discrimination, the systematic composition of training data and the resulting teacher score distribution have received relatively less attention. In this work, we highlight that focusing solely on hard negatives prevents the student from learning the comprehensive preference structure of the teacher, potentially hampering generalization. To effectively emulate the teacher score distribution, we propose a Stratified Sampling strategy that uniformly covers the entire score spectrum. Experiments on in-domain and out-of-domain benchmarks confirm that Stratified Sampling, which preserves the variance and entropy of teacher scores, serves as a robust baseline, significantly outperforming top-K and random sampling in diverse settings. These findings suggest that the essence of distillation lies in preserving the diverse range of relative scores perceived by the teacher.

P024 · **CRED: Calibrated Relational Enhanced Distillation for LLM-Based Pointwise Reranking** [Short]

Qingran Yang (Harbin Institute of Technology); Wenxuan Zhang (Harbin Institute of Technology); Yuting Wang (University of International Business and Economics); Xue Li (Harbin Institute of Technology); Lanshun Nie (Harbin Institute of Technology)

In document reranking, rerankers based on Large Language Models (LLMs) demonstrate superior performance but are constrained by high memory consumption and latency. To develop lightweight yet high-performance LLM-based pointwise rerankers through knowledge distillation, we identify two critical limitations: teachers often yield over-smoothed and inaccurate supervision on hard negative samples, thereby hindering the student's optimization; furthermore, traditional methods underutilize the relevance score differences between candidates, which are crucial for ranking tasks. To address these challenges, we propose CRED (Calibrated Relational Enhanced Distillation), which integrates Adaptive Teacher Calibration (ATC) to calibrate teacher predictions and amplify score margins, while employing Preference Relation Alignment (PRA) to align the distributional patterns of relevance score differences, enabling the student to capture precise ranking structures. To support this approach, we also construct FineDistill, a dataset of 1M samples providing fine-grained score supervision. We distill an 8B teacher into a 0.6B pointwise student. Extensive experiments on TREC and BEIR benchmarks show that our model outperforms leading baselines in both performance and generalization.

P025 · **From Continuous Pretraining to Domain-Adaptive Reranking via Task Vector Adaptation** [Short]

Sanghyun Cho (Electronics and Telecommunications Research Institute); Myeongjin Lee (Electronics and Telecommunications Research Institute); Jong-hun Shin (Electronics and Telecommunications Research Institute); Jeong Heo (Electronics and Telecommunications Research Institute); Kiyoung Lee (Electronics and Telecommunications Research Institute); Soojong Lim (Electronics and Telecommunications Research Institute)

Large language model (LLM)-based rerankers have demonstrated strong performance in information retrieval tasks, but adapting them to specialized domains remains challenging due to the substantial cost and effort required to construct high-quality domain-specific training datasets. To address this limitation, we leverage task vectors derived from continuous pretraining as a mechanism for transferring domain knowledge. However, existing task vector integration methods are highly sensitive to scaling factors and can lead to unstable performance across domains and model scales. In this work, we propose a fine-grained task vector adaptation method that learns parameter-wise scaling coefficients for the task vector. These coefficients are optimized using a language modeling objective while keeping all model parameters fixed, enabling effective integration of domain-specific knowledge without degrading reranking capabilities. Experiments on a general-domain benchmark and five specialized domains across two model scales demonstrate that our method provides stable improvements across most domains and avoids the degradation observed with fixed task vector scaling.

P026 · **CroSearch-R1: Better Leveraging Cross-lingual Knowledge for Retrieval-Augmented Generation** [Short]

Rui Qi (Beijing Jiaotong University); Fengran Mo (Université de Montréal); Sijin Lu (Beijing Jiaotong University); Yufeng Chen (Beijing jiaotong univercity); Jian-Yun Nie (Université de Montréal); Kaiyu Huang (Beijing Jiaotong University)

A multilingual collection may contain useful knowledge in other languages to supplement and correct the facts in the original language for Retrieval-Augmented Generation (RAG). However, the vanilla approach that simply concatenates multiple pieces of knowledge from different languages into the context may fail to improve effectiveness due to the potential disparities across languages. To better leverage multilingual knowledge, we propose CroSearch-R1, a search-augmented reinforcement learning framework to integrate multilingual knowledge into the Group Relative Policy Optimization (GRPO) process. In particular, the approach adopts a multi-turn retrieval strategy with cross-lingual knowledge integration to dynamically align the knowledge from other languages as supplementary evidence into a unified representation space. Furthermore, we introduce a multilingual rollout mechanism to optimize reasoning transferability across languages. Experimental results demonstrate that our framework effectively leverages cross-lingual complementarity and improves the effectiveness of RAG with multilingual collections.

P027 · **Conv-FinRe: A Conversational and Longitudinal Benchmark for Utility-Grounded Financial Recommendation** [Resource]

Yan Wang (The Fin AI); Yi Han (Georgia Institute of Technology); Lingfei Qian (The Fin AI); Yueru He (Columbia University); Xueqing Peng (The Fin AI); Dongji Feng (California State University, Monterey Bay); Zhuohan Xie (Mohamed bin Zayed University of Artificial Intelligence); Vincent Jim Zhang (The Fin AI); Yuqing Guo (The Fin AI); Fengran Mo (University of Montreal); Jimin Huang (The Fin AI); Yankai Chen (Mohamed bin Zayed University of Artificial Intelligence); Jian-Yun Nie (University of Montreal)

Most recommendation benchmarks evaluate how well a model imitates user behavior. In financial advisory, however, observed actions can be noisy or short-sighted under market volatility and may conflict with a user's long-term goals. Treating what users chose as the sole ground truth, therefore, conflates behavioral imitation with decision quality. We introduce Conv-FinRe, a conversational and longitudinal benchmark for stock recommendation that evaluates LLMs beyond behavior matching. Given an onboarding interview, step-wise market context, and advisory dialogues, models must generate rankings over a fixed investment horizon. Crucially, Conv-FinRe provides multi-view references that distinguish descriptive behavior from normative utility grounded in investor-specific risk preferences, enabling diagnosis of whether an LLM follows rational analysis, mimics user noise, or is driven by market momentum. We build the benchmark from real market data and human decision trajectories, instantiate controlled advisory conversations, and evaluate a suite of state-of-the-art LLMs. Results reveal a persistent tension between rational decision quality and behavioral alignment: models that perform well on utility-based ranking often fail to match user choices, whereas behaviorally aligned models can overfit short-term noise. The dataset is publicly released on Hugging Face. https://huggingface.co/collections/TheFinAI/conv-finre, and the codebase is available on GitHub. https://github.com/The-FinAI/Conv-FinRe.

P028 · **Learning to Route Queries to Heads for Attention-based Re-ranking with Large Language Models** [Short]

Yuxing Tian (Université de Montréal); Fengran Mo (Université de Montréal); Zhiqi Huang (CapitalOne); Weixu Zhang (McGill University); Jian-Yun Nie (Université de Montréal)

Large Language Models (LLMs) have recently been explored as fine-grained zero-shot re-rankers by leveraging attention signals to estimate document relevance. However, existing methods either aggregate attention signals across all heads or rely on a statically selected subset identified by heuristic rules. This solution can be suboptimal because the informative heads can vary across queries or domains. Moreover, naively combining multiple heads can degrade performance due to redundancy or conflicting ranking signals. In this paper, we propose a query-dependent head selection method, RouteHead, for attention-based re-ranking with LLMs. Specifically, we learn a lightweight router that can map each query to an optimal head set, and relevance scores are computed by aggregating attention signals only from these heads. Since query-to-head optimal labels are unavailable, we first construct pseudo labels via an offline search. The router represents each head with a learnable embedding and represents each query using an embedding extracted from the hidden states of the frozen LLM. Then it is trained on the pseudo labels with a sparsity regularizer. Experiments on diverse benchmarks and multiple LLM backbones show that the proposed method consistently outperforms strong baselines.

P029 · **ItemRAG: Item-Based Retrieval-Augmented Generation for LLM-Based Recommendation** [Short]

Sunwoo Kim (KAIST); Geon Lee (KAIST); Kyungho Kim (KAIST); Jaemin Yoo (Seoul National University); Kijung Shin (KAIST)

Recently, large language models (LLMs) have been widely used as recommender systems, owing to their reasoning capability and effectiveness

in handling cold-start items. A common approach prompts an LLM with a target user's purchase history to recommend items from a candidate set, often enhanced with retrieval-augmented generation (RAG). Most existing RAG approaches retrieve purchase histories of users similar to the target user; however, these histories often contain noisy or weakly relevant information and provide little or no useful information for candidate items. To address these limitations, we propose ItemRAG, a novel RAG approach that shifts focus from coarse user-history retrieval to fine-grained item-level retrieval. ItemRAG augments the description of each item in the target user's history or the candidate set by retrieving items relevant to each. To retrieve items not merely semantically similar but informative for recommendation, ItemRAG leverages co-purchase information alongside semantic information. Especially, through their careful combination, ItemRAG prioritizes more informative retrievals and also benefits cold-start items. Through extensive experiments, we demonstrate that ItemRAG consistently outperforms existing RAG approaches under both standard and cold-start item recommendation settings.

P030 · **TRACE: A Conversational Framework for Sustainable Tourism Recommendation with Agentic Counterfactual Explanations** [Demo]

Ashmi Banerjee (Technical University of Munich); Adithi Satish (Technical University of Munich); Wolfgang Wörndl (Technical University of Munich (TUM)); Yashar Deldjoo (Polytechnic University of Bari)

Traditional conversational travel recommender systems primarily optimize for user relevance and convenience, often reinforcing popular, overcrowded destinations and carbon-intensive travel choices. To address this, we present TRACE (Tourism Recommendation with Agentic Counterfactual Explanations), a multi-agent, LLM-based framework that promotes sustainable tourism through interactive nudging. TRACE uses a modular orchestrator-worker architecture where specialized agents elicit latent sustainability preferences, construct structured user personas, and generate recommendations that balance relevance with environmental impact. A key innovation lies in its use of agentic counterfactual explanations and LLM-driven clarifying questions, which together surface greener alternatives and refine understanding of intent, fostering user reflection without coercion. User studies and semantic alignment analyses demonstrate that TRACE effectively supports sustainable decision-making while preserving recommendation quality and interactive responsiveness. TRACE is implemented on Google's Agent Development Kit, with full code, Docker setup, prompts, and a publicly available demo video to ensure reproducibility. A project summary, including all resources, prompts, and demo access, is available at https://ashmibanerjee.github.io/trace-chatbot.

P031 · **XFoodRec: An Explainable Mobile Recommender for Personalized Healthy Eating** [Demo]

Amir Mollazadeh (University of Oulu, CMVS); Mourad Oussalah (University of Oulu, CMVS); Mehrdad Rostami (University of Oulu, CMVS)

Food recommender systems often struggle with popularity bias that favors unhealthy recipes, making the promotion of healthier items a significant communication challenge. While explainable recommendations can help justify healthy trade-offs, evaluating their real-world impact is often hindered by the technical complexity of conducting online user studies. We present XFoodRec, an extensible mobile research platform designed to bridge the gap between offline algorithmic evaluation and online behavioral studies. The system integrates personalized recipe recommendation with an LLM-based explanation pipeline that translates structured nutritional data into persuasive natural-language explanations. Its built-in A/B testing framework automatically assigns users to experimental groups to measure how explanations influence real-world user behavior and engagement metrics. By providing a modular and deployable infrastructure for controlled experimentation, XFoodRec supports end-users in making healthier meal choices while enabling researchers to investigate how transparency mechanisms influence decision-making in real-world mobile settings. The demonstration video and all code resources are publicly available at: https://github.com/Amir-Mol/XFoodRec.

P032 · **DisCoRec: Disentangled Conformity-aware Recommendation with LLM-Guided Multi-View Learning** [Short]

Minkyung Song (Chungnam National University); Soyoung Park (Chungnam National University); Sungsu Lim (Chungnam National University)

User–item interactions in recommender systems are shaped by intertwined factors, such as genuine intent and social conformity. Recent Large Language Model (LLM)-based recommenders leverage textual semantics to improve accuracy and explainability, but often fail to disentangle these factors, limiting personalized recommendations. Even in explicitly disentangled models, attention-based explanations can be unfaithful, as attention weights do not reliably indicate importance. To address these issues, we propose DisCoRec, a multi-view framework that integrates LLM semantics into each view and calibrates view weights via LLM-guided gating. Experiments on Amazon benchmarks show DisCoRec achieves consistent gains over disentanglement-based and LLM-based recommenders, with an average improvement of about 5% over the strongest baseline, while improving robustness and explainability.

P033 · **MuPP: Multi-Preference Padding for Sequential Recommendation** [Short]

Wooseung Kang (Gyeongsang National University); Minje Kim (Gyeongsang National University); Suwon Lee (Gyeongsang National University); Gun-Woo Kim (Gyeongsang National University); Sang-Min Choi (Gyeongsang National University)

Sequential recommendation (SR) systems aim to provide personalized suggestions by analyzing users' historical behavior. However, SR faces significant challenges due to data sparsity that arises from insufficient user interaction data. To address the sparsity in behavior sequences, existing research has applied padding techniques and developed various augmentation methods to enhance recommendation performance. Despite these improvements, SR models still face challenges from data sparsity, especially for users with highly sparse interaction histories. Recent studies have explored augmenting behavior sequences or filling padded portions with repeated identical items. However, the former approach demands high computational resources due to redundant generation of user representations, while the latter reinforces redundant item embeddings, which limits the expressiveness of user preferences and reduces representational diversity by repeatedly inserting identical items. To overcome these limitations, we propose a novel padding method called Multi-Preference Padding (MuPP). Specifically, we employ a Multi-Preference Layer to generate diverse sub-preference representations from each item in the user's sequence to fill padded sequences. By generating representations that explicitly capture items' diverse sub-preferences, our approach alleviates representational redundancy caused by repetitive insertions of identical items and enriches the expressiveness of user preference representations. Extensive experiments conducted on four benchmark datasets demonstrate the effectiveness of MuPP across various SR models. Our code is available at https://github.com/kangvic0615/MuPP

P034 · **REMICA: Reflective Memory and Interventional Context Alignment with Multi-Agent LLMs for Inappropriate Utterance Detection** [Short]

Juyoung Kim (Gyeongsang National University); Ji-Hong Park (Gyeongsang National University); Sang-Min Choi (Gyeongsang National University); Gun-Woo Kim (Gyeongsang National University)

Recently, multi-agent debate-based inference methods have been widely adopted for inappropriate utterance detection to aggregate diverse perspectives and improve explainability. However, performing real-time interactions for every input incurs substantial latency and cost, and may amplify error entrenchment as early misjudgments are repeatedly justified. In this paper, we propose the Reflective Memory and Interventional Context Alignment (REMICA) framework. Specifically, REM stores multi-agent predictions and rationales as offline memory for efficient retrieval-based inference. During this process, ICA enforces alignment constraints to ensure consistency between rationales and predictions. Experiments on four datasets with three commercial LLMs show that REMICA reduces the average inference time relative to online multi-agent inference while improving the average F1 score, with GPT-5.1-mini achieving a relative average F1 improvement of +2.594% and an average inference-time reduction of 7.80× in the best-performing setting. These results suggest that the proposed framework can help mitigate the trade-off between accuracy and latency by reusing offline memory and reducing the risk of error entrenchment. Furthermore, the memory and code developed in this paper are publicly available at https://github.com/wndudwkd003/REMICA.

P035 · **Visual-RAG: Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries** [Resource]

Yin Wu (Nanyang Technological University); Quanyu Long (Nanyang Technological University); Jing Li (Harbin Institute of Technology (Shenzhen)); Jianfei Yu (Nanjing University of Science and Technology); Wenya Wang (Nanyang Technological University)

Retrieval-augmented generation (RAG) augments large language models (LLMs) with external knowledge to tackle knowledge-intensive question answering.While several benchmarks evaluate multimodal LLMs (MLLMs) under multimodal RAG settings, they predominantly retrieve from textual corpora and do not explicitly assess how models exploit visual evidence during answer generation. Consequently, there still lacks benchmark that cleanly isolates and measures the contribution of retrieved images in a visual knowledge-intensive RAG pipeline.We introduce Visual-RAG, a question-answering benchmark that targets visually-grounded, knowledge-intensive questions in a visual evidence-centric manner. Unlike prior work, Visual-RAG requires text-to-image retrieval and the integration of retrieved clue images whose pixel content explicitly encodes the visual knowledge necessary for answer generation.With Visual-RAG, we evaluate five open-source and three proprietary MLLMs and find that current systems still substantially underutilize the visual information available in retrieved images. Despite clear opportunities for multimodal evidence integration, state-of-the-art models struggle to extract and exploit fine-grained visual knowledge, and text-to-image retrieval itself remains challenging even under constrained entity-level corpora. These results underscore the need for improved visual retrieval, grounding, and attribution in multimodal RAG.\footnote{\Visual-RAG is publicly available at: \url{github.com/visual-rag/visual-rag}}

P036 · **MEG-RAG: Quantifying Multi-modal Evidence Grounding for Evidence Selection in RAG** [Short]

Xihang Wang (Zhejiang University); Zihan Wang (Peking University); Chengkai Huang (University of New South Wales); Quan Z. Sheng (Macquarie University); Lina Yao (University of New South Wales)

Multimodal Retrieval-Augmented Generation (MRAG) addresses key limitations of Multimodal Large Language Models (MLLMs), such as hallucination and outdated knowledge. However, current MRAG systems struggle to distinguish whether retrieved multimodal data truly supports the

semantic core of an answer or merely provides superficial relevance. Existing metrics often rely on heuristic position-based confidence, which fails to capture the informational density of multimodal entities. To address this, we propose Multi-modal Evidence Grounding (MEG), a semantic-aware metric that quantifies the contribution of retrieved evidence. Unlike standard confidence measures, MEG utilizes Semantic Certainty Anchoring, which dynamically filters out high-frequency stopwords via Inverse Document Frequency (IDF) to focus strictly on information-bearing tokens. Building on MEG, we introduce MEG-RAG, a framework that trains a multimodal reranker to align retrieved evidence with the semantic anchors of the ground truth. By prioritizing high-value content based on semantic grounding rather than token probability distributions, MEG-RAG improves the accuracy and multimodal consistency of generated outputs. Extensive experiments on the ???2RAG benchmark show that MEG-RAG consistently outperforms strong baselines and demonstrates robust generalization across different teacher models. The data and code are available at here.

P037 · **SocialDropout: Dynamic Agent Dropout for Social Simulation** [Short]

Huajie Wang (Harbin Institute of Technology); Geng Tu (Harbin Institute of Technology); Xi Zeng (The 30th Research Institute of China Electronics Technology Group Corporation); Ruifeng Xu (Harbin Institute of Technology); Min Zhang (Harbin Institute of Technology)

Large language model driven multi-agent social simulation frameworks enable realistic modeling of complex societal dynamics but incur substantial computational overhead due to dense agent participation and extensive interaction costs. To address this limitation, we propose SocialDropout, a reinforcement learning–based agent selection strategy within the AgentSociety framework, inspired by the AgentDropout paradigm, which dynamically identifies and samples informative agent subsets for each simulation round. Each agent is assigned an adaptive importance weight optimized to jointly minimize agent sparsity and computational cost measured by LLM calls, token consumption, and execution time—while preserving social interaction intensity within the environment. Extensive performance evaluation demonstrates that the proposed method significantly improves simulation efficiency and scalability. Moreover, ablation studies verify that high-level behavioral realism and outcome consistency are largely maintained despite substantial agent reduction. Our approach offers a practical and general optimization mechanism for large-scale LLM-based multi-agent social simulations under constrained computational budgets.

P038 · **Towards Unified Affective Reasoning via In-Context Reinforcement Learning** [Short]

Feng Xiong (Harbin Institute of Technology); Jun Wang (Harbin Institute of Technology); Ruifeng Xu (Harbin Institute of Technology)

Recent advances in Large Language Models (LLMs) have shifted affective computing from direct label prediction to deliberative reasoning, with Reinforcement Learning (RL) emerging as a promising means of cultivating such capabilities. However, existing RL-based methods are largely confined to single-task settings, requiring separate training pipelines for different affective domains and limiting transfer across related problems. To address this limitation, we propose Unified Affective Reasoning (UAR), an in-context reinforcement learning framework for learning generalizable affective reasoning under a unified multi-task paradigm. UAR jointly optimizes multiple affective tasks and leverages in-context demonstrations to reconcile heterogeneous annotation schemas across datasets, thereby aligning disparate label spaces while preserving task-specific supervision. Experiments on 16 standard benchmarks across emotion recognition, aspect-based sentiment analysis, stance detection, and sarcasm detection show that UAR consistently improves the underlying base models and generalizes effectively to held-out out-of-domain tasks. These results demonstrate that in-context reinforcement learning provides an effective path toward unified and generalizable affective reasoning. Code is available at https://github.com/farisxiong/UAR.

P039 · **PeaCap: Patch-Level Retrieval for Lightweight Retrieval-Augmented Image Captioning** [Short]

Robin Viltoriano (Adelaide University); Wei Emma Zhang (Adelaide University); Hu Wang (Mohamed bin Zayed University of Artificial Intelligence); Mong Yuan Sim (Adelaide University); Yanjun Shu (Harbin Institute of Technology)

Retrieval-augmented image captioning aims to improve caption quality by grounding generation in external evidence, but most prior systems retrieve evidence using coarse whole-image similarity, which can miss small or rare objects in cluttered scenes. We propose PeaCap, a patch-based retrieval-augmented captioning framework that explicitly studies how retrieval granularity affects the quality of retrieved object evidence and downstream caption generation. PeaCap decomposes a query image into patches, performs patch-level image-to-image retrieval to obtain object tags, and fuses the retrieved tags with the whole-image embedding via a lightweight cross-attention module and an alignment loss to robustly prompt a frozen LLM. Analyses on retrieval (encoder choice, patch-vs.-whole retrieval, and patch-grid ablations) show that patch-level retrieval can improve object coverage, and experiments on COCO and out-of-domain benchmarks demonstrate competitive captioning performance under a lightweight training setup.

P040 · **H-MAPS: Hierarchical Memory-Augmented Proactive Search Assistant for Scientific Literature** [Demo]

Koji Nishikawa (University of Tsukuba); Makoto P. Kato (University of Tsukuba)

Scientific reading is an active process that frequently requires consulting external resources, but manual keyword searching interrupts the reading flow and imposes a high cognitive load. Existing proactive information retrieval systems often suffer from context ambiguity, as they rely solely on on-screen text and ignore the reader's specific background and intent. In this demonstration, we present H-MAPS (Hierarchical Memory-Augmented Proactive Search Assistant), a proactive literature exploration assistant that resolves this ambiguity by leveraging a three-layered hierarchical memory. Triggered by implicit reading behaviors, H-MAPS articulates the user's latent information needs into explicit natural language questions and performs neural retrieval entirely on the local device to ensure privacy. We demonstrate H-MAPS using a scenario where two researchers, specializing in NLP and HCI, read the same paper. In response, the system generates profile-specific questions and retrieves distinct literature tailored to each user.

P041 · **EviRAG: Evidence-Guided Retrieval-Augmented Generation for Medical Vision-Language Models** [Short]

Yiyang Gu (Peking University); Jiayue Fan (Peking University); Kaili Liu (Xidian University); Bohan Wu (Peking University); Binqi Chen (Peking University); Zequn Liu (Zhongguancun Academy); Zhiping Xiao (University of Washington); Rong-Cheng Tu (Nanyang Technological University); Xiao Luo (University of Wisconsin-Madison); Ming Zhang (Peking University)

Retrieval-augmented generation (RAG) is widely adopted for radiology report generation with medical vision-language models, leveraging external reports as linguistic references. However, existing RAG methods rely primarily on dense embedding similarity, which may retrieve reports that are semantically related yet clinically inconsistent with respect to presence or laterality constraints. Such inconsistencies are often propagated into generation, resulting in contradictory or unsupported findings. We propose an evidence-guided retrieval-augmented framework EviRAG that decomposes retrieval into structured and unstructured alignment levels. First, we induce structured clinical triplets from both query and database cases through targeted visual interrogation, projecting images into a shared evidence space. Triplet-level alignment enforces explicit agreement over presence and laterality variables, yielding a clinically admissible candidate set via structural ranking. Within this constrained space, we perform semantic alignment in a shared multimodal embedding space to capture nuanced descriptive correspondence. The top-ranked reports and query image are jointly fed into a medical vision-language model for report generation. Comprehensive experiments on radiology report generation benchmarks show that EviRAG substantially reduces clinical inconsistencies compared to strong medical vision-language baselines. The source code is available at https://github.com/liamgu06/EviRAG.

P042 · **GeoSearch: Augmenting Worldwide Geolocalization with Web-Scale Reverse Image Search and Image Matching** [Short]

Tung-Duong Le-Duc (University of Science, VNU-HCM); Hoang-Quoc Nguyen-Son (National Institute of Information and Communications Technology); Minh-Son Dao (National Institute of Information and Communications Technology)

Worldwide image geolocalization, which aims to predict the GPS coordinates of any image on Earth, remains challenging due to global visual diversity. Recent generative approaches based on Retrieval-Augmented Generation (RAG) and Large Multimodal Models (LMMs) leverage candidates retrieved from fixed databases for reasoning, but often struggle with scenes that are absent from the reference set. In this work, we propose \textbf{GeoSearch}, an open-world geolocation framework that integrates web-scale reverse image search into the RAG pipeline. GeoSearch augments LMM prompts with database-retrieved coordinates and textual evidence extracted from web pages. To mitigate noise from irrelevant content, we introduce a two-layer filtering mechanism consisting of image matching, followed by confidence-based gating. Experiments on standard benchmarks Im2GPS3k and YFCC4k demonstrate the superiority of GeoSearch under leakage-aware evaluation. Our code and data are publicly available to support reproducibility.

P043 · **Retrieval-Augmented Contrastive Learning for Knowledge Tracing** [Short]

Kamal Berahmand (RMIT University); Mehrnoush Mohammadi (The University of Queensland); Homa Babai (Monash University); Hassan Khosravi (The University of Queensland)

Knowledge Tracing (KT) models aim to predict student performance from interaction histories in order to support personalised learning. However, many learners generate only limited interaction data, making reliable knowledge-state estimation difficult. Recent contrastive KT methods attempt to address this data sparsity through self-supervised representation learning from augmented versions of individual learner sequences, operating within an intra-learner paradigm, where contrastive signals are derived solely from variations of a single learner's trajectory. We advance prior work by proposing RACL (Retrieval-Augmented Contrastive Learning), a knowledge tracing framework that introduces an inter-learner contrastive paradigm, leveraging the observation that students with similar skill profiles often exhibit comparable learning trajectories. Cross-learner structure therefore provides naturally occurring positive and negative examples that are more pedagogically meaningful than synthetic augmentations. Experiments on four benchmarks demonstrate that RACL achieves +1.2% average AUC improvement over state-of-the-art methods, with 97% performance retention at 20% training data, indicating improved robustness under sparse-learning conditions.

P044 · **Task-Adaptive Retrieval over Agentic Multi-Modal Web Histories via Learned Graph Memory** [Short]

Saman Forouzandeh (School of Engineering, Royal Melbourne Institute of Technology University); Kamal Berahmand (School of Engineering, Royal Melbourne Institute of Technology University); Mahdi Jalili (School of Engineering, Royal Melbourne Institute of Technology University)

Retrieving relevant observations from long multi-modal web interaction histories is challenging because relevance depends on the evolving task state, modality (screenshots, HTML text, structured signals), and temporal distance. Prior approaches typically rely on static similarity thresholds or fixed-capacity buffers, which fail to adapt relevance to the current task context. We propose \textbf{ACGM}, a learned graph-memory retriever that constructs \emph{task-adaptive} relevance graphs over agent histories using policy-gradient optimization from downstream task success. ACGM captures heterogeneous temporal dynamics with modality-specific decay (visual decays $4.3\times$ faster than text: $\lambda_v{=}0.47$ vs.\ $\lambda_x{=}0.11$) and learns sparse connectivity (3.2 edges/node), enabling efficient $O(\log T)$ retrieval. Across WebShop, VisualWebArena, and Mind2Web, ACGM improves retrieval quality to \textbf{82.7 nDCG@10} (+9.3 over GPT-4o, $p{<}0.001$) and \textbf{89.2\% Precision@10} (+7.7), outperforming 19 strong dense, re-ranking, multi-modal, and graph-based baselines. Code to reproduce our results is available at{\color{blue}\href{https://github.com/S-Forouzandeh/ACGM-Agentic-Web}{Saman Forouzandeh}}.

P045 · **GATHER: Convergence-Centric Hyper-Entity Retrieval for Zero-Shot Cell-Type Annotation** [Short]

Zhonghui Zhang (Shenzhen University of Advanced Technology); Feng Jiang (Shenzhen University of Advanced Technology); Shaowei Qin (Shenzhen University of Advanced Technology); Jiahao Zhao (Northeastern University); Min Yang (Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences)

Zero-shot single-cell cell-type annotation aims to determine a cell's type from a given set of expressed genes without any training. Existing knowledge-graph-based RAG approaches retrieve evidence by expanding from source entities and relying on iterative LLM reasoning. However, in this setting each query contains tens to hundreds of genes, where no single gene is decisive and the label emerges only from their collective co-occurrence. Such hyper-entity queries fundamentally challenge local, entity-wise exploration strategies, which reason from individual genes, leading to poor scalability and substantial LLM cost. We propose GATHER (Graph-Aware Traversal with Hyper-Entity Retrieval), a convergence-centric retriever tailored to hyper-entity queries. It performs global multi-source graph traversal and identifies topological convergence points, nodes jointly reachable from many input genes. These convergence nodes act as high-information hyper-entities that capture entity synergy. By incorporating node- and path-importance scoring, GATHER selects informative evidence entirely without LLM involvement during retrieval. Instantiated on a self-constructed cell-centric biological knowledge graph (VCKG), GATHER outperforms strong KG-RAG baselines (ToG, ToG-2, RoG, PoG) on two datasets (Immune and Lung), achieving the highest exact-match accuracy (27.45% and 59.64%) with only a single LLM call per sample, compared to 2 to 61 calls for KG-RAG baselines. Our results demonstrate that convergence nodes compress multi-entity signals into compact, high-information evidence that conveys more per item than multi-hop paths, providing an efficient global alternative to local entity-wise reasoning.

P046 · **GONets: A First-Look Into GitHub Organisation Networks** [Resource]

Hridoy Sankar Dutta (Deakin University); Biswadeep Khan (Stanford University); Parth Mitesh Shah (Ahmedabad University); Amit A. Nanavati (Ahmedabad University)

The rapid growth of open-source software has made GitHub a central platform for studying large-scale collaboration for software development. Existing datasets and event logs are typically limited to individual projects or user–repository collaboration at the GitHub level. We introduce GONets, a publicly available dataset that provides a direct bipartite representation of user–repository interactions at the {\it organisational} level. GONets comprises over 230K GitHub organisations, 2.52M unique users, 4.37M repositories, and nearly 30M contribution edges, enriched with detailed organisational metadata including organisation size, creation date and follower count. We present an automated data collection pipeline that retrieves organisational and membership information via GitHub's API, aggregates user-repository contribution data, and constructs large-scale bipartite networks suitable for network analysis and modelling. Using this dataset, we conduct the first large-scale characterisation of the GitHub organisational ecosystem, analysing organisational attributes and user-repository contribution patterns across a diverse range of organisations. We release the complete dataset and collection pipeline to support reproducible research and facilitate future studies on topics including code reuse diffusion, collaboration dynamics and the security ecosystem of GitHub organisations. The entire dataset can be accessed at: \url{https://zenodo.org/records/18472549}.

P047 · **TDHGNN: A Temporal Directed Hypergraph Neural Network for Bitcoin Fraud Detection** [Short]

Zheng Gong (The Hong Kong University of Science and Technology (Guangzhou)); Shuheng Shen (Ant Group); Changhua Meng (Ant Group); Ying Sun (Hong Kong University of Science and Technology (Guangzhou))

The rapid expansion of cryptocurrencies, especially Bitcoin, has heightened the need for efficient detection of illicit transactions, including ransomware payments and money laundering. To effectively address the unique challenges of Bitcoin fraud detection, such as handling heterogeneous, directional, and time-sensitive transaction patterns within the UTXO model, we propose a novel hypergraph-based representation learning framework. Central to our approach is a temporal directed hypergraph data structure that comprehensively encapsulates the multifaceted interactions among transactions and addresses. Building upon this formulation, we develop the Temporal Directed HyperGraph Neural Network (TDHGNN), which learns expressive and integrated embeddings for both transactions and addresses. Extensive experiments demonstrate that our method substantially enhances fraud detection performance, achieving improvements of 12.0% in AUROC for illicit address detection and 5.2% for transaction detection compared to existing approaches. Our code, data and scripts can be found here.

P048 · **F2-Gen: An Open-Source Web Platform for Scenario-Driven Financial Fraud Data Simulation** [Resource]

Junquan Gu (Shanghai University); Zehao Gong (Shanghai University); Runchen Ji (Shanghai University); Kun Shi (Shanghai University); Xiangfeng Luo (Shanghai University); Hang Yu (Shanghai University)

Real bank-transaction data are rarely available to non-regulatory parties due to privacy and compliance constraints, while public substitutes often mismatch financial semantics, limiting the transferability and reproducibility of anomaly-detection research in financial settings. To address this, we release F2-Gen: an open-source, scenario-driven, configurable web platform that generates synthetic data covering both tabular transaction records and transaction graphs for modeling and evaluating financial-fraud behaviors. F2-Gen provides parameterized fraud scenarios, interactive visual analysis, automated screening to remove label-revealing fields, tutorials, and batch data export. Use cases include anomaly detection, graph learning, and teaching demonstrations; standard benchmarking is supported via a screened feature subset, ready-to-run baseline scripts, and sanity/controllability checks. Project page: https://sethgu.github.io/FinancialFraudDataGenerator/. Code: https://github.com/sethGu/FinancialFraudDataGenerator.

P049 · **Prototype-Calibrated Graph Prompting for Few-Shot Graph Adaptation** [Short]

Yumeng Zhao (University of Science and Technology of China); Yan Yan (Southeast University); Chuang Zhao (University of Science and Technology of China); Yingzi Shi (Nanjing University); Bei Hua (University of Science and Technology of China); Shuo Wen (McGill University)

Graph Neural Networks (GNNs) increasingly follow the "pre-training, adaptation" paradigm, where a GNN is pre-trained on large-scale graphs and then adapted to downstream tasks. Graph prompting adapts to the frozen encoder by modifying the input graph structure, rather than fine-tuning the model parameters. However, existing graph prompting methods often rely on probabilistic rewiring and auxiliary regularizers to control sparsity, which makes the prompting process sensitive to hyperparameters and can introduce instability in few-shot settings. To address the issue, we propose ProtoCalib, a lightweight local graph prompt for few-shot adaptation of frozen GNNs. ProtoCalib uses prototypes from the support set to score candidate edges for each anchor node. It then calibrates these scores with a per-node edge budget, which keeps the prompted graph sparse and removes the need for extra sparsity or entropy losses. We further adopt a deterministic construction of the prompted adjacency to reduce sampling noise at inference time. Extensive experiments on five graph datasets under four pre-training strategies demonstrate that our proposed ProtoCalib outshines baselines on multiple node classification datasets.

P050 · **PatentMerit: A Holistic System for In-depth Technology Understanding** [Demo]

Tianxiang Xie (Dalian University of Technology); Jingxuan Wu (Dalian University of Technology); Jiaying Liu (Dalian University of Technology); Junxiang Zhang (The Hong Kong University of Science and Technology (Guangzhou)); Shuo Yu (Dalian University of Technology)

Demand for precise evaluation and in-depth analysis of patent technology value has become increasingly urgent with the accelerated iteration of innovation, while existing patent platforms primarily focus on surface-level bibliographic information, offering limited support for understanding the technological value and evolutionary trajectories of patents. Hence, we develop PatentMerit, a comprehensive system for in-depth analysis of the value of patents, based on large language models, network analysis, and data mining techniques. PatentMerit features five core functionalities, including multidimensional patent classification retrieval, disruptive technology identification, citation network and technology evolution visualization, topic semantic analysis, and automated comprehensive report generation. Its core strength lies in transcending single data element analysis to conduct in-depth exploration, such as accurate technology assessment of individual patents and their technological clusters, which reduces the cognitive and operational burden for R&D personnel. PatentMerit serves as an intelligent decision-making tool for technology trend forecasting, R&D decision-making, and patent strategy planning. The PatentMerit system is accessible via the following link: https://patentmerit.com/.

P051 · **Labadain Chat: A Conversational Agent for the Tetun Language** [Demo]

Gabriel de Jesus (INESC TEC); Sérgio Nunes (INESC TEC)

Large language model (LLM)-based conversational assistants are designed for general-purpose conversation tasks and are primarily optimized for high-resource languages. Although these systems support some low-resource languages (LRLs), their responses often fall short of user expectations. Consequently, speakers of LRLs remain marginalized and unable to fully benefit from advances in LLMs. These challenges underscore

the need for targeted, language-specific solutions that can effectively serve underrepresented language communities. This study presents Labadain Chat, a conversational agent for Tetun, a low-resource language spoken by over 932,000 people in Timor-Leste. We adapt existing LLMs to Tetun using language-specific prompting strategies and report on the system's architecture, features, applications, and utility for the Tetun-speaking community. Results from the user study show a high task success rate for Labadain Chat (91%, with substantial inter-annotator agreement, Cohen's κ=0.67) and high user satisfaction (4.30 out of 5, with Cohen's weighted κ=0.75), demonstrating the effectiveness of language-specific LLM customization for Tetun. Overall, this study provides a practical pathway toward promoting equitable access to AI-powered information services for the Tetun-speaking community and suggests an adaptable methodology that can be applied to other under-resourced languages in similar contexts. The system is publicly available at https://www.labadain.com, with mobile applications for both iOS and Android.

P052 · **Semantic Recall for Vector Search** [Short]

Leonardo Kuffo (CWI); Ioanna Tsakalidou (EPFL); Roberta De Viti (MPI-SWS); Albert Angel (Google); Ji■í Iša (Google); Rastislav Lenhardt (Google)

We introduce Semantic Recall, a novel metric to assess the quality of approximate nearest neighbor search algorithms by considering only semantically relevant objects that are theoretically retrievable via exact nearest neighbor search. Unlike traditional recall, semantic recall does not penalize algorithms for failing to retrieve objects that are semantically irrelevant to the query, even if those objects are among their nearest neighbors. We demonstrate that semantic recall is particularly useful for assessing retrieval quality on queries that have few relevant results among their nearest neighbors---a scenario we uncover to be common within embedding datasets. Additionally, we introduce Tolerant Recall, a proxy metric that approximates semantic recall when semantically relevant objects cannot be identified. We empirically show that our metrics are more effective indicators of retrieval quality, and that optimizing search algorithms for these metrics can lead to improved cost-quality tradeoffs.

P053 · **The Matryoshka Hypencoder** [Short]

Majd Alkawaas (University of Glasgow); Sean MacAvaney (University of Glasgow)

The Hypencoder is a recently-proposed retrieval approach that encodes queries as shallow neural networks ("Q-Nets") that estimate relevance over pre-computed document embeddings. Inspired by Matryoshka Representation Learning, we show that the Hypencoder can be extended to support multiple sizes of Q-Nets, allowing trade-offs between effectiveness and efficiency when deployed. We find that this "Matryoshka Hypencoder" achieves comparable in-domain effectiveness with approximately 7× fewer active parameters in-domain and half as many active parameters out-of-domain, which corresponds to a 1.6–3.4× increase in scoring throughput. This work paves the way for practical deployment of Hypencoders. https://github.com/MajdAlkawaas/hypencoder-paper

P054 · **Load-sensitive Selective Pruning in Dense Retrieval** [Short]

Maria Diana Calugaru (Sapienza University of Rome); Federico Siciliano (National Institute of Geophysics and Volcanology); Francesca Pezzuti (University of Pisa); Nicola Tonellotto (University of Pisa); Fabrizio Silvestri (Sapienza University of Rome)

We introduce a load-aware selective embedding-pruning method to be used in dense retrieval systems. Our method adapts the embedding dimensionality at inference time based on per-query deadlines and the current queue occupancy. Under light load, full-dimensional embeddings maximise effectiveness; as load increases, dimensionality is adaptively reduced to maintain throughput and avoid query drops. Our approach is orthogonal to the specific dimensionality reduction strategy employed and operates independently of the embedding model. We evaluate our approach using both PCA-reduced embeddings and nested Matryoshka embeddings. Empirical results show that our load-aware strategy consistently achieves a better effectiveness–efficiency trade-off than static dimensionality reduction baselines across varying load conditions. As a by-product of this study, our experiments show that under selective pruning, the PCA-based approach consistently outperforms Matryoshka, indicating that specialised multi-representation training is not strictly required for robust load balancing in dense retrieval. Our code is available at: https://github.com/mariadianacalugaru/selective_pruning.

P055 · **PrepRet: Automated Data Preparation Pipeline Selection for Neural Retrieval** [Short]

Jing Chang (Shenzhen University); Chang Liu (Shenzhen University)

Neural retrieval models typically rely on fixed, hand-crafted preprocessing pipelines designed independently of the retrieval task, leading to suboptimal performance that varies across datasets and architectures. We propose PrepRet, a framework that jointly optimizes preprocessing pipeline selection and neural retrieval through differentiable optimization. We formulate preprocessing selection as a differentiable discrete choice problem using Gumbel-Softmax relaxation, enabling end-to-end gradient-based learning over a search space of 700 configurations spanning text cleaning, pre-tokenization, chunking, and augmentation. A hierarchical selection mechanism captures inter-stage dependencies between preprocessing operations. On MS MARCO, PrepRet achieves 0.533 nDCG@10, improving over Grid Search by 7.7\% and Contriever by 11.7\%, while requiring only 1.1 GPU-hours. Zero-shot evaluation on eight BEIR datasets confirms robust cross-domain generalization, with particularly strong gains on scientific and entity-rich domains.

P056 · **Revisiting Poison Pills for Neural Information Retrieval** [Short]

Moriya Menachem (Technion - Israel Institute of Technology); Oren Kurland (Technion - Israel Institute of Technology); Fiana Raiber (Technion - Israel Institute of Technology)

Relevance feedback can substantially improve retrieval effectiveness. However, the performance depends on the specific documents used; some relevant documents, termed poison pills, may even yield worse performance than that attained without feedback. While prior research has focused on lexical feedback methods and the use of a single feedback document, in this work, we expand the definition of poison pills to sets of documents and study both lexical and neural relevance feedback methods. Experiments across multiple datasets and retrieval approaches show that neural feedback methods achieve higher average effectiveness than lexical methods but are more sensitive to the choice of feedback documents. We further observe notable interactions among feedback documents. Poison pills that degrade performance when used alone can combine to form sets that improve performance. In addition, the performance of combining a relevant document with pseudo-relevant documents depends on whether the relevant document is itself a poison pill.

P057 · **RankEvolve: Automating the Discovery of Retrieval Algorithms via LLM-Driven Evolution** [Short]

Jinming Nian (Santa Clara University); Fangchen Li (Independent Researcher); Dae Hoon Park (melten.ai); Yi Fang (Santa Clara University)

Retrieval algorithms like BM25 and query likelihood with Dirichlet smoothing remain strong and efficient first-stage rankers, yet improvements have mostly relied on parameter tuning and human intuition. We investigate whether a large language model, guided by an evaluator and evolutionary search, can automatically discover improved lexical retrieval algorithms. We introduce RankEvolve, a program evolution setup based on AlphaEvolve, in which candidate ranking algorithms are represented as executable code and iteratively mutated, recombined, and selected based on retrieval performance across 12 IR datasets from BEIR and BRIGHT. RankEvolve starts from two seed programs: BM25 and query likelihood with Dirichlet smoothing. The evolved algorithms are novel, effective, and show promising transfer to the full BEIR and BRIGHT benchmarks as well as TREC DL 19 and 20. Our results suggest that evaluator-guided LLM program evolution is a practical path towards automatic discovery of novel ranking algorithms.

P058 · **PeerPrism: Peer Evaluation Expertise vs Review-writing AI** [Resource]

Soroush Sadeghian (Reviewerly); Alireza Daghighfarsoodeh (Reviewerly); Radin Cheraghi (Reviewerly); Sajad Ebrahimi (University of Toronto); Negar Arabzadeh (UC Berkeley); Ebrahim Bagheri (University of Toronto)

Large Language Models (LLMs) are increasingly used in scientific peer review, assisting with drafting, rewriting, and refinement. However, existing peer-review LLM detection methods largely treat authorship as a binary problem (human vs. AI) without accounting for the hybrid nature of modern review workflows. In practice, evaluative ideas and surface realization may originate from different sources, creating a spectrum of human-AI collaboration. To address this, we introduce PeerPrism, a large-scale benchmark of 20,690 peer reviews explicitly designed to disentangle idea provenance from text provenance. We construct controlled generation regimes spanning fully human, fully synthetic, and multiple hybrid transformations. We benchmark state-of-the-art LLM text detection methods on PeerPrism. While several methods achieve high accuracy on the standard binary task, their predictions diverge sharply under hybrid regimes. In particular, when ideas originate from humans but the surface text is AI-generated, detectors frequently disagree and produce contradictory classifications. Our results show that current detection methods conflate surface realization with intellectual contribution. Rather than relying on this binary, authorship must be modeled as a multidimensional construct spanning semantic reasoning and stylistic realization. PeerPrism is the first benchmark evaluating human-AI collaboration in these settings. We release all code, data, prompts, and evaluation scripts to facilitate reproducible research at https://github.com/Reviewerly-Inc/PeerPrism.

P059 · **Failing Forward: Understanding Query Failure in Retrieval, Judgment, and Generation** [Short]

Seyed Mohammad Hosseini (Toronto Metropolitan University); Negar Arabzadeh (University of California, Berkeley); Mohammad Hossein Saliminabi (Toronto Metropolitan University); Dimitrios Androutsos (Toronto Metropolitan University); Morteza Zihayat (Toronto Metropolitan University); Ebrahim Bagheri (University of Toronto)

Modern information retrieval pipelines combine retrieval, LLM-based generation, and LLM-based judgment, and a poor outcome may originate in any of the three stages. Existing work studies these failures in isolation. This paper instead asks whether query difficulty itself transfers across the three stages: are the same queries hard to retrieve, hard to generate for, and hard to judge? Using four years of TREC Deep Learning benchmarks (2019–2022), we define hard-to-retrieve, hard-to-generate, and hard-to-judge query sets under a unified quartile-based operationalization and analyze their overlap, their stability across system configurations, and the linguistic and semantic causes of failure in each task. We find that the three sets overlap only weakly; three-way overlap is at or below the level expected under independence, indicating that difficulty is largely task-conditioned and does not transfer reliably across stages. The overlap structure is nonetheless stable across retrievers, generators, and judging setups, suggesting that task-specific difficulty is driven by query characteristics interacting with each

task's inductive biases rather than by model choice. We further induce a data-driven typology of failure causes and show that conditioning generation on task-relevant difficulty cues yields consistent gains in answer quality.

P060 · **Resources for Automated Evaluation of Assistive RAG Systems that Help Readers with News Trustworthiness Assessment** [Resource]

Dake Zhang (University of Waterloo); Mark D. Smucker (University of Waterloo); Charles L. A. Clarke (University of Waterloo)

Many readers today struggle to assess the trustworthiness of online news because reliable reporting coexists with misinformation. The TREC 2025 DRAGUN (Detection, Retrieval, and Augmented Generation for Understanding News) Track provided a venue for researchers to develop and evaluate assistive RAG systems that support readers' news trustworthiness assessment by producing reader-oriented, well-attributed reports. As the organizers of the DRAGUN track, we describe the resources that we have newly developed to allow for the reuse of the track's tasks. The track had two tasks: (Task 1) Question Generation, producing 10 ranked investigative questions; and (Task 2, the main task) Report Generation, producing a 250-word report grounded in the MS MARCO V2.1 Segmented Corpus. As part of the track's evaluation, we had TREC assessors create importance-weighted rubrics of questions with expected short answers for 30 different news articles. These rubrics represent the information that assessors believe is important for readers to assess an article's trustworthiness. The assessors then used their rubrics to manually judge the participating teams' submitted runs. To make these tasks and their rubrics reusable, we have created an automated process to judge runs not part of the original assessing. We show that our AutoJudge ranks existing runs well compared to the TREC human-assessed evaluation (Kendall's $\tau = 0.678$ for Task 1 and $\tau = 0.872$ for Task 2). These resources enable both the evaluation of RAG systems for assistive news trustworthiness assessment and, with the human evaluation as a benchmark, research on improving automated RAG evaluation.

P061 · **C3Flow: SIMD-Style Concurrent Claude Code Workflow for Scaling Deep Research** [Demo]

Yunfan Gao (Tongji University); Xinyi Huang (Fudan University); Yijie Zhong (Tongji University); Zhidong Fan (Ant Group); Zhengke Gui (Ant Group); Lei Liang (Ant Group); Yun Xiong (Fudan University); Haofen Wang (Tongji University)

Deep research is a retrieval-intensive task that requires iteratively retrieving evidence, reading across sources, and synthesizing source-grounded outputs. In practice, real-world deep research applications are of high workloads that require generating massive reports or conducting large-scale literature surveys. Such applications increasingly require batch processing capabilities, which are missing from traditional chat-oriented agents, limiting throughput when processing large volumes of structurally similar jobs. We introduce C3Flow (Concurrent Claude Code Workflow), a framework that transforms Claude Code from an interactive assistant into a SIMD-style (Single Instruction, Multiple Data) concurrent compute engine. C³Flow treats each agent instance as an isolated, schedulable unit capable of handling declarative multi-step tasks, multi-model routing, and comprehensive trajectory logging. On BrowseComp-zh, C³Flow improves pass@1 from 48.44% to 61.59% and pass@3 from 70.24% to 77.51% compared to standard function calling, while reducing average latency. For multi-hop fact verification, C3Flow achieves a 5.9 speedup over human annotators while maintaining 87.5% accuracy, demonstrating its effectiveness for production-scale deep research pipelines. Code is available at~ https://github.com/RAGenius/C3Flow.

P062 · **Optimal Re-Ranking Depth** [Short]

Siqing Huo (University of Waterloo); Andrew Parry (University of Glasgow); Debasis Ganguly (University of Glasgow); Charles L. A. Clarke (University of Waterloo)

Second-stage neural rankers are commonly applied with a fixed re-ranking depth, assuming that retrieval effectiveness saturates as depth increases. Prior research has questioned the assumption that increasing the re-ranking depth yields linear performance gains, further suggesting that optimal re-ranking depth varies considerably from query to query. In the past, studying such phenomena was methodologically difficult given the scale of manual annotation required. With the advent of LLM-based relevance judgments we can now more easily undertake such studies, in this case to pinpoint the optimal re-ranking depth on a per-query basis. Using dense LLM-based relevance judgments over a typical re-ranking pipeline, we show that many queries exhibit a well-defined optimal re-ranking depth, beyond which effectiveness stagnates or degrades. We formulate re-ranking depth as a query-specific property and study whether it can be predicted a priori from first-stage retrieval characteristics. Through a large-scale analysis, we find that most standard query performance prediction (QPP) methods are ineffective for this task. In contrast, a predictor derived from LLM-assessed first-stage ranking quality, which we term IR-DCG@10, can reduce the average re-ranking depth by up to a factor of 3 while preserving overall effectiveness, depending on the specific first- and second-stage rankers used. Under oracle selection of optimal depths, we further show that retrieval effectiveness can improve by more than 7% while reducing the average re-ranking depth by a factor of 5 on the MSMARCO DEV collection. Given our promising preliminary findings, we would encourage the use of automatic judgments to facilitate research otherwise infeasible under manual annotation. Towards this point, we release relevance judgments, our codebase, and experimental artefacts to support reproducibility and further research.

P063 · **Auditing Query Drift: Do Users Actually Benefit from Pseudo-Relevance Feedback?** [Short]

Zeyan Liang (University of Glasgow); Graham McDonald (University of Glasgow); Iadh Ounis (University of Glasgow)

Pseudo-Relevance Feedback (PRF) aims to improve performance, but it can suffer from query drift, creating asymmetric impacts across queries. Current Selective PRF (sPRF) strategies attempt to mitigate this by predicting when PRF will help, but they rely on offline metrics that may not reflect actual user preferences. To bridge this gap, we propose a participatory auditing strategy that evaluates PRF's real-world impact through natural user interactions. Specifically, we design a 3×3 grid interface using Latin square ordering to mitigate position bias, and deploy Team Draft Interleaving (TDI) to collect implicit user preferences without disrupting normal search behaviour. We conduct two user studies: in our first user study, we find that only 20.9% of queries benefit from PRF, while 25.6% result in a degraded user experience. Our second study demonstrates that avoiding harmful PRF yields a 68.5% relative improvement, nearly twice the 37.0% gained from amplifying beneficial PRF, highlighting that harm avoidance is more valuable than benefit amplification. Our findings highlight the necessity for reliable sPRF and suggest that future sPRF research could benefit from incorporating interaction-aware participatory auditing when developing sPRF approaches.

P064 · **When RAG Disagrees: Detecting Latent Epistemic Conflict via Logit Interactions** [Short]

Saisab Sadhu (Indian Institute of Science Education and Research Bhopal); Dwaipayan Roy (Indian Institute of Science Education and Research Kolkata); Tanmay Basu (Indian Institute of Science Education and Research Bhopal)

Retrieval-Augmented Generation (RAG) is frequently constrained by knowledge conflicts where external evidence contradicts the internal parametric memory of a model. Current approaches to predicting these conflicts rely on computationally expensive supervised probing or black-box adjudication, introducing significant latency. We propose a training-free mechanistic predictor, the Interaction Score, derived from the parametric confidence of the model and its semantic alignment with the retrieved context. Using high-resolution logit analysis on 70B-parameter models, we show that this score serves as a statistically robust, training-free predictor of internal epistemic conflict. Across Llama, Qwen, and Mistral families, we further observe that instruction tuning (RLHF) increases internal predictability while decoupling it from textual behavior. This creates a state of latent dissonance — a condition where models generate compliant text despite harboring significant internal tension detectable only in logit space. We introduce NQ-Temporal-2K, a curated diagnostic benchmark for real-world update scenarios. Our method enables conflict prediction (binary classification of whether retrieval helps or hurts) with 25ms of overhead, offering a scalable gatekeeper for production-grade retrieval systems.

P065 · **Context Convergence Improves Answering Inferential Questions** [Short]

Jamshid Mozafari (University of Innsbruck); Bhawna Piryani (University of Innsbruck); Adam Jatowt (University of Innsbruck)

While Large Language Models (LLMs) are widely used in open-domain Question Answering (QA), their ability to handle inferential questions—where answers must be derived rather than directly retrieved—remains still underexplored. This study investigates how the structure and quality of passages influence LLM performance on such questions. We focus on convergence, a measure of how effectively sentences (hints) eliminate incorrect answers, as a criterion for constructing passages. Using subsets of the TriviaHG dataset, we form passages by combining sentences with varying convergence levels and evaluate six LLMs of different sizes and architectures. Our results show that passages built from higher-convergence sentences lead to substantially better answer accuracy than those selected by cosine similarity, indicating that convergence captures meaningful relevance for inferential reasoning. Additionally, ordering sentences by descending convergence slightly improves performance, suggesting that LLMs tend to prioritize earlier, information-rich cues. These findings highlight convergence as a practical signal for guiding passage construction and analyzing inferential reasoning behavior in LLMs.

P066 · **Reward Shaping for Robust Refusal in Small Language Models for Retrieval-Augmented Question Answering** [Short]

Thilina C. Rajapakse (Eindhoven University of Technology); Maarten de Rijke (University of Amsterdam)

We focus on smaller open-source LMs (2–7B parameters), which are attractive for practical deployment due to their lower computational cost and greater accessibility than frontier-scale models. We show that instruction-tuned models generate answers even when explicitly prompted to refuse when the answer is not supported by the documents. In the presence of distractor documents, instruction-tuned models demonstrate inconsistent performance, with answer accuracy metrics deteriorating in most cases. To mitigate this behavior, we introduce Reward Shaping for Refusal and Reasoning (RSRR), a reinforcement learning framework that teaches LMs to reason step-by-step over multiple documents and to refuse to answer when evidence is insufficient. Models trained with RSRR achieve substantial improvements in robustness to distractor documents and in correct refusal accuracy, with gains of 39.8% and 43.3%, respectively. We release code and data to reproduce all results. https://github.com/ThilinaRajapakse/rsrr

P067 · **AdversarialCoT: Single-Document Retrieval Poisoning for LLM Reasoning** [Short]

Hongru Song (State Key Laboratory of AI Safety); Yu-An Liu (State Key Laboratory of AI Safety); Ruqing Zhang (State Key Laboratory of AI Safety); Jiafeng Guo (State Key Laboratory of AI Safety); Maarten de Rijke (University of Amsterdam); Yixing Fan (State Key Laboratory of AI Safety); Xueqi Cheng (State Key Laboratory of AI Safety)
Retrieval-augmented generation (RAG) enhances large language model (LLM) reasoning by retrieving external documents, but also opens up new attack surfaces. We study knowledge-base poisoning attacks in RAG, where an attacker injects malicious content into the retrieval corpus, which is then surfaced by the retriever and consumed by the LLM during reasoning. Unlike prior work that floods the corpus with poisoned documents, we propose AdversarialCoT, a query-specific attack that poisons only a single document in the corpus. AdversarialCoT first extracts the target LLM's reasoning framework to guide the construction of an initial adversarial chain-of-thought. The adversarial document is iteratively refined through interactions with the LLM, progressively exposing and exploiting critical reasoning vulnerabilities. Experiments on benchmark LLMs show that a single adversarial document can significantly degrade reasoning accuracy, revealing subtle yet impactful weaknesses. Our study exposes security risks in RAG systems and provides actionable insights for designing more robust LLM reasoning pipelines.

P068 · ERASE - A Real-World Aligned Benchmark for Unlearning in Recommender Systems [Resource]

Pierre Sicco Lubitzsch (BIFOLD & TU Berlin); Maarten de Rijke (University of Amsterdam); Sebastian Schelter (BIFOLD & TU Berlin)
Machine unlearning (MU) enables the removal of selected training data from models to address privacy compliance, security, and liability issues in recommender systems. Existing MU benchmarks poorly reflect real-world recommender settings: they focus primarily on collaborative filtering, assume unrealistically large deletion requests, and overlook practical constraints such as sequential unlearning and efficiency. We present ERASE, a large-scale benchmark for MU in recommender systems designed to align with real-world usage. ERASE spans three core tasks---collaborative filtering, session-based recommendation, and next-basket recommendation---and includes unlearning realistic scenarios such as sequentially removing sensitive interactions or spam. The benchmark covers seven unlearning algorithms, including general-purpose and recommender-specific methods, across nine public datasets and nine state-of-the-art models, producing over 600 GB of reusable artifacts, such as extensive logs and over a thousand model checkpoints. The artifacts that we release enable systematic analyses of where current unlearning methods succeed and where they fail. ERASE shows that approximate unlearning matches retraining in some settings, but robustness varies widely across datasets and architectures. Repeated unlearning exposes weaknesses in general-purpose methods, especially for attention-based and recurrent models, while recommender-specific approaches behave more reliably. ERASE offers an empirical basis for the community to assess, drive, and track progress toward practical MU in recommender systems.

P069 · One-Pass Decoding for Generative Recommendation with WFST-Constrained A* Search [Short]

Puji Wang (University of Chinese Academy of Sciences); Yingchen Zhang (State Key Laboratory of AI Safety); Ruqing Zhang (State Key Laboratory of AI Safety); Jiafeng Guo (State Key Laboratory of AI Safety); Xueqi Cheng (State Key Laboratory of AI Safety)
Generative recommendation (GR) represents items as discrete semantic identifiers (SIDs) and performs next-item prediction via identifier sequence generation. Most existing methods perform prefix-constrained decoding over a trie, where hard structural constraints lead to locally greedy decisions. An early suboptimal prefix irrevocably restricts subsequent decoding, causing error propagation and suboptimal retrieval. In this work, we propose OneGR, a novel generative recommendation framework that formulates SID prediction as a one-pass, structure-constrained decoding problem. OneGR computes globally consistent, position-wise SID scores in a single neural forward pass, eliminating iterative hypothesis expansion and repeated inference. To address the extreme sparsity of valid SIDs, we encode the space of valid SIDs as a deterministic weighted finite-state transducer and perform A^* search with the one-pass scores as additive path costs. This structured decoding strategy explores only reachable valid prefixes while guaranteeing the optimal SID prediction. Experiments on multiple benchmarks show that OneGR consistently outperforms strong baselines.

P070 · Mitigating Collaborative Semantic ID Staleness in Generative Retrieval [Short]

Vladimir Baikalov (AI VK); Iskander Bagautdinov (ITMO University); Sergey Muravyov (ITMO University)
Generative retrieval with Semantic IDs (SIDs) assigns each item a discrete identifier and treats retrieval as a sequence generation problem rather than a nearest-neighbor search. While content-only SIDs are stable, they do not take into account user-item interaction patterns, so recent systems construct interaction-informed SIDs. However, as interaction patterns drift over time, these identifiers become stale, i.e., their collaborative semantics no longer match recent logs. Prior work typically assumes a fixed SID vocabulary during fine-tuning, or treats SID refresh as a full rebuild that requires retraining. However, SID staleness under temporal drift is rarely analyzed explicitly. To bridge this gap, we study SID staleness under strict chronological evaluation and propose a lightweight, model-agnostic SID alignment update. Given refreshed SIDs derived from recent logs, we align them to the existing SID vocabulary so the retriever checkpoint remains compatible, enabling standard warm-start fine-tuning without a full rebuild-and-retrain pipeline. Across three public benchmarks, our update consistently improves Recall@K and nDCG@K at high cutoffs over naive fine-tuning with stale SIDs and reduces retriever-training compute by approximately 8-9 times compared to full retraining.

P071 · RoTE: Coarse-to-Fine Multi-Level Rotary Time Embedding for Sequential Recommendation [Short]

Haolin Zhang (Tsinghua University); Longtao Xiao (Huazhong University of Science and Technology); Guohao Cai (Huawei Noah's Ark Lab); Ruixuan Li (Huazhong University of Science and Technology); Xiu Li (Tsinghua University)
Sequential recommendation models have been widely adopted for modeling user behavior. Existing approaches typically construct user interaction sequences by sorting items according to timestamps and then model user preferences from historical behaviors. While effective, such a process only considers the order of temporal information but overlooks the actual time spans between interactions, resulting in a coarse representation of users' temporal dynamics and limiting the model's ability to capture long-term and short-term interest evolution. To address this limitation, we propose RoTE, a novel multi-level temporal embedding module that explicitly models time span information in sequential recommendation. RoTE decomposes each interaction timestamp into multiple temporal granularities, ranging from coarse to fine, and incorporates the resulting temporal representations into item embeddings. This design enables models to capture heterogeneous temporal patterns and better perceive temporal distances among user interactions during sequence modeling. RoTE is a lightweight, plug-and-play module that can be seamlessly integrated into existing Transformer-based sequential recommendation models without modifying their backbone architectures. We apply RoTE to several representative models and conduct extensive experiments on three public benchmarks. Experimental results demonstrate that RoTE consistently enhances the corresponding backbone models, achieving up to a 20.11% improvement in NDCG@5, which confirms the effectiveness and generality of the proposed approach. Our code is available at https://github.com/XiaoLongtaoo/RoTE.

P072 · Revisiting the Role of Learned Attention Weighting in SASRec [Short]

Keito Kozaki (Hokkaido University); Keigo Sakurai (Hokkaido University); Ren Togo (Hokkaido University); Takahiro Ogawa (Hokkaido University); Miki Haseyama (Hokkaido University)
Causal self-attention models such as SASRec are widely used in sequential recommendation, where learned attention weights are often assumed to provide crucial importance weighting over past interactions. Yet it is unclear when predictive performance truly depends on such non-uniform weighting. We study a controlled SASRec variant that replaces learned attention weights with uniform aggregation and is trained under an otherwise identical block structure and training recipe. Across fourteen benchmark datasets, this modification often yields performance comparable to the original model, with clear dataset-dependent exceptions. To explain this heterogeneity, we introduce a stage-wise norm-based decomposition that quantifies self-preserving vs. cross-position mixing within attention blocks. Across datasets, we find distinct regimes: low mixing yields robustness to uniformization; higher mixing tends to coincide with sensitivity, while some datasets exhibit substantial mixing without dependence on learned weighting. Our results provide a practical diagnostic for identifying when attention weighting is functionally utilized in sequential recommendation. The code is available at: https://github.com/keito0329/revisiting-sasrec.

P073 · SplitLight: An Exploratory Toolkit for Recommender Systems Datasets and Splits [Resource]

Anna Volodkevich (SB AI Lab); Dmitry Anikin (Applied AI Institute); Danil Gusak (AXXX); Anton Klenitskiy (SB AI Lab); Evgeny Frolov (AXXX); Alexey Vasilev (SB AI Lab)
Offline evaluation of recommender systems is often affected by hidden, under-documented choices in data preparation. Seemingly minor decisions in filtering, handling repeats, cold-start treatment, and splitting strategy design can substantially reorder model rankings and undermine reproducibility and cross-paper comparability. In this paper, we introduce SplitLight, an open-source exploratory toolkit that enables researchers and practitioners designing preprocessing and splitting pipelines or reviewing external artifacts to make these decisions measurable, comparable, and reportable. Given an interaction log and derived split subsets, SplitLight analyzes core and temporal dataset statistics, characterizes repeat consumption patterns and timestamp anomalies, and diagnoses split validity, including temporal leakage, cold-user/item exposure, and distribution shifts. SplitLight further allows side-by-side comparison of alternative splitting strategies through comprehensive aggregated summaries and interactive visualizations. Delivered as both a Python toolkit and an interactive no-code interface, SplitLight produces audit summaries that justify evaluation protocols and support transparent, reliable, and comparable experimentation in recommender systems research and industry. Code is available at https://github.com/monkey0head/SplitLight.

P074 · Differentiable Dual Anchor Negative Sampling for Graph-based Recommendation [Short]

Xi Wu (Yanshan University); Wenzhe Zhang (Yanshan University); Kexin Zhao (Yanshan University); Jiquan Peng (Yanshan University); Jibing Gong (Yanshan University)
Negative sampling plays a pivotal role in training recommendation systems with implicit feedback, where the effectiveness of negatives directly impacts model convergence and recommendation quality. The key challenge is to efficiently mine high-quality hard negatives from the massive item space. Existing strategies typically rely on a single user perspective as the sampling

anchor and use discrete $\arg\max$ operations to select negatives. However, the single-anchor design introduces noisy negatives, and the discrete hard selection prevents end-to-end optimization. To address these limitations, we propose a differentiable dual-anchor negative sampling framework for graph-based recommendation. Our framework introduces a differentiable cross-hop sampling mechanism based on the Gumbel-Softmax trick, enabling hard negative selection while preserving gradient flow. Furthermore, we incorporate both the user and the corresponding positive item as complementary sampling anchors to improve the quality and stability of negative samples. Extensive experiments on three benchmark datasets demonstrate that our approach consistently improves recommendation performance.

P075 · **SimEval-IR: A Unified Toolkit and Benchmark Suite for Evaluating User Simulators and Search Sessions** [Resource]

Saber Zerhoudi (University of Passau)

User simulators are increasingly central to interactive information retrieval, yet the community lacks standardized evaluation tools. Simulators serve two objectives, behavioral realism (matching real user behavior) and tester reliability (producing valid system rankings), and these are often conflated despite being distinct and sometimes conflicting. We present SimEval-IR, an open-source toolkit and benchmark suite that makes this distinction measurable. SimEval-IR provides: (1) a canonical session schema unifying session search and conversational interactions, with validated dataset adapters and explicit loss accounting; (2) three executable benchmarks covering behavioral realism, tester reliability with RATE-style estimation, and an analysis linking the two; and (3) baseline results across four real datasets in two languages and four simulator families. Our key finding: the classifier-discriminator "human-likeness" check, the dominant realism test in the literature, has essentially no pooled predictive power for system-ranking validity (r= + 0.09, n= 48), while marginal click-depth distance and Fré chet distance over session embeddings give a much stronger signal (|r|= 0.43 and 0.40, p■eq 0.005). SimEval-IR is released with all configurations and scripts to reproduce the reported analysis.

P076 · **WebFAQ 2.0: A Multilingual QA Dataset with Mined Hard Negatives for Dense Retrieval** [Resource]

Michael Dinzinger (University of Passau); Laura Caspari (University of Passau); Ali Salman (University of Passau); Irvin Topi (University of Passau); Jelena Mitrovi■ (University of Passau); Michael Granitzer (University of Passau)

We introduce WebFAQ 2.0, a new version of the WebFAQ dataset, containing 198 million FAQ-based natural question-answer pairs across 108 languages. Compared to the previous version, it significantly expands multilingual coverage and the number of bilingual aligned QA pairs to over 14.3M, making it the largest FAQ-based resource. Unlike the original release, WebFAQ 2.0 uses a novel data collection strategy that directly crawls and extracts relevant web content, resulting in a substantially more diverse and multilingual dataset with richer context through page titles and descriptions. In response to community feedback, we also release a hard negatives dataset for training dense retrievers, with 1.25M queries across 20 languages. These hard negatives were mined using a two-stage retrieval pipeline and include cross-encoder scores for 200 negatives per query. We further show how this resource enables two primary fine-tuning strategies for dense retrievers: Contrastive Learning with MultipleNegativesRanking loss, and Knowledge Distillation with MarginMSE loss. WebFAQ 2.0 is not a static resource but part of a long-term effort. Since late 2025, structured FAQs are being regularly released through the Open Web Index, enabling continuous expansion and refinement. We publish the datasets and training scripts to facilitate further research in multilingual and cross-lingual IR. The dataset itself and all related resources are publicly available on GitHub and HuggingFace.

P077 · **Who Is Shopping With You? How Persona Design Shapes Cognitive and Social Engagement in AI Shopping Agents** [Short]

Hyungwoo Song (Seoul National University); Kyusik Kim (Seoul National University); Hyeonseok Jeon (Seoul National University); Minjeong Shin (Seoul National University); Bongwon Suh (Seoul National University)

Conversational shopping agents powered by large language models are increasingly used for online product exploration, yet the role of interaction style in shaping shopping behavior and user experience remains underexplored in shopping IR. To address the gap, we conducted two studies in experience-goods domains. Study 1 involved 24 participants and compared a neutral conversational agent with a traditional product search interface, confirming functional adequacy and identifying two unmet needs, self-reflective preference structuring and socially grounded relational guidance. Study 2 involved 30 participants and evaluated two personas derived from these needs, \textit{Self-Mirroring} and \textit{Relational Peer}, against the same neutral agent in a within-subjects design with information availability held constant. \textit{Self-Mirroring} increased critical thinking scores and sustained follow-up questioning of retrieved content, whereas \textit{Relational Peer} increased social presence while reducing explicit verification behaviors such as comparing alternatives and checking conditions. The discussion outlines implications for AI shopping agents that adapt interaction style to decision context, balancing efficient exploration with user-led evaluation.

P078 · **CoverageBench: Evaluating Information Coverage across Tasks and Domains** [Resource]

Saron Samuel (Johns Hopkins University); Andrew Yates (Johns Hopkins University); Dawn Lawrie (Johns Hopkins University); Ian Soboroff (National Institute of Standards and Technology); Trevor Adriaanse (Johns Hopkins University); Benjamin Van Durme (Johns Hopkins University); Eugene Yang (Johns Hopkins University)

We wish to measure the information coverage of an ad hoc retrieval algorithm, that is, how much of the range of available relevant information is covered by the search results.Information coverage is a central aspect for retrieval, especially when the retrieval system is integrated with generative models in a retrieval-augmented generation (RAG) system.The classic metrics for ad hoc retrieval, precision and recall, reward a system as more relevant documents are retrieved. However, since relevance in ad hoc test collections is defined for a document without any relation to other documents that might contain the same information, high recall is sufficient but not necessary to ensure coverage. The same is true for other metrics such as rank-biased precision (RBP), normalized discounted cumulative gain (nDCG), and mean average precision (MAP). Test collections developed around the notion of diversity ranking in web search incorporate multiple aspects that support a concept of coverage in the web domain. In this work, we construct a benchmark, CoverageBench, for evaluating information coverage made from existing collections. This suite offers researchers a unified testbed spanning multiple genres and tasks. All topics, nuggets, relevance labels, and baseline rankings are released on Hugging Face Datasets, along with instructions for accessing the publicly available document collections.

P079 · **Too Many Questions: Deriving Concise and Effective Nugget Banks** [Short]

Laura Dietz (University of New Hampshire); Naghmeh Farzi (University of New Hampshire); Eugene Yang (Johns Hopkins University); Dawn Lawrie (Johns Hopkins University)

Nugget-based LLM judges evaluate Retrieval-Augmented Generation (RAG) systems using a bank of questions that capture the key facts and criteria an answer should address. These nugget banks are typically constructed through a combination of human input and LLM generation. System outputs are graded by how well they cover the nuggets.For cost and scalability reasons, the nugget bank should be small. However, a major limitation of current nugget generation approaches is that many questions are overly generic and fail to discriminate between top-performing RAG systems. Grounding nuggets in system responses or source documents can increase specificity, but typically leads to an explosion in the number of questions. Since every response is graded for every nugget question, a higher number of questions directly increases the amount of LLM prompts and/or tokens required, contributing to costs.Inspired by preference-based evaluation, we derive differential nuggets from winner-loser passage pairs, focusing on information that captures differences in topicality, level of detail, and evidential support between responses under an automatic preference judge. We examine how these contrastive signals can be leveraged to construct nugget banks that are both compact and discriminative, enabling reliable separation among top-performing RAG systems.

P080 · **Auto-ARGUE: LLM-Based Report Generation Evaluation** [Demo]

William Walden (Johns Hopkins University); Marc Mason (Solerity); Orion Weller (Johns Hopkins University); Laura Dietz (University of New Hampshire); John M. Conroy (IDA Center for Computing Services); Neil Molino (IDA Center for Computing Services); Hannah Recknor (Johns Hopkins University); Bryan Li (Google); Gabrielle Kaili-May Liu (Yale University); Yu Hou (University of Maryland); Dawn Lawrie (Johns Hopkins University); James Mayfield (Johns Hopkins University); Eugene Yang (Johns Hopkins University)

Generation of citation-backed reports is a primary use case for retrieval-augmented generation (RAG) systems. While open-source evaluation tools exist for various RAG tasks, tools designed for report generation are lacking. Accordingly, we introduce Auto-ARGUE, a robust LLM-based implementation of the recently proposed ARGUE framework for report generation evaluation. We present analysis of Auto-ARGUE on the report generation pilot task from the TREC 2024 NeuCLIR track and on two tasks from the TREC 2024 RAG track, showing good system-level correlations with human judgments. Additionally, we release ARGUE-viz, a web app for visualization and fine-grained analysis of Auto-ARGUE judgments and scores.1

P081 · **RecNextEval: A Reference Implementation for Temporal Next-Batch Recommendation Evaluation** [Demo]

Tze-Kean Ng (Nanyang Technological University); Joshua Teng-Khing Khoo (Nanyang Technological University); Aixin Sun (Nanyang Technological University)

A good number of toolkits have been developed in Recommender Systems (RecSys) research to promote fair evaluation and reproducibility. However, recent critical examinations of RecSys evaluation protocols have raised concerns regarding the validity of existing evaluation pipelines. In this demonstration, we present RecNextEval, a reference implementation of an evaluation framework specifically designed for next-batch recommendation. RecNextEval utilizes a time-window data split to ensure models are evaluated along a global timeline, effectively minimizing data leakage. Our implementation highlights the inherent complexities of RecSys evaluation and encourages a shift toward model development that more accurately simulates production environments. The RecNextEval library and its accompanying GUI interface are open-source and publicly accessible.

P082 · **Better than Dense? Investigating the Natural Backward Compatibility of Learned Sparse Representations** [Short]

Jingfen Qiao (University of Amsterdam); Gabrielle Poerwawinata (University of Amsterdam); Thong Nguyen (University of Amsterdam); Jia-Huei Ju (University of Amsterdam); Eugene Yang (Johns Hopkins University); Evangelos Kanoulas (University of Amsterdam); Andrew Yates (Johns Hopkins University)

Advancements in retrieval models necessitate re-indexing, a computationally expensive process for large-scale production environments. While updating only the query encoder and continuing to use the old index could be a promising middle ground, dense retrieval systems suffer severe performance drops in this setting. We investigate whether Learned Sparse Retrieval (LSR) can mitigate this backward compatibility issue, as its lexical matching may provide a stable term-based anchor to preserve compatibility across model versions. Experiments on BEIR and the streaming settings of LoTTE show that upgrading only the query encoder causes only a small effectiveness drop in LSR when no mitigation applied, whereas dense retrieval fails severely. We explore lightweight query adaptation methods including ranking fusion, representation fusion, and minimal-training adapters to further improve compatibility. These approaches significantly improve backward compatibility on BEIR and effectively reduce performance loss in streaming retrieval. Code: https://github.com/JingfenQiao/LSR-BC.git

P083 · **R&F-Inventory: A Large-Scale Dataset for Monotonic Inventory Estimation in Reach and Frequency Advertising** [Resource]

Yunshan Peng (Kuaishou Technology); Ji Wu (Kuaishou Technology); Wentao Bai (Kuaishou Technology); Yunke Bai (Kuaishou Technology); Jinan Pang (Kuaishou Technology); Wenzheng Shu (Kuaishou Technology); Yanxiang Zeng (Kuaishou Technology); Xialong Liu (Kuaishou Technology); Peng Jiang (Kuaishou Technology)

Reach and Frequency (R&F) contract advertising is an important form of widely used brand advertising. Unlike performance advertising, R&F contracts emphasize controllable delivery of UV and PV under given targeting, scheduling, and frequency control constraints. In practical systems, advertisers typically need to view the UV, PV change curves at different budget levels in real time when creating an R&F contract. However, most existing publicly available advertising datasets are based on independent samples, lacking a characterization of the core structure of the "budget-performance curve" (including UV and PV) in R&F contracts.This paper proposes and releases a large-scale R&F contract inventory estimation dataset. This dataset uses the R&F contract context consisting of "targeting-scheduling-frequency control" as the basic context, providing observations of UV and PV corresponding to multiple budget points within the same context, thus forming a complete budget-performance curve. The dataset explicitly includes a time-window-based frequency control mechanism (e.g., "no more than 3 times within 5 days") and naturally satisfies the monotonicity and diminishing marginal returns characteristics in the budget and scheduling dimensions. We further derive the theoretical maximum exposure ceiling and use it as a consistency check to evaluate data quality and the feasibility of model predictions. Using this data set, this paper defines two standardized benchmark tasks: single-point performance prediction and reconstruction of budget-performance curves, and provides a set of reproducible baseline methods and evaluation protocols. This dataset can support systematic research on problems such as structural constraint learning, monotonic regression, curve consistency modeling, and R&F contract planning.The code for our experiments can be found at https://github.com/pengyunshan/RF-Inventory.

P084 · **Counterfactual Multi-task Learning for Delayed Conversion Modeling in E-commerce Sales Pre-Promotion** [Short]

Xin Song (Alibaba Group); Kaiyuan Li (Alibaba Group); Jinxin Hu (Alibaba Group)

Sales promotions, as short-term incentives to stimulate product purchases, play a pivotal role in modern e-commerce marketing strategies. During promotional events, user behavior patterns exhibit distinct characteristics compared to regular periods. Notably, in the pre-promotion phase, users typically engage in product search and browsing without immediate purchases, often adding items to carts in anticipation of promotional discounts. This behavior leads to delayed conversions, resulting in significantly lower conversion rates (CVR) before the promotion day.Although existing research has made progress in CVR prediction for promotion days using historical data, it largely overlooks the critical pre-promotion period. Although delayed feedback modeling has been extensively studied, current approaches fail to account for the unique distribution shifts in conversion behavior before promotional events, where delayed conversions predominantly occur on the promotion day rather than over continuous time windows.To address these limitations, we propose the Counterfactual Multi-task Delayed Conversion Model (CM-DCM), which leverages historical pre-promotion data to enhance CVR prediction for both delayed and direct conversions. Our model incorporates three key innovations: (i) A multi-task architecture that jointly models direct and delayed conversions using historical pre-promotion data; (ii) A personalized user behavior gating module to mitigate data sparsity issues during brief pre-promotion periods; (iii) A counterfactual causal approach to model the transition probability from add-to-cart (ATC) to delayed conversion.Extensive experiments demonstrate that CM-DCM outperforms state-of-the-art delayed CVR models in pre-promotion scenarios. Online A/B tests during major promotional events showed significant improvements in advertising revenue, delayed conversion GMV, and overall GMV, validating the effectiveness of our approach.

P085 · **SCORE-RAG: Self-Correcting Exploration-Exploitation Retrieval for Multi-hop Question Answering** [Short]

Shu Zhou (Nanjing University); Rui Ling (Nanjing University); Junan Chen (Nanjing University); Tao Fan (Nanjing University of Finance and Economics); Hao Wang (Nanjing University)

Retrieval-augmented generation (RAG) has emerged as a promising paradigm to enhance Large Language Models (LLMs) with external knowledge, effectively mitigating hallucinations and broadening the model's knowledge coverage. Despite recent advances, existing RAG methods fundamentally assume static query understanding, where the query is interpreted once before retrieval. This assumption proves inadequate for multi-hop questions, where comprehending the query itself often requires retrieval support, creating a chicken-and-egg dilemma between query understanding and information retrieval. To address this challenge, we propose SCORE-RAG Self-COrrecting Exploration-Exploitation REtrieval, a novel framework inspired by the explore-exploit paradigm in decision theory. SCORE-RAG reformulates multi-hop RAG as a two-phase adaptive process: exploration for dynamic query understanding, followed by exploitation for precise evidence gathering. Specifically, SCORE-RAG first performs exploratory retrieval with multi-perspective queries to resolve ambiguities and discover key entities and relations, then conducts targeted exploitation retrieval guided by the refined understanding to construct coherent evidence chains, and finally applies self-correction mechanisms to verify consistency and repair potential errors. Through this integrated approach, SCORE-RAG enables adaptive query comprehension, reduces error accumulation via self-verification, and produces interpretable reasoning chains for accurate answer generation. Extensive experiments on HotPotQA and 2WikiMultihopQA demonstrate that SCORE-RAG significantly outperforms existing state-of-the-art RAG frameworks, achieving substantial improvements particularly on complex multi-hop questions requiring deep reasoning.

P086 · **Why Knowledge Distillation Fails to Scale in Neural Retrieval** [Short]

Shu Zhou (Nanjing University); Rui Ling (Nanjing University); Junan Chen (Nanjing University); Tao Fan (Nanjing University of Finance and Economics); Hao Wang (Nanjing University)

Knowledge distillation (KD) from cross-encoder teachers is a widely adopted technique for training effective neural retrieval models. However, recent studies have revealed a puzzling phenomenon: while retrieval models trained with contrastive loss (CL) exhibit clear scaling behavior with larger language models, KD-trained models show minimal performance gains as model size increases from 1B to 8B parameters. The underlying cause of this scaling failure remains unexplored. In this work, we hypothesize that the teacher model's capacity acts as an information bottleneck, limiting how much large student models can learn. To test this hypothesis, we conduct systematic experiments using decoder-only LLMs (Llama-3: 1B, 3B, 8B) as student retrievers and cross-encoder teachers ranging from 66M to 3B parameters. Our experiments on MSMARCO and BEIR benchmarks reveal that: (1) Small teachers severely constrain student scaling, with 1B, 3B, and 8B students performing nearly identically. (2) Larger teachers progressively restore scaling behavior, enabling significant performance gains at the 8B scale. (3) A teacher-to-student parameter ratio above a critical threshold appears necessary for effective knowledge transfer. Our findings provide practical guidance for selecting appropriate teacher models when training large-scale neural retrievers with knowledge distillation.

P087 · **Understanding DNNs in Feature Interaction Models: A Dimensional Collapse Perspective** [Short]

Jiancheng Wang (University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence); Mingjia Yin (University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence); Hao Wang (University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence); Enhong Chen (University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence)

DNNs have gained widespread adoption in feature interaction recommendation models. However, there has been a longstanding debate on their roles. On one hand, some works claim that DNNs possess the ability to implicitly capture high-order feature interactions. Conversely, recent studies have highlighted the limitations of DNNs in effectively learning dot products, specifically second-order interactions, let alone higher-order interactions. In this paper, we present a novel perspective to understand the effectiveness of DNNs: their impact on the dimensional robustness of the representations. In particular, we conduct extensive experiments involving both parallel DNNs and stacked DNNs. Our evaluation encompasses an overall study of complete DNN on two feature interaction models, alongside a fine-grained ablation analysis of components within DNNs. Experimental results demonstrate that both parallel and stacked DNNs can effectively mitigate the dimensional collapse of embeddings. Furthermore, a gradient-based theoretical analysis, supported by empirical evidence, uncovers the underlying mechanisms of dimensional collapse. The code is accessible for reproduction.
https://github.com/USTC-StarTeam/Dimensional-Collapse-Analysis

P088 · **Sparse Contrastive Learning for Content-Based Cold Item Recommendation** [Short]

Gregor Meehan (Queen Mary University of London); Johan Pauwels (Queen Mary University of London)

Item cold-start is a pervasive challenge for collaborative filtering (CF) recommender systems. Existing methods often train cold-start models by

mapping auxiliary item content, such as images or text descriptions, into the embedding space of a CF model. However, such approaches can be limited by the fundamental information gap between CF signals and content features. In this work, we propose to avoid this limitation with purely content-based modeling of cold items, i.e. without alignment with CF user or item embeddings. We instead frame cold-start prediction in terms of item-item similarity, training a content encoder to project into a latent space where similarity correlates with user preferences. We define our training objective as a sparse generalization of sampled softmax loss with the α-entmax family of activation functions, which allows for sharper estimation of item relevance by zeroing gradients for uninformative negatives. We then describe how this Sampled Entmax for Cold-start (SEMCo) training regime can be extended via knowledge distillation, and show that it outperforms existing cold-start methods and standard sampled softmax in ranking accuracy. We also discuss the advantages of purely content-based modeling, particularly in terms of equity of item outcomes.

P089 · **Reformulating Post-Training as Matrix Factorization for Joint Embedding Refinement** [Short]

YeoJun Choi (Chung-Ang University); Yoon-Sik Cho (Chung-Ang University)

Post-training embedding refinement has recently emerged as a promising approach to enhance recommendation quality. Despite its promise, existing methods refine only item embeddings while keeping user embeddings fixed. Since both embeddings are jointly learned under the low-rank factorization assumption, this one-sided refinement disrupts their learned alignment. As a result, the joint embedding space loses representation capacity and recommends a narrower set of items. In this paper, we propose REFORM (REformulating post-training as matrix FactOR ization Method), a novel framework that reformulates post-training as a matrix factorization, enabling the joint realignment of user and item embeddings. However, the post-training setting provides pre-trained embeddings, not a target matrix designed for factorization. Therefore, REFORM constructs a target matrix with two objectives: (i) preserving the angular structure of pre-trained embeddings, which reliably encodes user-item preferences regardless of frequency, and (ii) incorporating frequency signals that carry confidence about preference. Experiments on three benchmark datasets demonstrate that REFORM achieves up to 26.9% Recall@20 and 29.2% NDCG@20 improvements over post-training baselines, while consistently maintaining higher effective rank and item coverage, mitigating the trade-off between accuracy and diversity observed in prior methods.

P090 · **MonacGraph: A Monadic Second-Order Logic Extended Graph Database System with Community-Aware Storage**

[Demo]

Yuntao Jin (Tsinghua University); Songyao Wang (Tsinghua University); Chaokun Wang (Tsinghua University)

Graph database systems play a vital role in graph structure analysis across a wide range of application domains. Queries with set-level constraints on community structures are increasingly demanded in real-world applications. However, existing graph databases lack native support for both efficient monadic second-order logic (MSOL) query processing and fast community retrieval, hindering their applicability to such analytical tasks. In this paper, we present Monac- Graph, a graph database system that enables practical MSOL queries. MonacGraph features an efficient two-phase execution engine that minimizes redundant first-order clause evaluations. We propose SO-Gremlin, an extension of the Gremlin graph traversal language with intuitive syntax for set quantification. The system adopts LSM-Community as its storage backend, enabling efficient queries over precomputed graph structures. Additionally, MonacGraph provides a user-friendlyWeb interface for composing complex set-level queries and visualizing results in real time. A demonstration video can be found at https://www.youtube.com/watch?v=Eezdq9tzbJE.

P091 · **IR Lens: A Tool for Interpreting Cross-Encoder Models**

[Demo]

Mihai Branga-Peicu (Sorbonne Université); Mathias Vast (Sinequa by ChapsVision); Basile Van Cooten (Sinequa by ChapsVision); Laure Soulier (Sorbonne Université); Jules Françoise (Université Paris-Saclay); Benjamin Piwowarski (Sorbonne Université); Baptiste Caramiaux (Sorbonne Université)

Transformer-based ranking models, such as MonoBERT, are central to Information Retrieval; yet their inner workings remain largely opaque. This hinders not only our understanding of the systems implementing them, but also our ability to improve them. To alleviate this limitation, we introduce IR Lens, a new interpretability tool tailored to cross-encoders based on two key components: 1) Neuron Integrated Gradients to expose the contributions of model parts at multiple levels, and 2) targeted ablations to support hypothesis tracking. With its interactive graphical interface, IR Lens enables IR practitioners to explore, analyze, and manipulate neuron-level mechanisms in cross-encoders, facilitating a deeper understanding of neural ranking models. By extending the reach of existing interpretability methods, we believe IR Lens has the potential to support the improvement of cross-encoders.

P092 · **ExODRec: An Explainable Framework for Outlier Detection Model Recommendation** [Short]

Saba Fathi Rabooki (RMIT University); Ziqi Xu (RMIT University); Elham Naghizade (RMIT University)

The rapid growth of large-scale datasets in outlier detection (OD) makes exhaustive evaluation of diverse candidate models for each new dataset impractical. The unsupervised outlier model recommendation (UOMR) problem remains underexplored and presents two major challenges: severe cold-start constraints and the lack of explainable model recommendation. To address these issues, we propose ExODRec, an explainable two-stage meta-learning framework for UOMR. In the first stage, a meta-learner is trained on benchmark OD datasets using their meta-features and historical model performances. The meta-learner integrates an item response theory (IRT)-based latent trait extractor with a performance predictor, enabling interpretable modelling of model difficulty, discrimination ability, and dataset challenge level. In the second stage, for a new dataset, ExODRec predicts latent traits from meta-features to estimate model performance and generate ranked recommendations without executing candidate models. Experiments on benchmark datasets demonstrate that ExODRec outperforms state-of-the-art baselines while providing interpretable model recommendations. The code is available at https://github.com/SabaFathi/ExODRec-SIGIR2026.

P093 · **From Single- to Cross-Document: Benchmarking Multi-Granularity Event Analysis of Large Language Models**

[Resource]

Tao Wen (University of Electronic Science and Technology of China); Shuai Shao (University of Electronic Science and Technology of China); Pei Ke (University of Electronic Science and Technology of China); Xu Han (Tsinghua University); Jie Zou (University of Electronic Science and Technology of China); Guannan Li (University of Electronic Science and Technology of China); Tao Tian (University of Electronic Science and Technology of China); Jinjie Qiu (University of Electronic Science and Technology of China); Lan Wang (University of Electronic Science and Technology of China); Ke Qin (University of Electronic Science and Technology of China)

Event analysis is an essential and fundamental direction of information extraction, involving various event-centric tasks at different granularity of documents. While large language models (LLMs) have preliminarily achieved promising performance in part of these tasks individually, their capability in event analysis still lacks comprehensive understanding due to restricted document granularity, task designs, and data source of existing benchmarks. To address these limitations, we introduce MiGUE-Bench, a systematic benchmark for assessing the performance of LLMs in multi-granularity event analysis. To support large-scale evaluation, we first develop an LLM-driven self-correcting annotation framework called MiGUE-Pipeline, enabling scalable acquisition of high-quality source data of events with automatic labels. Then, we design four core tasks in our benchmark, i.e., event detection, relation reasoning, structure induction, and future prediction, to probe model competence at different levels, from atomic event details to complex cross-document narratives. Extensive experiments on state-of-the-art LLMs and retrieval-augmented generation (RAG) methods delineate the current capability boundary and identify critical deficiencies, providing insights into the the future improvement of LLMs in challenging event analysis tasks.

P094 · **TReB: A Comprehensive Benchmark for Evaluating Table Reasoning Capabilities of Large Language Models**

[Resource]

Ce Li (JIUTIAN Research); Xiaofan Liu (JIUTIAN Research); Zhiyan Song (JIUTIAN Research); Ce Chi (JIUTIAN Research); Boshen Shi (JIUTIAN Research); Chen Zhao (JIUTIAN Research); Guanguang Chang (JIUTIAN Research); Zhendong Wang (JIUTIAN Research); Kexin Yang (JIUTIAN Research); Xing Wang (JIUTIAN Research); Chao Deng (JIUTIAN Research); Junlan Feng (JIUTIAN Research)

The majority of data in businesses and industries is stored in tables, databases, and data warehouses. Reasoning with table-structured data poses significant challenges for large language models (LLMs) due to its hidden semantics, inherent complexity, and structured nature. One of these challenges is lacking an effective evaluation benchmark fairly reflecting the performances of LLMs on broad table reasoning abilities. In this paper, we fill in this gap by presenting a comprehensive table reasoning benchmark, TReB. Firstly, we propose a taxonomy to systematically measure both shallow table understanding abilities and deep table reasoning abilities, covering a total of 26 sub-tasks. We then construct a high quality dataset through a dedicated data processing and synthesis procedure. Based on these well-constructed samples, we design an evaluation framework to robustly measure table reasoning capabilities with three distinct inference modes. Experimental results with our data and framework reveal that existing LLMs still have significant room for improvement in addressing the complex and real world table related tasks. Both the dataset and evaluation framework are publicly available, with the dataset hosted on huggingface.co/datasets/JT-LM/JIUTIAN-TReB, and the framework on github.com/JT-LM/jiutian-treb.

P095 · **CommonWhy: A Dataset for Evaluating Entity-Based Causal Commonsense Reasoning in Large Language Models**

[Resource]

Armin Toroghi (University of Toronto); Faeze Moradi Kalarde (University of Toronto); Scott Sanner (University of Toronto)

To effectively interact with the real world, Large Language Models (LLMs) require entity-based commonsense reasoning, a challenging task that necessitates integrating factual knowledge about specific entities with commonsense inference. Existing datasets for evaluating LLM entity-based commonsense reasoning have largely focused on True/False or multiple-choice questions, leaving the explicit assessment of the model's ability in abductive reasoning about causes and effects and generating explanations largely unexamined. In this work, we introduce CommonWhy, a

dataset of 15,000 why questions designed to evaluate entity-based commonsense reasoning about causal relationships in LLMs. CommonWhy also serves as a Knowledge Graph Question Answering (KGQA) benchmark, as all supporting knowledge required to answer its queries is available in the Wikidata knowledge graph. Unlike existing KGQA datasets, which primarily test fact retrieval, CommonWhy targets causal commonsense reasoning, establishing a new paradigm for KGQA evaluation. Experiments with state-of-the-art LLMs and LLM-based KGQA methods reveal their significant shortcomings, including frequent factual hallucinations and failures in causal reasoning.

P096 · **Content-Based Dataset Knowledge Graphs for Dataset Search** [Resource]

Qing Shi (Nanjing University); Jing He (Nanjing University); Xintian Pan (Nanjing University); Jialiang Wan (Nanjing University); Gong Cheng (Nanjing University)

Dataset search remains hindered by a reliance on sparse, incomplete metadata that overlooks semantic connections in the data itself. We introduce CoDaKG, the first content-enriched dataset knowledge graph aligned with the NTCIR benchmark. Unlike metadata-only methods, CoDaKG systematically extracts relationships from both metadata and heterogeneous data files. We also release an open-source Python toolkit for easily building CoDaKG instances for any dataset collection. Experiments demonstrate that CoDaKG's enriched signals exhibit a positive trend in ad hoc retrieval and yield significant improvements in graph-based re-ranking of datasets. All resources are publicly available.

P097 · **rerankers: A Lightweight Unified Toolkit for Reranking Approaches** [Resource]

Benjamin Clavié (Mixedbread AI and National Institute of Informatics (NII))

This paper presents \texttt{rerankers}, a Python library which provides an easy-to-use interface to the most commonly used re-ranking approaches. Re-ranking is an integral component of many retrieval pipelines; however, there exist numerous approaches to it, relying on different implementation methods. \texttt{rerankers} unifies these methods into a single user-friendly interface, allowing practitioners and researchers alike to explore different methods while only changing a single line of Python code. Moreover, \texttt{rerankers} ensures that its implementations are done with the fewest dependencies possible, facilitating its use in any development environment. Methods within rerankers are tested on common benchmarks, guaranteeing that our simplified implementations and interfaces results in no performance degradation compared to more complex ones. Finally, rerankers' extensibility facilitates its integration with broader IR toolkits and supports its application for new domain, with multimodal, text-to-image rerankers being integrated in the library.The full source code and list of supported models are updated regularly and available at \url{https://github.com/answerdotai/rerankers}.

P098 · **Targeted Parameter Selection for Rank-Biased Measurement** [Short]

Alistair Moffat (The University of Melbourne)

The rank-biased family of measurements allows ordered rankings to be compared in a top-weighted manner, with the differential emphasis on early items derived from a decaying geometric distribution across positions in the ranking. In this short paper we consider a range of other weighting functions, including both infinite and finite inverse power law distributions. We show that the introduction of a second governing parameter provides a commensurate increase in flexibility, and describe principles and algorithmic techniques for selecting suitable parameters so as to achieve defined experimental goals. The result is greater flexibility of measurement, and tools that suit a wider range of experimental scenarios

P099 · **Modalities of Expert Search: How Users Transition Between Names and Topics** [Short]

Marjan Azimi (The University of Melbourne); Alistair Moffat (The University of Melbourne); Justin Zobel (The University of Melbourne)

Expert finding systems support two distinct search modalities: name-based queries targeting known individuals, and topic-based queries exploring expertise areas, with users often moving between these modes as their search progresses. In this study, we analyze nearly half a million queries submitted to an institutional expert search system, using a lexicon-based approach to classify each query as "name", "topic", or "unclear". We find that most queries in multi-query sessions have the same mode as the previous query, but that more than a fifth of all such sessions begin and end in different modes. Moreover, mode switching is partially success-driven: queries followed by clicks then switch modes more often than do queries without clicks. Together, these findings show that expert search behavior is shaped by interactions between search modalities, session-level dynamics, and observed engagement outcomes, suggesting several directions for improving expert search systems.

P100 · **FactOWL: A Cost-Efficient Tool for Long-Form Factuality Evaluation** [Demo]

Andrey Sakhovskiy (Applied AI Institute); Nikita Sushko (Applied AI Institute); Maria Marina (Applied AI Institute); Vasily Konovalov (Institute for Big Data Analysis and Machine Learning); Elena Tutubalina (Institute for Big Data Analysis and Machine Learning); Alexander Panchenko (Institute for Big Data Analysis and Machine Learning); Pavel Braslavski (HSE University)

The presence of factual hallucinations in large language model (LLM) generations drives the development of retrieval-augmented factuality evaluation tools. However, existing implementations suffer from slow inference, outdated knowledge bases, and the use of paid search and LLM APIs, which hinders research and practical applications. To fill this gap, we propose FactOWL, a FActScore-based Factuality evaluation tool which adopts an Open LLM and real-time Wikipedia search for evaluation of Long-form LLM responses. FactOWL effectively addresses unreliable evidence, incomplete contexts, and entity ambiguity by performing multi-source aggregation over Wikipedia pages and Wikidata triples. FactOWL is at least 10× faster than the pioneering FActScore fact checking tool and its key modifications methods and is freely available on GitHub: https://github.com/s-nlp/factowl.

P101 · **G-CoS: An Interpretable Gain-Cost Framework for User Satisfaction Estimation in Generative Information Retrieval** [Short]

Jia-Ling Shi (Beijing Institute of Technology); Zhijing Wu (Beijing Institute of Technology); Yidong Liang (Beijing Institute of Technology); Xian-Ling Mao (Beijing Institute of Technology)

User satisfaction serves as a key indicator of search experience in Generative Information Retrieval (GenIR) systems, and its accurate estimation is essential for system optimization. Extensive research in traditional IR has established interaction signals (e.g., dwell time, clicks, query reformulations) as reliable indicators of user satisfaction. However, prevailing approaches for satisfaction estimation in GenIR (e.g., LLM-as-a-judge) primarily rely on textual content and fail to account for user interactions during the search process. In this work, empirical analysis of real-world data shows that user satisfaction correlates negatively with interaction signals reflecting interaction cost, and positively with response quality. Building on these findings, we propose the Gain-aware Cost-sensitive Satisfaction estimator (G-CoS), an interpretable gain-cost framework for user satisfaction estimation in GenIR. G-CoS models user satisfaction as a dynamic trade-off between Response Quality and multidimensional Interaction Cost. Experimental results demonstrate that G-CoS outperforms LLM-as-a-judge methods, interaction sequence models, as well as machine learning models using the same gain and cost features. Moreover, the learned parameters reveal interpretable associations between gain-cost dynamics and user satisfaction. This work contributes an interpretable framework for user satisfaction estimation and offers insights for GenIR system optimization.

P102 · **Unifying On- and Off-Policy Variance Reduction Methods** [Short]

Olivier Jeunen (aampe)

Continuous and efficient experimentation is key to the practical success of user-facing applications on the web, both through online A/B-tests and off-policy evaluation. Despite their shared objective---estimating the incremental value of a treatment---these domains often operate in isolation, utilising distinct terminologies and statistical toolkits. This paper bridges that divide by establishing a formal equivalence between their canonical variance reduction methods. We prove that the standard online Difference-in-Means estimator is mathematically identical to an off-policy Inverse Propensity Scoring estimator equipped with an optimal (variance-minimising) additive control variate. Extending this unification, we demonstrate that widespread regression adjustment methods (such as CUPED, CUPAC, and ML-RATE) are structurally equivalent to Doubly Robust estimation. This unified view extends our understanding of commonly used approaches, and can guide practitioners and researchers working on either class of problems.

P103 · **Confidence-Based Stopping Methods for Systematic Reviews** [Short]

Aaron H.A. Fletcher (University of Sheffield); Mark Stevenson (University of Sheffield)

Technology Assisted Review stopping methods aim to ensure that no more documents are screened than necessary. Most existing approaches focus on achieving a target recall, which does not consider whether an information need has been met. This paper introduces two heuristic stopping methods that instead monitor whether screened documents contain enough information to make a decision. Evaluation on a standard dataset of Diagnostic Test Accuracy Systematic Reviews demonstrates that the proposed approaches substantially reduce the number of documents that need to be examined while, in the majority of cases, maintaining conclusions that are consistent with all evidence available.

P104 · **SQuTR: A Robustness Benchmark for Spoken Query to Text Retrieval under Acoustic Noise** [Resource]

Yuejie Li (Ant Group); Ke Yang (The University of Hong Kong); Yueying Hua (Soochow University); Bolin Chen (University of Science and Technology of China); Jianhao Nie (Wuhan University); Yueping He (Tsinghua University); Caixin Kang (The University of Tokyo)

Spoken query retrieval is an important interaction mode in modern information retrieval. However, existing evaluation datasets are often limited to simple queries under constrained noise conditions, making them inadequate for assessing the robustness of spoken query retrieval systems under complex acoustic perturbations. To address this limitation, we present SQuTR, a robustness benchmark for spoken query retrieval that includes a large-scale dataset and a unified evaluation protocol. SQuTR aggregates 37,317 unique queries from six commonly used English and Chinese text retrieval datasets, spanning multiple domains and diverse query types. We synthesize speech using voice profiles from 200 real speakers and mix 17 categories of real-world environmental noise under controlled SNR levels, enabling reproducible robustness evaluation from quiet to highly noisy conditions. Under the unified protocol, we conduct large-scale evaluations on representative cascaded and end-to-end retrieval systems. Experimental results show that retrieval performance decreases as noise increases, with

substantially different drops across systems. Even large-scale retrieval models struggle under extreme noise, indicating that robustness remains a critical bottleneck. Overall, SQuTR provides a reproducible testbed for benchmarking and diagnostic analysis, and facilitates future research on robustness in spoken query to text retrieval.

P105 · **QE-RAG: A Robust Retrieval-Augmented Generation Benchmark for Query Entry Errors** [Resource]

Kepu Zhang (Renmin University of China); Zhongxiang Sun (Renmin University of China); Weijie Yu (University of International Business and Economics); Xiaoxue Zang (Kuaishou Technology Co., Ltd); Kai Zheng (Kuaishou Technology Co., Ltd); Yang Song (Kuaishou Technology Co., Ltd); Han Li (Kuaishou Technology Co., Ltd); Jun Xu (Renmin University of China)

Current benchmarks evaluate the performance of RAG methods from various perspectives, they share a common assumption that user queries used for retrieval are error-free. However, in real-world interactions between users and LLMs, query entry errors are frequent. The impact of these errors on current RAG methods against such errors remains largely unexplored. To bridge this gap, we propose QE-RAG, the first robust RAG benchmark designed specifically to evaluate performance against query entry errors. We analyze the impact of these errors on LLM outputs and find that corrupted queries degrade model performance, which can be mitigated through query correction and training a robust retriever for retrieving relevant documents. Based on these insights, we propose a contrastive learning-based robust retriever training method and a retrieval-augmented query correction method. Extensive experiments reveal that: (1) state-of-the-art RAG methods including sequential, branching, and iterative methods, exhibit poor robustness to query entry errors; (2) our method enhances the robustness of RAG when handling query entry errors and it's compatible with existing RAG methods, further improving their robustness.

P106 · **Efficient Query Region Expansion and Decomposition based Spatial Range Query Algorithm** [Short]

Lianyin Jia (Kunming University of Science and Technology); Rongjin Wang (Kunming University of Science and Technology); Yingbin Su (Kunmimg University of Science and Technology); Suprio Ray (University of New Brunswick); Mengjuan Li (Yunnan Normal University); Jiaman Ding (Kunming University of Sciences and Technology)

Spatial range queries play a crucial role in spatial information retrieval. Existing Z-order curve based algorithms suffer from accessing a large number of invalid points outside the query region. To address this challenge, we design a simple yet efficient Z-order curve based learned index, ZPI. Building upon ZPI, we propose a novel spatial range query algorithm, ZPI-RQ. ZPI-RQ leverages efficient decomposition mechanism to address the invalid points issue. To avoid decomposing the query region into a large number of overly small blocks, a query region expansion strategy is further introduced to align each query border with a m-order dividing lines. Experimental results show that ZPI-RQ significantly outperforms state-of-the-art algorithms in query efficiency, achieving a 3.6× improvement over traditional Z-order curve-based algorithms while accessing only 1% invalid points.

P107 · **A Taxonomy and Catalog of SERP Elements Across Web Search Engines** [Resource]

Adelaide Miranda Santos (University of Porto); Carla Teixeira Lopes (INESC TEC)

Search engines serve as a primary gateway to the web, and their interfaces must effectively support users' search tasks. Over time, Search Engine Results Pages (SERPs) have evolved from the traditional "list of 10 blue links" into sophisticated and diversified interfaces, incorporating a wide range of features to facilitate the search process. However, analyses of SERP elements often suffer from inconsistent terminology, making cross-study comparisons difficult. To address this issue, this work proposes a comprehensive catalog of SERP elements that standardizes their classification. We examined leading web search engines (Google, Bing, Yandex, Yahoo!, Baidu and DuckDuckGo), assigned consistent names to elements, and organized them into groups. In total, 143 elements were identified and classified as organic results, sponsored results, or features. By publishing this catalog on a dedicated website and in a research repository, this work provides a valuable resource for future research on search interfaces.

P108 · **SAGE: Solver-Aligned Guided Exploration** [Short]

Nikita Sorokin (MWS AI); Ivan Sedykh (MWS AI); Timur Ionov (MWS AI); Valentin Malykh (MWS AI)

Modern code agents achieve strong results on challenging software engineering benchmarks such as SWE-bench, but solving each issue remains expensive: most inference cost is spent not on patch generation, but on repository exploration and search, accounting for up to 56% of tokens in our experiments. We propose SAGE, a modular approach that trains a small searcher model to handle codebase exploration as a tool for a frozen large solver model. We first distill the search trajectories from a strong agent to obtain a compact searcher with comparable retrieval quality. We then apply reinforcement learning to optimize the searcher for usefulness under the solver's fixed interface, using step-level feedback that directly evaluates whether retrieved context is actionable for downstream patching. On SWE-bench Verified, SAGE improves resolve rate while reducing search overhead by 67% and overall cost by 21%, demonstrating a practical path to cheaper, modular software engineering agents.

P109 · **CSyMR: Benchmarking Compositional Music Information Retrieval in Symbolic Music Reasoning** [Short]

Boyang Wang (University of California, San Diego); Yash Vishe (University of California, San Diego); Xin Xu (University of California, San Diego); Zachary Novack (University of California, San Diego); Xunyi Jiang (University of California, San Diego); Julian McAuley (University of California, San Diego); Junda Wu (University of California, San Diego)

Natural language information needs over symbolic music scores rarely reduce to a single-step lookup. Many queries require compositional Music Information Retrieval (MIR) that extracts multiple pieces of evidence from structured notation and aggregates them to answer the question. This setting remains challenging for Large Language Models due to the mismatch between natural language intents and symbolic representations, as well as the difficulty of reliably handling long structured contexts. Existing benchmarks only partially capture these retrieval demands, often emphasizing isolated theoretical knowledge or simplified settings. We introduce CSyMR-Bench, a benchmark for compositional MIR in symbolic music reasoning grounded in authentic user scenarios. It contains 126 multiple-choice questions curated from community discussions and professional examinations, where each item requires chaining multiple atomic analyses over a score to derive implicit musical evidence. To support diagnosis, we provide a taxonomy with six query intent categories and six analytical dimension tags. We further propose a tool-augmented retrieval and reasoning framework, CSyMR-Agent, that integrates a ReAct-style controller with deterministic symbolic analysis operators built with music21. Experiments across prompting baselines and agent variants show that tool-grounded compositional retrieval consistently outperforms Large Language Model-only approaches, yielding 5-7% absolute accuracy gains, with the largest improvements on analysis-heavy categories.

P110 · **Cross-Sensory Comparison of EEG Signals for Brain-Based Information Retrieval** [Short]

Niall McGuire (University of Strathclyde); Yashar Moshfeghi (University of Strathclyde)

Translating internal information needs into textual queries poses challenges for information retrieval, particularly for users with physical impairments or ill-defined search intentions. Brain Passage Retrieval (BPR) approaches map EEG signals directly to dense passage representations, bypassing text translation. Whilst recent work reports superior performance with auditory versus visual EEG, these findings emerge from unbalanced experimental conditions where sample sizes, vocabularies, and corpus characteristics differ substantially between modalities, making it unclear whether observed differences reflect genuine neural processing advantages or dataset artefacts. We investigate EEG-based retrieval performance comparing auditory and visual modalities under controlled dataset conditions. Using the Brennan (auditory, 49 subjects) and Nieuwland (visual, 51 subjects) datasets, balanced for vocabulary overlap and sample distributions, we train BPR models with transformer EEG encoders and BERT text encoders via contrastive learning. Under controlled conditions, we observe complementary performance characteristics: audio EEG demonstrates stronger recall (Hit@5: +74%, Hit@10: +140%) whilst visual EEG achieves better precision (Hit@1: +100%). These findings suggest that modality-specific strengths could inform the design of brain-computer interfaces for information retrieval.

P111 · **MSW-NTM: A Spherical Wasserstein Autoencoder for Multimodal Neural Topic Modeling with LLM-Guided Topic Refinement** [Short]

Dayu Guo (Beijing Normal-Hong Kong Baptist University); Zhiwen Luo (Concordia University); Nizar Bouguila (Concordia University); Wentao Fan (Beijing Normal-Hong Kong Baptist University)

Multimodal topic modeling aims to uncover interpretable semantic structures from documents containing both text and images, while also supporting downstream tasks such as multimodal retrieval. However, existing multimodal neural topic models, typically based on standard Variational Autoencoders (VAEs), often suffer from posterior collapse and fail to account for the directional nature of L2-normalized embeddings. To address these limitations, We propose MSW-NTM (Multimodal Spherical Wasserstein Neural Topic Model) which is built upon a Wasserstein Autoencoder that performs topic modeling on the unit hypersphere. It adopts a von Mises-Fisher mixture prior and aligns the aggregated posterior with the prior using the Spherical Sliced-Wasserstein distance, ensuring stable training and improved topic interpretability aligned with multimodal embedding geometry. Moreover, we incorporate an Large Language Model (LLM)-guided topic refinement module to further enhance topic coherence. Experimental results on standard multimodal benchmarks demonstrate that MSW-NTM delivers high-quality, interpretable topic representations, supporting retrieval-oriented semantic organization and exploration.

P112 · **A Multi-Granularity Game-Theoretic Approach to Weakly Supervised Temporal Article Grounding** [Short]

Shuyi He (Institute of Automation, Chinese Academy of Sciences); Hao Wen (Beijing Wenge Technology Co., Ltd.); Pu Zou (Tianwen Digital Media Technology (Hunan) Co., Ltd.); Song Zhou (Beijing Wenge Technology Co., Ltd.); Qingchao Kong (Institute of Automation, Chinese Academy of Sciences)

Weakly Supervised temporal Article Grounding (WSAG) is an important task in the field of video understanding and retrieval, which aims to ground sentences from the article with the corresponding video clips (segments), using only video-text pairwise annotations during training. Existing methods typically rely on a proposal-based approach that utilizes contrastive learning to score candidate video clips. However, these approaches often face two main drawbacks: (1) They mainly focus on global video-level alignment with the article while neglecting the multi-level relationships between video clips

and text hierarchy; (2) They overlook intra-modal feature learning, failing to adequately distinguish salient patterns. To address these issues, we propose a novel multi-granularity game-theoretic method, namely Hierarchical Cooperative Network (HCN), which models video clips and article units (i.e., at word, sentence, and paragraph layers) as players in a game. Specifically, our proposed method introduces two mechanisms: (1) Intra-modal feature cooperation: Within each modality, we encourage cooperation among the features to obtain a salient feature set as the effective representations, enabling a sharper distinction between semantically overlapping video clips and enhancing discriminative textual cues; (2) Layered cross-modal cooperation: For different semantic layers of the article, we employ the Shapley interaction index to quantify and promote synergistic effects to further enhance cross-modal matching. By decomposing multi-modal interactions into intra-modal and cross-modal cooperations with multi-layer semantics, HCN achieves fine-grained alignment and strengthens salient feature representations for the WSAG task. Experimental results demonstrate that HCN achieves the state-of-the-art performances on representative WSAG benchmarks, significantly outperforming existing weakly supervised methods and competitive video large language models.

P113 · **A Sketch+Text Composed Image Retrieval Dataset for Thangka** [Resource]

Jinyu Xu (Wuhan University of Technology); Yi Sun (Wuhan University of Technology); Jiangling Zhang (Wuhan University of Technology); Qing Xie (Wuhan University of Technology); Daomin Ji (RMIT University); Zhifeng Bao (The University of Queensland); Jiachen Li (Wuhan University of Technology); Yanchun Ma (Wuhan Vocational College of Software and Engineering); Yongjian Liu (Wuhan University of Technology)

Composed Image Retrieval (CIR) enables image retrieval by combining multiple query modalities, but existing benchmarks predominantly focus on general-domain imagery and rely on reference images with short textual modifications. As a result, they provide limited support for retrieval scenarios that require fine-grained semantic reasoning, structured visual understanding, and domain-specific knowledge. In this work, we introduce CIRThan, a sketch+text composed image retrieval dataset for Thangka imagery, a culturally grounded and knowledge-specific visual domain characterized by complex structures, dense symbolic elements, and domain-dependent semantic conventions. CIRThan contains 2,287 high-quality Thangka images, each paired with a human-drawn sketch and hierarchical textual descriptions at three semantic levels, enabling composed queries that jointly express structural intent and multi-level semantic specification. We provide standardized data splits, comprehensive dataset analysis, and benchmark evaluations of representative supervised and zero-shot CIR methods. Experimental results reveal that existing CIR approaches, largely developed for general-domain imagery, struggle to effectively align sketch-based abstractions and hierarchical textual semantics with fine-grained Thangka images, particularly without in-domain supervision. We believe CIRThan offers a valuable benchmark for advancing sketch+text CIR, hierarchical semantic modeling, and multimodal retrieval in cultural heritage and other knowledge-specific visual domains. The dataset is publicly available at https://github.com/jinyuxu-whut/CIRThan.

P114 · **Same Image, Different Meanings: Toward Retrieval of Context-Dependent Meanings** [Short]

Ayuto Tsutsumi (Tokyo Metropolitan University); Ryosuke Kohita (CyberAgent, Inc.)

A scene of two people in the rain can convey hope and warmth in a reunion story or sorrow and finality in a farewell story. We investigate this context-dependent nature of image meaning and its implications for retrieval. Our key observation is that context dependency correlates with semantic abstraction: concrete elements (objects, actions) remain stable across contexts, while abstract elements (atmosphere, intent) shift with context. We operationalize this as the L1-L4 framework, organizing image semantics from context-independent (L1) to maximally context-dependent (L4). Using synthetic story contexts and queries for controlled evaluation, we examine how injecting narrative context into embeddings affects retrieval across abstraction levels. Concrete queries are retrievable without context, while abstract levels increasingly depend on narrative grounding. Where context is injected also matters, with image-side enrichment proving particularly effective. The most abstract level, however, remains challenging even with full context, highlighting context-dependent image retrieval as an important open problem. Our framework and findings lay groundwork toward retrieval systems that handle the context-dependent meanings images acquire in narrative settings.

P115 · **Booking Funnel and Substitution-Aware User Behavior Modeling for Demand Prediction and Joint Room Pricing** [Short]

Zhikang Fan (Hong Kong University of Science and Technology); Zihao Chen (Xi'an Jiaotong University); Pin Gao (The Chinese University of Hong Kong); Ruohan Zhan (University College London); Shaowen Zhang (Meituan); Xingrui Li (Meituan); Jianpei Wen (Meituan); Su Zhao (Meituan); Shuai Chen (Meituan); Wei Lin (Meituan)

User interactions on online travel platforms (OTPs) naturally follow a two-stage booking funnel, spanning \emph{hotel click} from a multi-hotel listing page and \emph{booking conversion} through room selection within the clicked hotel. Crucially, booking decisions are set-dependent: users select among multiple room types within the same hotel, where price changes of one option can shift demand to others. Two challenges arise in modeling such behavior: (i) the cascading booking-funnel behavior from hotel click to booking conversion, and (ii) intra-hotel substitution across room types. While accurate modeling of such behavior is essential for demand prediction and downstream optimization, existing methods typically ignore this structure by collapsing user behavior into a single stage and overlooking intra-hotel substitution effects, resulting in biased demand estimations and suboptimal decisions. In this paper, we propose BFSNet, a Booking Funnel and Substitution-aware neural network that decomposes booking demand into \emph{hotel click} and \emph{booking conversion}, enabling interpretable and stage-wise learning of user behavior. To capture the intra-hotel substitution effect, we introduce a mixing network-based substitution effect module that incorporates economic structural information by enforcing monotonic substitution relationships between the demand of a focal room and the prices of competing room types. To bridge prediction and downstream decision-making, we further develop a lightweight pricing procedure that distills the learned prediction model to efficiently evaluate candidate price vectors under operational constraints. Experiments on large-scale production data from a major online travel platform demonstrate consistent improvements in demand prediction accuracy and significant gains in downstream revenue.

P116 · **Scalable K-Means Guided Partitioning for Block-based Sparse Document Retrieval** [Short]

Parker Carlson (University of California, Santa Barbara); Sammy Lesner (University of California, Santa Barbara); Antonio Mallia (Seltz); Tao Yang (University of California, Santa Barbara)

Document clustering is a common technique used in sparse retrieval to group similar documents together. The K-means method has been widely adopted to group similar sparse vectors together, but it is not scalable when dealing with a large number of clusters, and the bipartite graph partitioning (BP) method is a preferred choice for block-based document retrieval to partition a large document set efficiently. This paper revisits such a partitioning approach and proposes a balanced K-means-guided bisection method with log-linear complexity while maintaining a good similarity-based clustering quality. Our evaluation with several IR datasets for block-based sparse retrieval algorithms show that the proposed method outperforms the baseline variants of K-means in scalability with 29-414x faster partitioning even for small datasets and reduces retrieval latency by up to 32% compared to BP when clustering 8.8M MS MARCO passages using SPLADE++ embeddings.

P117 · **A Dataset of Cultural Heritage Manipulation on English Wikipedia in the Russo–Ukrainian Context** [Resource]

Maxime Garambois (EPFL); Hamest Tamrazyan (Digital Humanities Institute, EPFL); Emanuela Boros (Digital Humanities Laboratory, EPFL)

Digital knowledge platforms increasingly shape how cultural heritage and historical responsibility are represented and remembered, yet they are vulnerable to subtle forms of narrative manipulation that rarely appear as explicit misinformation. We present a new dataset of English Wikipedia revision pairs designed to support the systematic study of cultural heritage manipulation in the context of the Russo--Ukrainian conflict. The resource captures fine-grained textual changes through which cultural attribution, agency, and legitimacy are incrementally reframed within a collaborative encyclopedic infrastructure. By providing paired revisions, editorial context, and consistent annotations, the dataset enables analysis of discursive interventions that remain procedurally legitimate while producing cumulative cultural effects. The dataset is released to support research in information retrieval, natural language processing, and computational social science on detecting and understanding subtle narrative change in large-scale, collaborative knowledge systems. The resource is available at https://zenodo.org/records/19883068.

Industry Poster Session · *Thursday 15:00–16:15* ·

Exhibition Hall 21B/22B · 129 boards

P001 · **DPEO: Dynamic Preference Evolution Optimization for Self-Evolving CTR Prediction** [Industry]

Kun Yao (JD.com); Congcong Liu (JD.COM); Ziheng Ni (JD.com); Wenlong Chen (JD.com); Changping Peng (JD.com); Ching Law (JD.com)

Click-through rate (CTR) prediction is a pivotal component in large-scale industrial systems. Historically, CTR prediction paradigms have been confined to monolithic architectures governed by a single-policy optimization process. However, such isolated learning paths lack the intrinsic evolutionary mechanisms necessary for optimal convergence. Without policy diversity and internal competition, models tend to get trapped in local optima as performance reaches saturation, hindering further breakthroughs in modeling capacity. In this paper, we propose DPEO (Dynamic Preference Evolution Optimization), a co-evolutionary framework that transforms CTR modeling into a dynamic policy contention task. DPEO decouples the monolithic architecture into dual sub-learners to induce policy diversity, constructing an internal preference landscape without external rewards. A performance-driven Role Arbiter then dynamically designates the superior sub-learner as the Reference Policy and the other sub-learner as the Target Policy per batch, driving continuous model evolution. Through an asymmetric gradient flow, the target policy is optimized to surpass the reference policy in both probability and logit spaces. This process drives a co-evolution, enabling the sub-learners to serve as alternating evolutionary benchmarks and 'self-evolve' toward the global optimum. Extensive experiments on public benchmarks and a massive industrial dataset with over 10 billion samples demonstrate that DPEO significantly outperforms state-of-the-art models.

P002 · **Reasoning-Grounded Intent Injection for Generative Recommendation** [Industry]

Xusong Chen (JD.com); Peini Guo (JD.com); Fang Liu (JD.com); Yu Zhao (JD.com); Yiyang Hu (JD.com); Mengqin Que (JD.com); Peng Li (JD.com); Haoran Wang (JD.com); Zhiwei Fang (JD.com); Wenlong Chen (JD.com); Changping Peng (JD.com); Ching Law (JD.com)

Industrial generative recommendation systems operating over discrete Semantic IDs (SIDs) are largely behavior-driven, and thus struggle to proactively activate latent demand before explicit user signals emerge, leading to intent cold-start. To address this, we propose RIGER (Reasoning-grounded Intent injection for GE nerative Recommendation), a deployable two-stage framework that integrates offline large language model (LLM) reasoning into an online generative recommender under strict latency constraints. Offline, to ensure scalable deployment, we distill the latent-intent inference capability of a strong LLM into a lightweight forecasting model using an automated data curation pipeline---leveraging judge-guided prompt calibration and future-query-guided rejection filtering. Online, to bridge the representation mismatch between free-form textual intents and the discrete SID token space, predicted intents are converted into SID-native tokens through a behavior-grounded mapping and injected into the deployed decoder-only retrieval backbone. We further fine-tune the model with beam-aware GRPO, introducing a hierarchical intent-alignment exploration reward in SID space while preserving exploitation behavior through KL regularization. Offline evaluations demonstrate a substantial increase in intent-aligned density and diversity with only a marginal reduction in hindsight recall, indicating that RIGER effectively enhances proactive intent exploration while preserving its capability to exploit historical behaviors. In a large-scale e-commerce display advertising system, RIGER improves clicks by 1.6% and advertiser spend by 1.3%.

P003 · AIPO: Adaptive Anchored Intent-aware Policy Optimization for Generative Recommendation [Industry]

Kun Yao (JD.COM); Congcong Liu (JD.COM); Ziheng Ni (JD.COM); Cai Shang (JD.COM); Wenlong Chen (JD.COM); Changping Peng (JD.COM); Ching Law (JD.COM)

Generative Recommendation (GenRec) has emerged as a significant evolutionary direction in the field of recommendation systems in recent years. However, during the reinforcement learning (RL) alignment stage of end-to-end GenRec, a core challenge is that the low signal-to-noise ratio (SNR) of feedback signals triggers reward hacking, which in turn leads to severe distributional drift. Massive click noise and sparse rewards result in highly unstable policy gradients, making it difficult to balance reward optimization and generative stability. To address this, we propose AIPO (Adaptive Anchored Intent-aware Policy Optimization) framework. AIPO synergistically enhances both optimization stability and performance from both data and prior anchoring perspectives. First, it introduces an intent-aware asymmetric resampling mechanism to purify high-confidence conversion intent data, thereby amplifying the gradient contribution of high-value paths. Second, it introduces a prior anchoring mechanism, which dynamically regulates its intensity through a barrier mapping based on the degree of policy distributional drift, mitigating distributional drift and reward hacking caused by low-SNR data. Offline experiments on both public benchmarks and JD's industrial datasets validated the superior performance of AIPO.

P004 · Cheaper is Better: A Discount-Aware Network for Conversion Rate Prediction in E-commerce Recommendation System [Industry]

Ruocong Tang (Alibaba Group); Yang Huang (Alibaba Group); Xing Fang (Alibaba Group); Chenyi Yan (Alibaba Group); Chuike Sun (Alibaba Group); Jing Wang (Alibaba Group)

Post-click conversion rate (CVR) is a crucial element in online recommendation systems, which addresses significant challenges such as data sparsity (DS), sample selection bias (SSB), and delayed feedback. However, the impact of item discount rate-a key factor influencing both pricing and user purchasing behavior, has received limited attention. In this paper, we introduce the Discount-Aware Network (DANet) to model the relationship between item discount rates and CVR. DANet comprises three main components: 1) a time-frequency transformation module that utilizes Fourier transform to derive the frequency spectrum and capture the long-term discount rate trends of items; 2) a distribution de-bias module designed to mitigate the biases in user-specific discount rates caused by various purchase combinations and promotional activities, as well as periodic deviations linked to different promotion periods on e-commerce platforms; and 3) a supervised regression auxiliary task that establishes the explicit item discount labels to enhance the model's performance in terms of value accuracy, facilitating an effective representation of item discount rates. Experimental results on real datasets demonstrate the superiority of DANet, with offline AUC improving by 1.61%, and online A/B test also shows that DANet achieves impressive gains of 3.63% on pCVR and 2.23% on GMV. DANet has been successfully deployed on Alibaba Tmall APP. The code is available at https://github.com/tangrc/DANet.

P005 · PGSF: Profile-Guided Semantic Fusion for Decoupled Industrial Pre-Ranking [Industry]

Chuike Sun (Taobao & Tmall Group of Alibaba); Nan Xu (Taobao & Tmall Group of Alibaba); Xing Fang (Taobao & Tmall Group of Alibaba); Yang Huang (Taobao & Tmall Group of Alibaba); Jing Wang (Taobao & Tmall Group of Alibaba); Junzhou Chen (Guangdong Provincial Key Laboratory of Intelligent Transportation Systems)

Pre-ranking is a critical stage in large-scale recommendation systems, strictly demanding both high efficiency and effectiveness. Despite being the industry benchmark for pre-ranking owing to its serving efficiency, the two-tower structure, by virtue of its decoupled architecture, inherently restricts fine-grained user-item semantic interactions. In this paper, we propose \textbf{P}rofile-\textbf{G}uided \textbf{S}emantic \textbf{F}usion (\textbf{PGSF}), a plug-and-play framework that injects target-item guidance during training while preserving tower decoupling at inference. PGSF comprises three modules: (1) \textbf{P}rofile-\textbf{I}tem \textbf{S}emantic \textbf{A}lignment (\textbf{PISA}) aligns user profile semantics with the target item via contrastive learning to generate a profile-guided query; (2) \textbf{M}ulti-\textbf{I}nterest \textbf{S}emantic \textbf{M}odeling (\textbf{MISM}) constructs a discrete semantic space using RQ-VAE to explicitly model heterogeneous user interests; (3) \textbf{M}ulti-\textbf{L}evel \textbf{A}daptive \textbf{F}usion (\textbf{MLAF}) hierarchically retrieves target-aware interests from historical behaviors and semantic candidates, then fuses them with the backbone user representation to produce a fine-grained enhanced user representation. Training jointly optimizes ranking, alignment, and interest prediction losses. Evaluated on Tmall App, PGSF integrates smoothly into existing decoupled pre-ranking systems, achieving +2.66\% uCTR and +3.13\% IPV gains. The implementation is available at \url{https://github.com/SunChuike/PGSF}.

P006 · Learning to Forget: Satiation-Aware Long-Sequence Transducers for Mitigating Post-Purchase Redundancy [Industry]

Yipin Dai (Alibaba Group); Ruocong Tang (Alibaba Group); Xing Fang (Alibaba Group); Yang Huang (Alibaba Group); Jing Wang (Alibaba Group); Zhentao Song (Alibaba Group); He Guo (Alibaba Group)

Sequential recommendation models predominantly interpret user interactions as positive signals for preference accumulation. However, in e-commerce scenarios, a purchase action often signifies the termination of a specific intent ("Interest Exit") rather than its continuation. Existing models overlook this distinction, suffering from Action-Intent Asymmetry, which leads to severe post-purchase redundancy. In this paper, we propose the Satiation-Aware Mechanism (SAM), an end-to-end framework designed to explicitly model the lifecycle of user interests. SAM incorporates three key components: (1) A Dual-path Cross-Attention architecture that retroactively suppresses historical clicks associated with a fulfilled intent while simultaneously retrieving personalized replenishment rhythms from long-term purchase history; (2) An Adaptive Satiation Gating Unit (ASGU) that generates a time-sensitive soft mask to inhibit satisfied interests immediately after purchase and gradually "re-awaken" them as the predicted repurchase cycle approaches; and (3) A self-supervised Time-to-Next-Purchase (TTNP) auxiliary task to learn latent product lifecycles without manual annotation. Extensive offline experiments on industrial datasets and online A/B testing demonstrate that SAM significantly reduces the Post-Purchase Repeat Rate (PPRR) by over 60%.

P007 · Automated Feature Lifecycle Management in Large-Scale Continual Ranking Systems [Industry]

Bhavani Shankar P S V N (Flipkart); Thejus Vm (Flipkart); Surender Kumar (Flipkart)

Response modeling is central to e-commerce content discovery. Data distributions in these environments are highly non-stationary. To counteract this concept drift, response models rely on continual learning frameworks to perform frequent updates. However, this rapid iteration breeds "feature creep", which is an accumulation of redundant input signals that bloat models and increase inference latency. Traditional offline feature selection is computationally prohibitive and incompatible with streaming, non-stationary data. In this paper, we present an end-to-end differentiable feature selection mechanism deployed within a production-scale continual learning framework at Flipkart. We introduce a Feature Gated Dense Layer that uses Straight-Through Estimators (STE) to learn discrete binary masks during training. To bridge the train-inference disparity and extract feature importance, we relax these gates to continuous probabilities during serving, yielding a deterministic importance score. We demonstrate that STE's stochastic gating outperforms Gumbel-Sigmoid relaxation in production when stabilized by Probability-Scaled Average Normalization and a Two-Stage Freezing protocol. Evaluated on e-commerce search traffic, our method pruned approximately 20% of the feature space and down-weighted another 25%, with an average AUC drop of just 0.019%. This establishes a scalable and interpretable paradigm for automated feature lifecycle management.

P008 · Calibrate Your Items! Train CTR Models in Two Steps for Robust Recommendations [Industry]

Yi Han (Amazon); Jay Fleischman (Amazon)

Click-through rate (CTR) prediction is a central task in recommender and information retrieval (IR) systems, with studies primarily benchmarking on AUC and LogLoss. However, when combined with other downstream components in a larger optimization pipeline to generate final ranking scores (e.g., ad auctions: pCTR × bid), calibration of predicted CTR is critical for ensuring overall system performance. In this paper, we explore a particularly important and understudied calibration slice: item-level calibration. We present group-wise absolute bias ratio (gaBR) as a robust metric for quantifying item-level prediction calibration and demonstrate its importance for recommendation quality. We also introduce TALENT - a simple add-on learning strategy - for training modern CTR neural networks to achieve robust item-level calibration and improve upon state-of-the-art AUC, without the need for a post-processing calibration step. We demonstrate the impact and robust performance of the TALENT framework beyond current calibration techniques in offline model experiments (including on the public Criteo, Avazu, and KKBox datasets and across features of widely different cardinalities) and multiple A/B tests with millions of live users, boosting user

response over methods that optimize for AUC or item-level calibration alone.

P009 · **Next Interest Flow: A Generative Pre-training Paradigm for Recommender Systems by Modeling All-domain Movelines** [Industry]

Chen Gao (Alibaba Group); Zixin Zhao (Alibaba Group); Lv Shao (Alibaba Group); Tong Liu (Alibaba Group)

Click-Through Rate (CTR) prediction has long been dominated by discriminative paradigms that optimize local decision boundaries within candidate-specific subspaces. However, these models often fail to capture the global joint distribution and the continuous structural evolution of user intent across all-domain movelines. While generative approaches attempt to model global transition patterns, existing methods suffer from discretization-induced information collapse by remapping nuanced e-commerce signals into discrete linguistic or categorical spaces, failing to preserve the topological fidelity of interest trajectories. To overcome these limitations, we propose a novel generative pre-training paradigm that models user intent as a continuous evolutionary trajectory on a high-dimensional latent interest manifold, termed the Next Interest Flow (NIF). We introduce kinematic constraints to govern this flow: Interest Diversity is achieved via tangent space decomposition, while Evolution Velocity ensures trajectory smoothness through geodesic regularization. To bridge the objective mismatch between generative pre-training and discriminative fine-tuning, we propose a bidirectional alignment strategy to synchronize semantic spaces. Furthermore, we develop a Temporal Sequential Pairwise (TSP) mechanism to instill temporal causality within the discriminative framework. We present the All-domain Moveline Evolution Network (AMEN), a unified framework implementing this pipeline. Extensive experiments on a 6.7-billion instance industrial dataset and online A/B tests on Taobao validate AMEN's superiority, achieving +0.87pt AUC gain and +11.6% CTCVR lift.

P010 · **NBA-Net: Next-Behavior-Aware Network for Intent Recommendation** [Industry]

Haoxin Shen (Mashang Consumer Finance Co., Ltd.); Peng Ying (Mashang Consumer Finance Co., Ltd.); Wanjie Tao (Mashang Consumer Finance Co., Ltd.); Jie Liang (Mashang Consumer Finance Co., Ltd.); Quan Lu (Mashang Consumer Finance Co., Ltd.); Ning Jiang (Mashang Consumer Finance Co., Ltd.)

Intent recommendation serves as a precision conduit aligning user demands with content supply, shifting recommendations from passive matching to proactive need comprehension. It is widely used in Intelligent Customer Service, especially during session initiation and conversation conclusion. Existing methods predominantly exploit historical behaviors to infer user's current intent, yet overlook the influence of forthcoming user behaviors—since current intent being intrinsically coupled with the behaviors users are poised to perform. Motivated by this insight, we propose the Next-Behavior-Aware Network (NBA-Net), grounded in the accurate generation of the next-behavior and its efficient guidance of the main learning process. NBA-Net incorporates two core modules: (1) Next-Behavior Generation Module (NBGen), which mitigates insufficient behavioral intent mining via forward deduction and reverse tracing of behavioral migration vectors under closed-loop supervision. (2) Next-Behavior Guidance Module (NBGui), which progressively activates latent representations and adaptively integrates the next-behavior information to better align the model with user's true intent. Extensive offline experiments confirm the effectiveness of our model, while online A/B testing demonstrates a 5.6% relative improvement in Click-Through Rate.

P011 · **CMSL: Constructive Multi-Sequence Learning for Recommendation Systems** [Industry]

Zikun Cui (Meta Recommendation System); Renzhi Wu (Meta Recommendation System); Junjie Yang (Meta Recommendation System); Li Sheng (Meta Recommendation System); Jijie Wei (Meta Recommendation System); Linfeng Liu (Meta Recommendation System); Tai Guo (Meta Recommendation System); Tao Jia (Meta Recommendation System); Xiaodong Wang (Meta Pytorch); Hong Li (Meta Recommendation System); Li Yu (Meta FB Monetization); Sri Reddy (Meta Recommendation System); Hong Yan (Meta Recommendation System)

Sequence learning has emerged as the promising paradigm in recommendation systems, surpassing traditional Deep Learning Recommendation Models (DLRM) by capturing the temporal nuances of user behavior. However, current state-of-the-art architectures operate under a limiting analogy: they treat user history as a monolithic chronological sequence like a sentence in a Large Language Model (LLM). We observe a fundamental divergence between natural language and recommendation data: unlike the linear, logical flow of text, user history is inherently multi-faceted. A user's journey is a fragmented reflection of diverse interests, resulting in much weaker coherence between items than is found in LLM training data. This lack of structural unity leads to context pollution. In single-sequence modeling, unrelated behaviors compete for the same attention budget. This "noisy" signal dilutes the model's focus, effectively capping its ability to discern high-intent patterns from background activity. To address this, we propose Constructive Multi-Sequence Learning (CMSL), a paradigm shift from passive sequence ingestion to active "context engineering" that constructs multiple coherent sequences in latent space. CMSL leverages a learnable Sequence Construction Module to disentangle user history into "pure" thematic strands, followed by a linear attention mechanism to efficiently model these strands at scale. CMSL has been deployed across ranking and retrieval tasks and across four major surfaces at Meta.

P012 · **From Time and Place to Preference: LLM-Driven Geo-Temporal Context in Recommendations** [Industry]

Yejin Kim (Comcast Technology AI); Shaghayegh Agah (Comcast Technology AI); Neeraj Sharma (Comcast Technology AI); Mayur Nankani (Comcast Technology AI); Maria Peifer (Comcast Technology AI); Feifei Peng (Sky); H. Howie Huang (George Washington University); Sardar Hamidian (Comcast Technology AI)

Recommender systems focus on timestamps as numeric or cyclical values, often ignoring real-world context like seasons, holidays and events. We present a scalable framework that utilizes large language models (LLMs) to create geo-temporal embeddings from timestamps and coarse locations, capturing holidays, seasonal trends, and local/global events. We then introduce a geo-temporal embedding informativeness test as a lightweight diagnostic, demonstrating on MovieLens, LastFM, and a large-scale production dataset that these embeddings provide a predictive signal consistent with the outcomes of full model integrations. Geo-temporal embeddings were integrated into sequential models via feature fusion with metadata. Our findings underscore the importance of adaptive and hybrid strategies for improving recommendations. We also release a context-enriched MovieLens dataset \footnote{https://github.com/yejinjennyK/movielens-1m-geo-temporal-context\\https://huggingface.co/datasets/yejinjennyK/movielens-1m-geo-temporal-context}.

P013 · **CASE: Cadence-Aware Set Encoding for Large-Scale Next Basket Repurchase Recommendation** [Industry]

Yanan Cao (Walmart Global Tech); Ashish Ranjan (Walmart Global Tech); Sinduja Subramaniam (Walmart Global Tech); Evren Korpeoglu (Walmart Global Tech); Kaushiki Nag (Walmart Global Tech); Kannan Achan (Walmart Global Tech)

Repurchase behavior is a primary signal in large-scale retail recommendation, particularly in categories with frequent replenishment: many items in a user's next basket were previously purchased, and their timing follows stable, item-specific cadences. Yet most next basket repurchase recommendation models represent history as a sequence of discrete basket events indexed by visit order, which cannot explicitly model elapsed calendar time or update item rankings as days pass between purchases. We present CASE (Cadence-Aware Set Encoding) for next basket repurchase recommendation, which decouples item-level cadence learning from cross-item interaction, enabling explicit calendar-time modeling while remaining production-scalable. CASE represents each item's purchase history as a calendar-time signal over a fixed horizon, applies shared multi-scale temporal convolutions to capture recurring rhythms, and uses induced set attention to model cross-item dependencies with sub-quadratic complexity, allowing efficient batch inference at scale. Across three public benchmarks and a proprietary dataset, CASE consistently improves precision, recall, and NDCG at multiple cutoffs compared to strong next basket recommendation baselines. In a production-scale evaluation with tens of millions of users and a large item catalog, CASE achieves up to 8.6% relative precision lift and 9.9% relative recall lift at top-5, showing that scalable cadence-aware modeling yields measurable gains in both benchmark and industrial settings.

P014 · **Full Retraining, Incremental Fine-tuning, and Hybrid Serving: Model Updating and Serving for Industrial Generative Recommender Systems** [Industry]

Lu Fan (The Hong Kong Polytechnic University); Qijiong Liu (The Hong Kong Polytechnic University); Zhongzhou Liu (Ltd.); Guoyuan An (Ltd.); Wei Guo (Ltd.); Yong Liu (Ltd.); Xiao-Ming Wu (The Hong Kong Polytechnic University)

Generative recommendation casts recommendation as conditional sequence generation over text- or token-based representations and has shown strong promise in industrial systems. However, keeping such models up to date in dynamic environments is difficult: full retraining on sliding windows is expensive and slow, while incremental fine-tuning on recent data may introduce distributional bias and catastrophic forgetting. In this work, we study the model updating problem in industrial generative recommender systems and propose a practical hybrid serving framework as one possible solution under suitable conditions. The framework maintains a stable base model trained on long-term historical data, applies lightweight fine-tuning on recent interactions to obtain an updated model, and routes users with fresh behavioral signals to the updated model at serving time, while serving others with the base model. This design supports rapid adaptation for active users while preserving stability for the broader population at a substantially reduced retraining cost. We evaluate several updating and serving strategies, including full retraining, incremental fine-tuning, base-model serving, and hybrid serving, on both a public benchmark dataset and a large-scale industrial dataset. The results show that their effectiveness depends strongly on the distribution shift between the user cohort used for updating and the user cohort encountered at serving time. Under certain practically relevant conditions, the proposed hybrid serving framework provides a useful balance between adaptation, stability, and efficiency. The source code will be made available at https://github.com/Jyonn/Incremmender.

P015 · **RedGR: Unified Generative Retrieval for Recommendation in REDnote** [Industry]

Mengcheng Fang (Xiaohongshu Technology Co., Ltd.); Hongyu Wang (Xiaohongshu Technology Co., Ltd.); Xichuan Niu (Xiaohongshu Technology Co., Ltd.); Dong Xu (Xiaohongshu Technology Co., Ltd.); Honggang Wang (Xiaohongshu Technology Co., Ltd.); Tao Zhuang (Xiaohongshu Technology Co., Ltd.); Ping Yang (Xiaohongshu Technology Co., Ltd.); Yao Hu (Xiaohongshu Technology Co., Ltd.)

Recently, the generative retrieval paradigm has emerged as a transformative framework that significantly enhances the efficiency of large-scale industrial recommendation systems. This innovative approach systematically maps items to meaningful semantic identifiers (SIDs) and employs advanced sequence generation techniques to construct high-quality candidate sets, thereby enabling more accurate modeling of users' evolving interests and behavioral patterns. Nevertheless, two critical challenges remain inadequately addressed in current research: (1) Existing methodologies predominantly focus on modeling a single task such as predicting users' click behavior, while overlooking other tasks including predicting users' dwell-time and engagement behaviors, which are very important for video/content recommendation at the same time. The independent modeling of each task inevitably results in substantial computational overhead, thereby raising the pivotal question of whether the sophisticated multi-task learning capabilities inherent in LLMs can be effectively leveraged to achieve unified and efficient multi-task learning for generative retrieval. (2) The mapping mechanism from SIDs to concrete items requires substantial refinement to ensure precise and reliable retrieval performance. To tackle these issues, we propose RedGR, a generative retrieval model that unifies the modeling of multiple complex retrieval tasks. RedGR first applies the RQ-Kmeans algorithm to map items into SIDs, and then conducts pre-training on large-scale user behavior datasets to learn general knowledge. Then the RedGR model is finetuned on multi-task retrieval data with a unique instruction prompt for each task. This enables RedGR to generate the corresponding set of SIDs for each task. And the union of all sets of SIDs is the multi-task retrieval result. Finally, the Swing algorithm incorporates explicit, high-quality collaborative signals to strengthen the mapping from SIDs to specific items, thereby facilitating efficient retrieval of high-quality items. RedGR has been fully depolyed in the homefeed recommendation scenario of RedNote,serving hundreds of millions of users every day. Online A/B test results show a 0.178% increase in pagetime, a 0.734% increase in average user engagement, and a 0.076% growth in homefeed active users (FAU). These metrics collectively validate the superior performance of RedGR's unified retrieval modeling approach in complex multi-task scenarios.

P016 · **Semantic IDs for Recommender Systems at Snapchat: Use Cases, Technical Challenges, and Design Choices** [Industry]

Clark Mingxuan Ju (Snap. Inc); Tong Zhao (Snap Inc.); Leonardo Neves (Snap Inc.); Liam Collins (Snap Inc.); Bhuvesh Kumar (Snap Inc.); Jiwen Ren (Snap Inc.); Lili Zhang (Snap Inc.); Wenfeng Zhuo (Snap Inc.); Wen Zhang (Snap Inc.); Xiao Bai (Snap Inc.); Jinchao Li (Snap Inc.); Karthik Iyer (Snap Inc.); Zihao Fan (Snap Inc.); Yilun Xu (Snap Inc.); Yiwen Chen (Snap Inc.); Peicheng Yu (Snap Inc.); Manish Malik (Snap Inc.); Neil Shah (Snap Inc.)

Effective item identifiers (IDs) are an important component for recommender systems (RecSys) in practice, and are commonly adopted in many use cases such as retrieval and ranking. IDs can encode collaborative filtering signals within training data, such that RecSys models can extrapolate during the inference and personalize the prediction based on users' behavioral histories. Recently, Semantic IDs (SIDs) have become a trending paradigm for RecSys. In comparison to the conventional atomic ID, an SID is an ordered list of codes, derived from tokenizers such as residual quantization, applied to semantic representations commonly extracted from foundation models or collaborative signals. SIDs have drastically smaller cardinality than the atomic counterpart, and induce semantic clustering in the ID space. At Snapchat, we apply SIDs as auxiliary features for ranking models, and also explore SIDs as additional retrieval sources in different ML applications. In this paper, we discuss practical technical challenges we encountered while applying SIDs, experiments we have conducted, and design choices we have iterated to mitigate these challenges. Backed by promising offline results on both internal data and academic benchmarks as well as online A/B studies, SID variants have been launched in multiple production models with positive metrics impact.

P017 · **LLM-Enhanced Topical Trend Detection at Snapchat** [Industry]

Hangqi Zhao (Snap Inc.); Jay Li (Snap Inc.); Abhiruchi Bhattacharya (Snap Inc.); Cong Ni (Snap Inc.); Jason Yeung (Snap Inc.); Jinchao Ye (Snap Inc.); Kai Yang (Snap Inc.); Akshat Malu (Snap Inc.); Manish Malik (Snap Inc.)

Automatic detection of topical trends at scale is both challenging and essential for maintaining a dynamic content ecosystem on social media platforms. In this work, we present a large-scale system for identifying emerging topical trends on Snapchat, one of the world's largest short-video social platforms. Our system integrates multimodal topic extraction, time-series burst detection, and LLM-based consolidation and enrichment to enable accurate and timely trend discovery. To the best of our knowledge, this is the first published end-to-end system for topical trend detection on short-video platforms at production scale. Continuous offline human evaluation over six months demonstrates high precision in identifying meaningful trends. The system has been deployed in production at global scale and applied to downstream surfaces including content ranking and search, driving measurable improvements in content freshness and user experience.

P018 · **OxygenREC: An Instruction-Following Generative Framework for E-commerce Recommendation** [Industry]

Zhiwei Zhang (JD.com); Qingyang Li (JD.com); Ming Zhang (JD.com); Yanchen Qiao (JD.com); Zhen Wang (JD.com); Ziyang Ji (JD.com); Xiangyu Qian (JD.com); Shijie Yang (JD.com); Yanlong Zang (JD.com); Weijie Ding (JD.com); Zhi Ma (JD.com); Zhen Li (JD.com); Yaqiang Zang (JD.com); Pinghua Gong (JD.com)

Traditional recommendation systems suffer from multi-stage objective misalignment. Generative recommendation offers end-to-end training, yet existing methods are largely inductive and miss complex intents that require world-knowledge-driven deductive reasoning. LLMs reason well but are impractical online due to latency and cost, and multi-scenario settings still demand separate training and deployment. We present OxygenREC, an industrial Fast-Slow Thinking recommender: a near-line LLM pipeline produces Contextual Reasoning Instructions, while a high-efficiency encoder-decoder backbone serves fast generation. We further introduce Instruction-Guided Retrieval (IGR) and a Query-to-Item (Q2I) loss for controllable semantic alignment. Post-training combines Monte Carlo Tree Search (MCTS)-based candidate search with Joint Pareto Optimization (JPO)-based cross-scenario policy alignment. For online deployment, our method delivers consistent Gross Merchandise Value (GMV) and order gains across core scenarios.

P019 · **OCP: Orthogonal Constrained Projection for Sparse Scaling in Industrial Commodity Recommendation** [Industry]

Chen Sun (JD.com); Beilin Xu (JD.com); Boheng Tan (JD.com); Jiacheng Wang (JD.com); Yuefeng Sun (JD.com); Rite Bo (JD.com); Ying He (JD.com); Yaqiang Zang (JD.com); Pinghua Gong (JD.com)

In industrial commodity recommendation systems, the representation quality of Item-Id vocabularies directly impacts the scalability and generalization ability of recommendation models. A key challenge is that traditional Item-Id vocabularies, when subjected to sparse scaling, suffer from low-frequency information interference, which restricts their expressive power for massive item sets and leads to representation collapse. To address this issue, we propose an Orthogonal Constrained Projection method to optimize embedding representation. By enforcing orthogonality, the projection constrains the backpropagation manifold, aligning the singular value spectrum of the learned embeddings with the orthogonal basis. This alignment ensures high singular entropy, thereby preserving isotropic generalized features while suppressing spurious correlations and overfitting to rare items. Empirical results demonstrate that OCP accelerates loss convergence and enhances the model's scalability; notably, it enables consistent performance gains when scaling up dense layers. Large-scale industrial deployment on JD.com further confirms its efficacy, yielding a 12.97\% increase in UCXR and an 8.9\% uplift in GMV, highlighting its robust utility for scaling up both sparse vocabularies and dense architectures.

P020 · **A Cascaded Generative Approach for e-Commerce Recommendations** [Industry]

Moein Hasani (Instacart); Hamidreza Shahidi (Instacart); Trace Levinson (Instacart); Yuan Zhong (The Pennsylvania State University); Guanghua Shu (EvenUp); Vinesh Gudla (Ambience Healthcare); Tejaswi Tenneti (Ambience Healthcare)

Personalized storefronts in large e-commerce marketplaces are often assembled from many independent components: static themes per page section (''placement''), retrieval systems to fetch eligible products per placement, and pointwise rankers to order content. While effective in optimizing for aggregate preferences, this paradigm is rigid and can limit personalization and semantic cohesion across the page. This makes it poorly suited to support dynamic objectives and merchandising requirements over time. To address this, we introduce a cascaded merchandising framework that decomposes storefront construction into two generative tasks: (i) placement-level theme generation and (ii) constrained keyword generation per placement to power product retrieval. Teacher–student fine-tuning is leveraged to improve scalability of this framework under production latency and cost constraints. Fine-tuned model ablations are shown to approach closed-weight LLM performance. We further contribute frameworks for AI-driven content evaluation and quality filtering, enabling safe and automated deployment of dynamic content at scale. Generative output is fused with traditional ranking models to preserve hybrid infrastructure. In online experiments, this framework yields an estimated +2.7% lift in cart adds per page view over a strong baseline.

P021 · **SIGMA: A Semantic-Grounded Instruction-Driven Generative Multi-Task Recommender at AliExpress** [Industry]

Yang Yu (Alibaba International Digital Commercial Group); Lei Kou (Alibaba International Digital Commercial Group); Huaikuan Yi (Alibaba International Digital Commercial Group); Bin Chen (Alibaba International Digital Commercial Group); Yayu Cao (Alibaba International Digital Commercial Group); Lei Shen (Alibaba International Digital Commercial Group); Chao Zhang (Alibaba International Digital Commercial Group); Bing Wang (Alibaba International Digital Commercial Group); Xiaoyi Zeng (Alibaba International Digital Commercial Group)

With the rapid evolution of Large Language Models (LLMs), generative recommendation is gradually reshaping the paradigm of recommender systems. However, most existing methods remain confined to the interaction-driven next-item prediction paradigm, struggling to keep pace with the latest evolving trends or address the diverse recommendation tasks along with business-specific requirements in real-world scenarios. To this end, we present SIGMA, a Semantic-Grounded Instruction-Driven Generative Multi-Task Recommender deployed at AliExpress. Specifically, we first ground item entities in a unified latent space capturing both general semantics and collaborative signals. Building upon this, we introduce a hybrid item tokenization method for both precise modeling and efficient generation. Moreover, we construct a large-scale multi-task supervised fine-tuning dataset empowering SIGMA to fulfill various recommendation demands via instruction-following. Finally, we design a three-step item generation procedure integrated with an adaptive probabilistic fusion

mechanism to calibrate the output distributions based on task-specific requirements for recommendation accuracy and diversity. Extensive offline experiments and online A/B tests demonstrate the effectiveness of SIGMA across various real-world recommendation tasks.

P022 · **Breaking the Relevance–Diversity Seesaw: Hierarchical LLM Reasoning with RL for Industrial Novelty Recommendation** [Industry]

Ying Sun (Harbin Institute of Technology); Yanyan Zou (JD.COM); Xiao Wang (Harbin Institute of Technology); Hanchuan Xu (Harbin Institute of Technology); Xuanhua Yang (JD.COM); Sulong Xu (JD.COM); Junbo Qi (Waseda University); Shengjie Li (JD.COM)

Novelty recommendation sustains long-term user engagement by exposing users to content that is both relevant and meaningfully different from their recent consumption. In large-scale e-commerce, this requires composing coherent yet non-redundant recommendation lists, a task fundamentally constrained by the relevance-diversity trade-off. Large language models (LLMs) offer a unified generative paradigm for inferring user intent and producing semantically coherent candidates, yet industrial deployment faces two critical challenges: (i) scarce supervision for modeling novelty transitions and diversity-aware list construction, and (ii) reward granularity mismatch, where standard RL assigns coarse sequence-level rewards that fail to capture item-level redundancy and complementarity. We present BALANCE, a hierarchical reasoning-and-generation framework that decomposes novelty recommendation into three structured stages: generating a Novelty Tag for exploration direction, refining an Interest Topic for intent specification, and constructing a Recommendation List for facet coverage. We address data scarcity through a self-reflection pipeline that synthesizes high-quality supervision by integrating real behavior logs with structured rationales. We resolve granularity mismatch through Sequence-Item Policy Optimization (SIPO), which jointly optimizes sequence- and item-level objectives via granularity-aware advantage fusion. Extensive offline experiments and online A/B test on the JD.com recommender system, validate the performance of our method, highlighting its superior novelty and diversity without compromising relevance.

P023 · **GenRec: A Preference-Oriented Generative Framework for Large-Scale Recommendation** [Industry]

Yanyan Zou (JD.com); Junbo Qi (Waseda University); Lunsong Huang (JD.com); Yu Li (JD.com); Kewei Xu (JD.com); Jiahao Gao (JD.com); Binglei Zhao (JD.com); Xuanhua Yang (JD.com); Sulong Xu (JD.com); Shengjie Li (JD.com)

Generative Retrieval (GR) offers a promising paradigm for recommendation through next-token prediction (NTP). However, scaling it to large-scale industrial systems introduces three challenges: (i) within a single request, the identical model inputs may produce inconsistent outputs due to the pagination request mechanism; (ii) the prohibitive cost of encoding long user behavior sequences with multi-token item representations based on semantic IDs, and (iii) aligning the generative policy with nuanced user preference signals. We present GenRec, a preference-oriented generative framework deployed on the JD App https://www.jd.com that addresses above challenges within a single decoder-only architecture. For training objective, we propose Page-wise NTP task, which supervises over an entire interaction page rather than each interacted item individually, providing denser gradient signal and resolving the one-to-many ambiguity of point-wise training. On the prefilling side, an asymmetric linear Token Merger compresses multi-token Semantic IDs in the prompt while preserving full-resolution decoding, reducing input length by ~2× with negligible accuracy loss. To further align outputs with user satisfaction, we introduce GRPO-SR, a reinforcement learning method that pairs Group Relative Policy Optimization with NLL regularization for training stability, and employs Hybrid Rewards combining a dense reward model with a relevance gate to mitigate reward hacking. In month-long online A/B tests serving production traffic, GenRec achieves 9.5% improvement in click count and 8.7% in transaction count over the existing pipeline.

P024 · **RAD-DPO: Robust Adaptive Denoising Direct Preference Optimization for Generative Retrieval in E-commerce** [Industry]

Zhiguo Chen (JD.com); Guohao Sun (Peking University); Yiming Qiu (JD.com); Xingzhi Yao (JD.com); Mingming Li (Institute of Information Engineering■Chinese Academy of Sciences); Huimu Wang (JD.com); Yangqi Zhang (JD.com); Songlin Wang (JD.com); Sulong Xu (JD.com)

Generative Retrieval (GR) is rapidly transforming e-commercesearch by replacing traditional multi-stage pipelines with the autoregressivedecoding of structured Semantic IDs (SIDs). Despitethis architectural efficiency, aligning GR models with nuanced, realworlduser preferences remains a critical challenge. While DirectPreference Optimization (DPO) offers an efficient alignment solution,its direct application to structured SIDs suffers from threelimitations: (i) it penalizes shared hierarchical prefixes, causinggradient conflicts; (ii) it is vulnerable to noisy pseudo-negativesfrom implicit feedback; and (iii) in multi-label queries with multiplerelevant items, it exacerbates a probability "squeezing effect" amongvalid candidates. To address these issues, we propose RAD-DPO,which introduces token-level gradient detachment to protect prefixstructures, similarity-based dynamic reward weighting to mitigatelabel noise, and a multi-label global contrastive objective integratedwith global SFT loss to explicitly expand positive coverage. Extensiveoffline evaluations and large-scale online A/B testing onJD.com's core search engine demonstrate that RAD-DPO achievessignificant improvements in both retrieval precision and training efficiency,proving its robustness for massive industrial deployments

P025 · **Bridging the Gap: Generative Retrieval via Query-to-Multi-Span Framework for Effective E-commerce Search** [Industry]

Huimu Wang (Institute of Automation, Chinese Academy of Sciences); Yiming Qiu (JD); Xingzhi Yao (JD); Guangtao Nie (JD); Zuxu Chen (Tsinghua University); Zhenlin He (Shangshi Aero Engine Co., Ltd); Songlin Wang (JD); Guoyu Tang (JD); Sulong Xu (JD); Jingwei Zhuo (JD); Mingming Li (Institute of Information Engineering, Chinese Academy of Sciences)

Generative retrieval formulates document retrieval as an identifiergeneration task. While prevailing methods increasingly adopt SemanticIDs (SIDs), their opaque nature and rigid mappings strugglewith the dynamic inventory and strict interpretability requirementsof E-commerce search. Furthermore, generating accurate targetsfrom brief queries against noisy, loosely structured item titles remainsa practical challenge. To address these issues, we propose aQuery-to-Multi-Span generative retrieval framework tailored forE-commerce. Instead of relying on opaque SIDs or raw titles, ourmethod simplifies the process by generating interpretable multispanidentifiers from queries. We align the autoregressive modelwith user preferences using click logs, and employ a constraintbasedbeam search to isolate key spans for final item retrieval.This approach explicitly bridges generative models with robustconstraint matching, ensuring both matching accuracy and transparency.Extensive offline evaluations demonstrate competitiveretrieval performance, and online A/B tests confirm its effectivenessin delivering measurable conversion gains in a productionenvironment.

P026 · **Towards Efficient and Generalizable Retrieval: Adaptive Semantic Quantization and Residual Knowledge Transfer** [Industry]

Huimu Wang (Institute of Automation, Chinese Academy of Sciences); Xingzhi Yao (JD.com); Yiming Qiu (JD.com); Qinghong Zhang (JD.com); Haotian Wang (Harbin Institute of Technology); Yufan Cui (Peking University); Songlin Wang (JD.com); Sulong Xu (JD.com); Mingming Li (Institute of Information Engineering■Chinese Academy of Sciences)

While semantic ID-based generative retrieval enables efficient end-to-end modeling in industrial applications, these methods face a persistent trade-off. On one hand, data-rich head items often suffer from ID collisions, which blur their distinct features and degrade downstream tasks. On the other hand, data-sparse tail items especially cold-start items are prone to semantic fragmentation during quantization; they are often mapped as isolated discrete points, which severely hinders their ability to generalize. To address this issue, we propose the Anchored Curriculum with Sequential Adaptive Quantization (SA2CRQ) framework. The framework introduces Sequential Adaptive Residual Quantization (SARQ) to dynamically allocate code lengths based on item path entropy, assigning longer, discriminative IDs to head items and shorter, generalizable IDs to tail items. To mitigate data sparsity, the Anchored Curriculum Residual Quantization (ACRQ) component utilizes a frozen semantic manifold learned from head items to regularize and accelerate the representation learning of tail items. Experimental results from a large-scale industrial search system and multiple public datasets indicate that SA2CRQ yields consistent improvements over existing baselines, particularly in cold-start retrieval scenarios.

P027 · **MMRM: A Multiplex Multimodal Representation Model for Product Ranking in E-commerce Search** [Industry]

Zhen-Lin Chen (JD.COM); Maosen Sheng (JD.COM); Peng Lin (JD.COM); Jianmin Chen (JD.COM); Zhuojian Xiao (JD.COM); Dongyue Wang (JD.COM); Xiwei Zhao (JD.COM)

Multimodal information is pivotal for e-commerce search ranking. Existing works leverage multimodal data typically by fine-tuning general Multimodal Large Language Models (MLLMs) via collaborative signals, subsequently integrating the derived representations into ranking models as item features. Despite their efficacy, these methods face two primary limitations: (1) they rely on a single collaborative signal for MLLM fine-tuning, failing to exploit the heterogeneous signals essential for multitask ranking; and (2) they treat multimodal representations as regular item features in ranking models, underutilizing their latent potential for user behavior modeling. To address these challenges, we propose the Multiplex Multimodal Representation Model (MMRM), a unified framework that aligns MLLMs with diverse collaborative signals. By employing a shared backbone with task-specific tokens and projection layers, MMRM simultaneously learns from multiple signals and generates comprehensive multiplex item representations in a single inference pass. Furthermore, we introduce a multiplex user representation strategy in ranking models, which derives task-specific user representations via search-based behavior sequence modeling leveraging multiplex item representations. Extensive experiments demonstrate MMRM's superior efficiency and effectiveness. Notably, MMRM has been successfully deployed in the JD e-commerce search engine, yielding significant performance gains for millions of daily users.

P028 · **Query-Attention Dual-Stream Framework with Cross-Category Transfer for Efficient Fine-Grained Interest Pre-Ranking** [Industry]

Huimu Wang (Institute of Automation■Chinese Academy of Sciences); Xujun Liu (JD.com); Yiming Qiu (JD.com); Zhenlin He (Shangshi Aero Engine Co., Ltd); Enqiang Xu (JD.com); Yihao Wang (JD.com); Jinyuan Zhao (JD.com); Guangtao Nie (JD.com); Songlin Wang (JD.com); Mingming Li (Institute of Information Engineering■Chinese Academy of Sciences)

Large-scale search and recommendation systems typically adopt a cascaded architecture of retrieval, pre-ranking, ranking, and re-ranking to balance efficiency and accuracy. However, pre-ranking still faces challenges of behavioral sparsity, limited interest diversity, and computational latency. We

propose the Query-Attention Dual-Stream (QADS) framework to address these issues. QADS partitions user behaviors into strongly and weakly correlated streams and further decomposes them into fine-grained subsequences guided by domain knowledge. A query-centric attention mechanism reduces complexity from O(N) to O(1), enabling efficient inter- and intra-sequence modeling. A contrastive cross-category transfer module propagates dense patterns from weakly correlated to sparse domains, while a latency-aware parallel inference architecture further reduces delay by 36%. Experiments on public and industrial datasets show that QADS delivers significant performance improvements and has been successfully deployed in large-scale e-commerce search systems.

P029 · **M²GR: Generative User Interest Modeling via Multi-Granularity Multi-Objective CoT for Industrial Recommendation** [Industry]

Yuqi Zhang (Jingdong Group); Jingwen Shi (Jingdong Group); Wen Shi (Jingdong Group); Jia Li (Jingdong Group); Zhen Chen (Jingdong Group); Jing Zhou (Jingdong Group); Dongyue Wang (Jingdong Group); Xiwei Zhao (Jingdong Group); Sulong Xu (Jingdong Group)

User interest modeling plays a vital role in industrial recommendation systems (RSs). Existing generative recommendation (GR) methods rely on single-step direct inference, which falls short of deeply modeling complex and dynamically evolving user interest. Recent chain-of-thought (CoT)-based GR methods attempt to address this, but either suffer from information loss during semantic space transformation in explicit reasoning or yield uncontrollable, homogeneous reasoning chains in implicit reasoning. To this end, we propose M2GR, a Generative user interest modeling method using a Multi-granularity Multi-objective CoT for industrial RSs. The multi-granularity chain first predicts coarse-grained interests in parallel (e.g., item attributes like brand and category) and then refines them into fine-grained interests, i.e., the exact next click item. Building on this click interest, the multi-objective chain further predicts more complex user order interest. This stepwise implicit reasoning, conducted entirely in the item space, preserves complete user behavior information while enhancing controllability and diversity in interest prediction. Comprehensive offline experiments and online A/B testing demonstrate that M2GR achieves superior performance, while exhibiting scalability and plug-and-play deployability in industrial-scale RSs.

P030 · **SID-Coord: Coordinating Semantic IDs for ID-based Ranking in Short-Video Search** [Industry]

Guowen Li (Kuaishou Technology); Yuepeng Zhang (Kuaishou Technology); Shunyu Zhang (Kuaishou Technology); Yi Zhang (Kuaishou Technology); Xiaoze Jiang (Kuaishou Technology); Yi Wang (Kuaishou Technology); Jingwei Zhuo (unaffiliated)

Large-scale short-video search ranking models are typically trained on sparse co-occurrence signals over hashed item identifiers (HIDs). While effective at memorizing frequent interactions, such ID-based models struggle to generalize to long-tailed items with limited exposure. This memorization–generalization trade-off remains a longstanding challenge in such industrial systems. We propose SID-Coord, a lightweight Semantic ID framework that incorporates discrete, trainable semantic IDs (SIDs) directly into ID-based ranking models. Instead of treating semantic signals as auxiliary dense features, SID-Coord represents semantics as structured identifiers and coordinates HID-based memorization with SID-based generalization within a unified modeling framework. To enable effective coordination, SID-Coord introduces three components: (1) an attention-based fusion module over hierarchical SIDs to capture multi-level semantics, (2) a target-aware HID–SID gating mechanism that adaptively balances memorization and generalization, and (3) a SID-driven interest alignment module that models the semantic similarity distribution between target items and user histories. SID-Coord can be integrated into existing production ranking systems without modifying the backbone model. Online A/B experiments in a real-world production environment show statistically significant improvements, with a +0.664% gain in long-play rate in search and a +0.369% increase in search playback duration.

P031 · **SAER: Scalable Assessment of E-commerce Recommendations using Large Language Models** [Industry]

Xiang Liu (Amazon); Manoj Srivatsav Regulagedda (Amazon); Sapan Patel (Amazon); Walter Wong (Amazon)

Selecting which recommendation algorithm variant to advance to online experimentation is a critical decision in industry practice. Manual evaluation is subjective and time-consuming, while offline metrics such as nDCG often fail to correlate with real-world customer preferences. We present SAER, a two-stage framework (pointwise filtering and pairwise comparison) that uses Large Language Models as judges to evaluate e-commerce recommendation quality. For pointwise evaluation, SAER achieves higher agreement with human consensus ($\kappa_W = 0.58$) than human annotators achieve with each other ($\kappa_W = 0.38$). For pairwise evaluation, SAER mitigates position bias (88.0% consistency) and demonstrates strong adversarial robustness (91.0% detection rate). In a large-scale online case study serving over 10M impressions, the algorithm SAER preferred offline also achieved a statistically significant click-through rate (CTR) lift ($p < 0.0001$), providing encouraging directional evidence, though this single observation cannot establish a general predictive relationship. SAER produces reproducible, explainable signals in approximately 30 minutes for $28, positioning it as a scalable pre-screening layer for recommendation development.

P032 · **A General Framework for Multimodal LLM-Based Multimedia Understanding in Large-Scale Recommendation Systems** [Industry]

Yiming Zhu (Meta Platforms); Xu Liu (Meta Platforms); Ziyun Xu (Meta Platforms); Zheng Wu (Meta Platforms); Joena Zhang (Meta Platforms); Sirius Chen (Meta Platforms); Chenheli Hua (Meta Platforms); Silvester Yao (Meta Platforms); Qichao Que (Meta Platforms); Wentao Shi (Meta Platforms); Junfeng Pan (Meta Platforms); Linhong Zhu (Meta Platforms)

Conventional recommendation systems frequently fail to fully exploit the high-dimensional semantic signals inherent in multimedia content, thereby limiting the fidelity of user preference modeling. While Multimodal Large Language Models (MM-LLMs) offer robust mechanisms for interpreting such complex data, their integration into latency-constrained, industrial-scale architectures remains a significant challenge. To address this, we propose a generalized framework for MM-LLM-driven multimedia understanding. Our methodology employs a tripartite architecture encompassing content interpretation, representation extraction, and systematic pipeline integration, instantiated via a LLaMA2-based model that generates descriptive captions subsequently ingested as tokenized categorical features. Empirical evaluation demonstrates the efficacy of this approach, yielding a 0.35% increase in offline AUC and a 0.02% improvement in online metrics at scale, substantiating the practical viability of leveraging MM-LLMs to enhance large-scale recommendation performance.

P033 · **COMET: Compatibility-Oriented Multi-modal Embedding Transformer for Visual Recommendations** [Industry]

Dween Rabius Sanny (Amazon); Prateek Sircar (Amazon); Deepak Gupta (Amazon)

Recommending visually compatible products in fashion and interior design is a significant challenge, as compatibility rules are nuanced, context-dependent, and reliant on fine-grained details that traditional models fail to capture. Existing methods often struggle with heterogeneous compatibility rules (e.g., sofa-table vs. sofa-curtain) and an over-reliance on global visual features, missing critical textual cues like style or material. To address these limitations, we introduce COMET (Compatibility-Oriented Multi-modal Embedding Transformer), a scalable, vision-language framework for visual recommendations. COMET replaces rigid, category-specific subspaces with an attribute-conditioned cross-attention mechanism, reframing compatibility as a conditional retrieval problem. By formulating a textual compatibility prompt that encodes relational context and structured attributes, COMET dynamically conditions which visual features are attended to, producing a joint, context-aware representation. This multi-modal approach allows the model to leverage descriptive text (e.g., "mid-century modern" or "oak finish") to understand visually ambiguous stylistic nuances. The model is trained using a triplet loss with hard negative sample strategy to effectively distinguish between compatible and incompatible item pairs. COMET was evaluated on established benchmarks for both fashion (Polyvore) and furniture (Bonn Furniture), in addition to our in-house datasets.

P034 · **CSMAD: Hallucination Detection via Multi-Agent Debate with NLI-Verified Contradictory Statements** [Industry]

Swapnil Gupta (Amazon); Akshay Verma (Amazon); Khushi Gupta (Amazon); Prateek Sircar (Amazon); Deepak Gupta (Amazon)

Large Language Models (LLMs) are prone to hallucinations, producing fluent but factually incorrect statements. Recent multi-agent debate methods improve hallucination detection by jointly improving reasoning and decision-making. However, existing approaches either collaborate which amplifies shared overconfidence, or adopt adversarial preset stances, that can inject incorrect information complicating decision making. To address this, we propose Contradictory Statement Multi-Agent Debate (CSMAD), a multi-agent framework that creates structured disagreement by generating a contradictory claim for each input claim. CSMAD asks independent agents to evaluate the claim and the contradictory claim, which encourages different lines of reasoning without assigning preset stances. When the outcome is non-discriminative; both the contradictory statements are either accepted or rejected; the agents exchange rationales and update their judgments after considering opposing evidence. A final judge then decides the truth of the original claim, using both arguments as context. To make contradictory statement generation reliable, we add a Natural Language Inference (NLI) based verifier that checks whether the generated statement actually contradicts the original claim; if it does not, the system falls back to an explicit negation-based contradiction. Across public benchmarks for question answering and scientific claim verification, as well as a proprietary e-commerce claims dataset, we show that CSMAD consistently outperforms the strongest baseline for both large (Claude-3.5 Sonnet) and medium-sized (Qwen3-8B) language models, improving F1 by +2.3 and +4.1 points, respectively, while reducing LLM token cost by 28%.

P035 · **Don't Contrast the Impossible: Region-Constrained Batching for Contrastive User Modeling on a Local Community Platform** [Industry]

Seungho Han (Danggeun Market Inc. (Karrot)); Byeongchang Kim (Danggeun Market Inc. (Karrot)); Jin Yu (Danggeun Market Inc. (Karrot))

Contrastive learning is widely used for user modeling in large-scale recommender systems, where standard in-batch negatives implicitly assume universal exposure that any user can be shown any item. On local community platforms such as Karrot, however, exposure is geographically constrained; many user-item pairs are impossible by design yet still treated as negatives during training, diluting the contrastive learning signal. We address this impossible negatives problem and propose Region-Constrained Batch Sampling (RCBS), a simple yet effective batching method that constructs region-homogeneous mini-batches so that users are contrasted primarily against items they could feasibly see. By replacing impossible

negatives with feasible ones, RCBS naturally introduces harder and more informative negatives under realistic exposure constraints. With offline evaluations and online A/B tests, we show that RCBS consistently improves user representation quality and consequently enhances home feed ranking, retrieval, and display ads ranking. The resulting user embeddings have been deployed in production across various applications.

P036 · **CS3: Efficient Online Capability Synergy for Two-Tower Recommendation** [Industry]

Lixiang Wang (Kuaishou Technology); Shaoyun Shi (Kuaishou Technology); Peng Wang (Kuaishou Technology); Wenjin Wu (Kuaishou Technology); Peng Jiang (Kuaishou Technology)

To balance effectiveness and efficiency in recommender systems, multi-stage pipelines commonly use lightweight two-tower models for large-scale candidate retrieval. However, the isolated two-tower architecture restricts representation capacity, embedding-space alignment, and cross-feature interactions. Existing solutions such as late interaction and knowledge distillation can mitigate these issues, but often increase latency or are difficult to deploy in online learning settings. We propose Capability Synergy (CS3), an efficient online framework that strengthens two-tower retrievers while preserving real-time constraints. CS3 introduces three mechanisms: (1) Cycle-Adaptive Structure for self-revision via adaptive feature denoising within each tower; (2) Cross-Tower Synchronization to improve alignment through lightweight mutual awareness between towers; and (3) Cascade-Model Sharing to enhance cross-stage consistency by reusing knowledge from downstream models. CS3 is plug-and-play with diverse two-tower backbones and compatible with online learning. Experiments on three public datasets show consistent gains over strong baselines, and deployment in a large-scale advertising system yields up to 8.36% revenue improvement across three scenarios while maintaining ms-level latency.

P037 · **Request-Only Optimization for Recommendation Systems** [Industry]

Liang Guo (Meta Platforms, Inc.); Wei Li (Meta Platforms, Inc.); Lucy Liao (Meta Platforms, Inc.); Huihui Cheng (Meta Platforms, Inc.); Rui Zhang (Meta Platforms, Inc.); Yu Shi (Meta Platforms, Inc.); Yueming Wang (Meta Platforms, Inc.); Yanzun Huang (Meta Platforms, Inc.); Keke Zhai (Meta Platforms, Inc.); Pengchao Wang (Meta Platforms, Inc.); Timothy Shi (Meta Platforms, Inc.); Xuan Cao (Meta Platforms, Inc.); Shengzhi Wang (Meta Platforms, Inc.); Renqin Cai (Meta Platforms, Inc.); Zhaojie Gong (Meta Platforms, Inc.); Omkar Vichare (Meta Platforms, Inc.); Rui Jian (Meta Platforms, Inc.); Leon Gao (Meta Platforms, Inc.); Shiyan Deng (Meta Platforms, Inc.); Xingyu Liu (Meta Platforms, Inc.); Xiong Zhang (Meta Platforms, Inc.); Fu Li (Meta Platforms, Inc.); Wenlei Xie (Meta Platforms, Inc.); Bin Wen (Meta Platforms, Inc.); Rui Li (Meta Platforms, Inc.); Lu Fang (Meta Platforms, Inc.); Xing Liu (Meta Platforms, Inc.); Jiaqi Zhai (Meta Platforms, Inc.)

\begin{abstract}Recommendation systems represent one of the largest machine learning applications on the planet -- industry-scale recommendation models are trained with petabytes of data and serve billions of users every day. To utilize the rich user signals in the long user history, these models have been scaled up to unprecedented complexity, up to trillions of floating-point operations (TFLOPs) per example. This scale, coupled with the huge amount of training data, necessitates new storage and training algorithms to efficiently improve the quality of these complex recommendation systems. In this paper, we present a Request-Only Optimizations (ROO) training and modeling paradigm. ROO simultaneously improves the storage and training efficiency as well as the model quality of recommendation systems. We holistically approach this challenge through co-designing data, infrastructure, and model architectures. Our ROO training and modeling paradigm treats a user request as a unit of the training data. Compared with the established practice of treating a user impression as a unit, this new design directly removes redundant features in data logging, saving data storage. Second, by de-duplicating computations and communications across multiple impressions in a request, this new paradigm enables highly scaled-up neural network architectures to better capture user interest signals, such as Generative Recommenders (GRs) and other request-only friendly architectures.Our proposed ROO training and modeling paradigm has been deployed to multiple major recommendation products, each with billions of active users. ROO data format allows for increasing the training sample volume up to 2.5x. ROO training optimization yields a substantial increase in training throughput -- up to 6.7x. Combined with ROO-based neural architectures like Hierarchical Sequential Transduction Units (HSTU), ROO scales model FLOPs by 7x utilizing the same amount of training compute, enabling significant offline and online metric wins. \end{abstract}

P038 · **ReST: A Plug-and-Play Spatially-Constrained Representation Enhancement Framework for Local-Life Recommendation** [Industry]

Hao Jiang (Kuaishou Technology); Long Zhang (Kuaishou Technology); Guoquan Wang (Kuaishou Technology); Sheng Yu (Kuaishou Technology); Yang Zeng (Kuaishou Technology); Wencong Zeng (Kuaishou Technology); Fei Pan (Kuaishou Technology); Peng Jiang (Kuaishou Technology); Guorui Zhou (Kuaishou Technology)

Local-life recommendation provides users with convenient access to daily essentials, yet faces two key challenges: (1) spatial constraints, where items are exposed only within limited geographic areas, and (2) long-tail sparsity, where popular items dominate user interactions. Existing methods predominantly adopt a user-centric view, often failing to address the representation degradation of long-tail items caused by these spatial limitations. We argue that an item-centric perspective is better suitable for this domain, as it focuses on enhancing long-tail item representations to align with the spatially-constrained characteristics of local lifestyle services. To this end, we propose ReST, a plug-and-play framework for Spatially-Constrained Representation Enhancement. ReST initializes item embeddings via attribute semantics and utilizes a spatially-constrained contrastive learning mechanism. This design adaptively strengthens weak ID representations by aligning them with spatial characteristics, while preserving performance on popular items. Experiments on two real-world datasets and online A/B testing demonstrate that ReST significantly improves performance and exhibits strong plug-and-play flexibility in industrial applications.

P039 · **RAG-Enhanced Large Language Models for Dynamic Content Expiration Prediction in Web Search** [Industry]

Tingyu Chen (Baidu Inc.); Wenkai Zhang (Baidu Inc.); Li Gao (Baidu Inc.); Lixin Su (Baidu Inc.); Ge Chen (Baidu Inc.); Dawei Yin (Baidu Inc.); Daiting Shi (Baidu Inc.)

In commercial web search, aligning content freshness with user intent remains challenging due to the highly varied lifespans of information. Traditional industrial approaches rely on static time-window filtering, resulting in "one-size-fits-all" rankings where content may be chronologically recent but semantically expired. To address this limitation, we present a novel Large Language Models (LLMs)-based Query Aware Dynamic Content Expiration Prediction Framework deployed in Baidu search, reformulating timeliness as a dynamic validity inference task. Our framework extracts fine-grained temporal contexts from documents and leverages LLMs to deduce a query-specific "validity horizon", a semantic boundary defining when information becomes obsolete based on user intent. Integrated with robust hallucination mitigation strategies to ensure reliability, our approach has been evaluated through offline and online A/B testing on live production traffic. Results demonstrate significant improvements in search freshness and user experience metrics, validating the effectiveness of LLM-driven reasoning for solving semantic expiration at an industrial scale.

P040 · **Population-Guided Intent-Aware Query Rewriting for Web Search** [Industry]

Yuanzhao Guo (School of Artificial Intelligence, Jilin University); Shuai Zhang (Baidu Inc); Anqi Li (Baidu Inc); Jiaming Zhang (Baidu Inc); Wei Li (Baidu Inc); Shihao Liu (Baidu Inc); Daiting Shi (Baidu Inc); Yuan Tian (School of Artificial Intelligence, Jilin University)

Query rewriting is a core component of web search, yet traditional methods mainly rely on large language model (LLM) prompting, fine-tuning, or personalized rewriting based on user history. These approaches often overlook population-level intent signals in large-scale query logs, leading to misalignment with mainstream search intent. Moreover, although large models achieve high rewriting quality, their computational demands and deployment complexity limit industrial applicability. To address this, we propose Population-Guided Intent-Aware Rewriting (PGIR), which captures dominant population intent via a Semantic Clustering Unit (SCU) and generates intent-aware rewrites through a Rewriting Unit (RU), aligning queries with mainstream search goals without requiring user history.Building upon PGIR, we further introduce PGIR-DPA, a dual-phase adaptation strategy that transfers capabilities from a teacher LLM to lightweight student models, achieving high-quality rewriting while ensuring industrial scalability. Extensive offline and online experiments, including A/B testing on Baidu Search, show substantial improvements in rewrite quality and a 6.28% relative increase in user satisfaction. The framework has been fully deployed in Baidu's production search system, operating stably at scale, validating its industrial feasibility and commercial value.

P041 · **Enhancing Multi-Valued Treatment Uplift Modeling with Knowledge Sharing and RCT Data** [Industry]

Jiaxiang Liu (Tsinghua University); Luo He (Kuaishou Technology); Zhenghao Zeng (Stanford University); Deqiang Kong (Kuaishou Technology); Funan Mu (Kuaishou Technology); Jiaming Zhang (Kuaishou Technology); Zilong Lu (Kuaishou Technology); Peng Jiang (Kuaishou Technology); Hongyan Liu (Tsinghua University)

Uplift modeling estimates individual-level treatment effects for personalized decisions in applications such as marketing and healthcare. Beyond binary treatments, many applications involve multi-valued treatments. As data are split across more groups, sparsity and imbalance pose challenges. We propose CAMU, which combines (i) a Codebook-enhanced representation network that anchors representations across mini-batches to stabilize learning and facilitate distribution alignment, and (ii) a Cross-Treatment Self-Attention (CTSA) enhanced prediction network for treatment-aware knowledge sharing. We extend CAMU to CAMUER, which transfers unbiased signals from randomized controlled trial (RCT) data via attention alignment and prediction-consistency distillation. The proposed model is deployed in Kuaishou's ad-serving system, serving tens of millions of users. Offline experiments and a large-scale online A/B test demonstrate consistent gains over baselines.

P042 · **Cross-Domain Preference Transfer for Promoting Engagement with Built-in Chatbots** [Industry]

Tianyu Wu (Bytedance); Guanyu Jiang (Bytedance); Hongwei Zheng (Bytedance); Haoming Li (Bytedance); Yongchun Zhu (Bytedance); Jingwu Chen (Bytedance); Feng Zhang (Bytedance)

Recent advances in large language models have led to the widespread integration of chatbots as built-in features in modern applications. To model user latent intent and provide topic guidance, the question recommendation task serves as the entry point for built-in chatbots. In content-consuming

applications, interaction data with built-in chatbots exhibit notable sparsity, while consuming signals account for the majority of user behaviors. Therefore, it is reasonable to incorporate content-consuming signals for user preference modeling. However, a significant domain shift exists between the preference in the content-consuming domain and that in the conversational topic domain. It is non-trivial to transfer those cross-domain signals into meaningful suggested topics that are aligned with users' real-time preference. In this paper, to promote user engagement with the built-in chatbot, we propose a cross-domain question generation framework via preference transfer, which consists of two stages: Cross-Domain Preference Adaptation and Target-Oriented Preference Alignment. The proposed framework is deployed on a large-scale industrial application. Extensive online experiments show significant improvements in built-in chatbot engagement, demonstrating the effectiveness of our approach.

P043 · **Bridging Passive and Active: Enhancing Conversation Starter Recommendation via Active Expression Modeling** [Industry]

Yiqing Wu (ByteDance); Haoming Li (ByteDance); Guanyu Jiang (ByteDance); Jiahao Liang (ByteDance); Yongchun Zhu (ByteDance); Jingwu Chen (ByteDance); Feng Zhang (ByteDance)

Large Language Model (LLM)-driven conversational search is shifting information retrieval from reactive keyword matching to proactive, open-ended dialogues. In this context, Conversation Starters are widely deployed to provide personalized query recommendations that help users initiate dialogues. Conventionally, recommending these starters relies on a closed "exposure-click" loop. Yet, this feedback loop mechanism traps the system in an echo chamber where, compounded by data sparsity, it fails to capture the dynamic nature of conversational search intents shaped by the open world. As a result, the system skews towards popular but generic suggestions. In this work, we uncover an untapped paradigm shift to shatter this harmful feedback loop: harnessing user "free will" through active user expressions. Unlike traditional recommendations, conversational search empowers users to bypass menus entirely through manually typed queries. The open-world intents in active queries hold the key to breaking this loop. However, incorporating them is non-trivial: (1) there exists an inherent distribution shift between active queries and formulated starters. (2) Furthermore, the "non-ID-able" nature of open text renders traditional item-based popularity statistics ineffective for large-scale industrial streaming training. To this end, we propose Passive-Active Bridge (PA-Bridge), a novel framework that employs an adversarial distribution aligner to bridge the distributional gap between passively recommended starters and active expressions. Moreover, we introduce a semantic discretizer to enable the deployment of popularity debiasing algorithms. Online A/B tests on our platform, which serves hundreds of millions of users, demonstrate that PA-Bridge significantly boosts the Feature Penetration Rate by 0.54% and User Active Days by 0.04%.

P044 · **Gesture Clustering for Real-Time User Disentanglement in Shared-Account Recommendation** [Industry]

Huiying Hu (Kuaishou Technology); Xinlang Yue (Kuaishou Technology); Kexin Yi (Kuaishou Technology); Lingzhen Xu (Kuaishou Technology); Yangyi Fang (Kuaishou Technology); Muyang Li (Kuaishou Technology); Yongqi Liu (Kuaishou Technology); Kaiqiao Zhan (Kuaishou Technology)

Shared-account usage is common on short-video platforms, especially on mobile and tablet devices, where a single device is accessed by multiple users. While existing industrial solutions generally focus on behavior sequence purification to disentangle mixed user preferences, such approaches inherently depend on behavior accumulation and therefore lack the capability for real-time user identification. To adapt to online recommendation, utilizing gesture interaction features is a natural and promising option, as they (1) are instantaneous without behavior collection and (2) naturally encode fine-grained user operation habits. Nevertheless, we empirically observe that directly incorporating raw gesture features into recommendation models yields limited gains. Identity-discriminative patterns embedded in gesture signals are largely entangled during the main model training, preventing them from being leveraged as explicit and reliable identity cues. As a result, efficiently utilizing gesture information to provide more distinct identity signals for recommendation models remains a critical challenge. To address this issue, we propose G-CORE (Gesture Clustering for Real-time RECommendation), an unsupervised framework that disentangles gesture representations via clustering before integrating them into the main recommendation model. By providing clearer and more identity-aware signals, G-CORE enables the main model with faster user switching without relying on a volume of behavior accumulation. Through extensive offline experiments and online A/B tests on Kuaishou platform, G-CORE demonstrates its effectiveness in various shared-account scenarios, and has been successfully deployed in the Mobile and Tablet system of the platform.

P045 · **Unifying User Satisfaction and Creator Incentive: A Constrained Optimization Framework for Short-Video Recommendation** [Industry]

Xiaoru Qu (Kuaishou Technology); Dingyi Zhang (Kuaishou Technology); Zhangxi Yan (Kuaishou Technology); Peng Zhang (Kuaishou Technology); Hu Liu (Kuaishou Technology); Yang Zou (Peking University); Jian Liang (Kuaishou Technology); Kaiqiao Zhan (Kuaishou Technology)

Short-video recommendation systems typically optimize for user satisfaction. However, allocating exposure to creators at critical growth stages incentivizes long-term content supply despite compromising immediate user engagement. Existing efforts concerning creator interests aim at either improving creator exposure fairness, matching creators with suitable audiences, or leveraging creator behavior to enhance user welfare. Directly maximizing the joint value of user satisfaction and creator incentive at recommendation time, however, remains largely unaddressed. This presents two challenges. First, the two objectives are heterogeneous in nature, making it non-trivial to formulate this joint optimization as a tractable problem. Second, optimizing creator incentive requires globally coordinated decisions across requests, making real-time serving infeasible. To address these challenges, we formulate the joint maximization as a constrained optimization problem that unifies the two heterogeneous objectives. We further derive an efficient online algorithm based on the primal-dual method, which decouples global incentive constraints into real-time decisions with theoretical guarantees. Experiments on a large-scale short-video platform demonstrate consistent improvements in joint user-creator value over existing baselines.

P046 · **PersonaPlugin: A Multi-Source Persona Framework for LLM Personalization in Telecommunications** [Industry]

Jinmo Kang (LG Uplus Corp.); Minseop Lee (LG Uplus Corp.); Songha Kim (LG Uplus Corp.); Junho Shin (LG Uplus Corp.); Changho Lee (LG Uplus Corp.); Yeonghwan Jeon (LG Uplus Corp.); Hyuncheol Jo (LG Uplus Corp.)

Personalizing large language model (LLM) responses in enterprise environments faces unique challenges: heterogeneous data sources, strict on-premise constraints, and unreliable LLM outputs. We present PersonaPlugin, a production-ready framework that transforms real-world telecom signals—call summaries, app usage, and location patterns—into structured personas for LLM personalization. Our system processes 1M+ behavioral records daily, entirely on-premise. The framework comprises three specialized engines, with the call extraction engine achieving >99% parsing success through cascaded fallback mechanisms. We conduct a comprehensive two-track evaluation: (1) general personalization across 800 scenarios, and (2) call-context personalization across 400 scenarios. Our LLM-as-a-Judge evaluation, validated against human judgments (?=0.71), demonstrates that persona injection significantly improves personalization: +24.0% in P-Score and +26.7% in Adherence over baseline. Integrating call context with personas yields +11% additional improvement, validating the synergistic value of multi-source persona construction.

P047 · **When Search Is Not Enough: Ranking Content for Query-Less Browse Surfaces** [Industry]

Shreya Mahapatra (Adobe); Diksha Bhardwaj (Adobe); Anandita Chopra (Adobe); Tracy Holloway King (Adobe)

Adobe Express is a content creation platform that enables users to easily create and customize visual content. Browse categories, such as Flyers, play a critical role in content discovery, serving as influential entry points where users explore templates without issuing explicit search queries. These browse categories are organized into collections, which represent curated and themed groups of content. Effectively ranking content within these collections is crucial. A simple formulation of using a collection name or category label as a search query may not adequately capture all aspects of a collection. We developed a machine learned browse ranker based on browse session data. Through offline experiments, online A/B testing, and production deployment, our results demonstrate that machine-learned browse ranking produces rankings that differ from collection-name-as-query approaches and that these differences result in measurable improvements to user engagement and product success metrics.

P048 · **Localization Boosting for Growth Markets: Mitigating Cross-Locale Behavioral Bias in Learning-to-Rank** [Industry]

Suryaa Veerabathiran Seran (Adobe); Ashwin Naresh Kumar (Adobe); Tracy Holloway King (Adobe); Jing Zheng (Adobe)

Adobe Express is expanding internationally, but the US has a disproportionately large content supply and interaction volume. Learning-to-rank (LTR) models trained primarily on behavioral feedback inherit this imbalance: templates popular in US are over-served in non-US locales. This cross-locale exposure bias suppresses local content discoverability and degrades ranking quality in growth locales. We show that click-only training suppresses semantically informative localization features. Adding vision-language model (VLM) graded relevance labels as auxiliary supervision alongside clicks improves semantic alignment but does not preserve local content visibility. We propose a multi-objective framework combining behavioral supervision, VLM-derived relevance signals, and locale-aware boosting. Across five locales, the resulting model improves relevance while restoring stable localization, demonstrating the importance of disentangling exposure from semantic supervision.

P049 · **When & How to Write for Personalized Demand-aware Query Rewriting in Video Search** [Industry]

Cheng Cheng (Weixin Group, Tencent); Chenxing Wang (Weixin Group, Tencent); Aolin Li (Weixin Group, Tencent); Haijun Wu (Weixin Group, Tencent); Huiyun Hu (Weixin Group, Tencent); Juyuan Wang (Weixin Group, Tencent); Dongliang Liao (Weixin Group, Tencent)

In video search systems, user historical behaviors provide rich context for identifying search intent and resolving ambiguity. However, traditional methods utilizing implicit history features often suffer from signal dilution and delayed feedback. To address these challenges, we propose WeWrite, a novel Personalized Demand-aware Query Rewriting framework. Specifically, WeWrite tackles three key challenges: (1) When to Write: An automated posterior-based mining strategy extracts high-quality samples from user logs, identifying scenarios where personalization is strictly necessary; (2) How to Write: A hybrid training paradigm combines Supervised Fine-Tuning (SFT) with Group Relative Policy Optimization (GRPO) to align the LLM's output style with the retrieval system; (3) Deployment: A parallel "Lightweight

Recall" architecture ensures low latency. Extensive offline experiments demonstrate that WeWrite significantly outperforms general-purpose LLMs (e.g., Qwen3-32B), improving rewriting accuracy by over 27%. We deployed WeWrite in the main scene of WeChat, a large-scale mobile platform with over 1 billion monthly active users, which improves the Click-Through Video Volume (VV>10s) by 1.07% and reduces the Query Reformulation Rate by 2.97% without increasing the serving cost.

P050 · **WeSEAL: Well-calibrated Search for Eliminating Attention-sink Leakage** [Industry]

Juyuan Wang (Tencent); Chenxing Wang (Tencent); Aolin Li (Tencent); Huiyun Hu (Tencent); Yuchen Fang (Tencent); Haijun Wu (Tencent); Jin Xu (South China University of Technology); Dongliang Liao (South China University of Technology)

Large Language Models (LLMs) have revolutionized relevance ranking, yet their deployment in high-concurrency search systems is stifled by a critical pathology: "Attention Sink Drift." In standard causal attention, models disproportionately allocate weight to the initial document regardless of relevance, creating a severe positional bias that traditionally necessitates costly test-time calibration, effectively halving system throughput. We challenge the consensus that this bias is immutable and present WeSEAL, a single-pass framework that internalizes calibration directly into model parameters. By employing a contrastive head selection mechanism and a novel attention-aligned Supervised Fine-Tuning objective, WeSEAL suppresses background noise and redirects discriminative signals to semantic retrieval heads, thereby intrinsically correcting positional bias. We deployed WeSEAL in the core search scene of WeChat, a large-scale mobile platform serving over 1 billion monthly active users. In rigorous online A/B testing, our approach not only reduced P99 latency to 26.56ms (a 3.3x speedup) and improved the Positive-to-Negative Ratio (PNR) to 2.08, but also delivered statistically significant gains in core engagement metrics—including a 0.402% increase in Click-Through Rate (CTR), a 0.185% growth in Average Video View Duration, and a 1.25% rise in overall user satisfaction—without increasing serving costs. This work establishes a new Pareto frontier for efficient LLM ranking, demonstrating that topologically optimized small-scale models can deliver robust, unbiased performance in strict industrial environments.

P051 · **Agentic Query Reformulation for Contextualized Hyper-Personalized Product Search** [Industry]

Raghav Gaggar (Walmart Global Tech); Sean D Rosario (Walmart Global Tech); Daniel Varivoda (Walmart Global Tech); Shahriar Golchin (Walmart Global Tech); Jayant Sachdev (Walmart Global Tech); Ali Lafzi (Walmart Global Tech); Siddharth Singh (Walmart Global Tech); Jason Cho (Walmart Global Tech); Yog Domlur (Walmart Global Tech); Swati Kirti (Walmart Global Tech); Chittaranjan Tripathy (Walmart Global Tech)

Traditional e-commerce search often suffers from high abandonment rates due to an intent gap where users provide broad or vague queries. We propose a novel framework using Agentic AI to bridge this gap through dynamic query reformulation. By leveraging the ReAct agentic framework, our system performs contextual reasoning by analyzing historical purchase data to transform generic queries into hyper-personalized search queries. Unlike personalization methods that require architectural overhauls, our solution operates on top of existing search systems, ensuring seamless integration with any search system without costly development or downtime. To ensure low-latency performance, we utilize offline pre-computation of top queries and real-time cosine similarity matching. Results from our experiments demonstrate that this agentic approach can significantly improve search ranking leading to improved customer engagement.

P052 · **Agentic Multi-Source Grounding for Enhanced Query Intent Understanding: A DoorDash Case Study** [Industry]

Emmanuel Aboah Boateng (DoorDash); Kyle MacDonald (DoorDash); Akshad Viswanathan (DoorDash); Sudeep Das (DoorDash)

Accurately mapping user queries to business categories is a fundamental Information Retrieval challenge for multi-category marketplaces, where context-sparse queries such as "Wildflower" exhibit intent ambiguity, simultaneously denoting a restaurant chain, a retail product, and a floral item. Traditional classifiers force a winner-takes-all assignment, while general-purpose LLMs hallucinate unavailable inventory. We introduce an Agentic Multi-Source Grounded system that addresses both failure modes by grounding LLM inference in (i) a staged catalog entity retrieval pipeline and (ii) an agentic web-search tool invoked autonomously for cold-start queries. Rather than predicting a single label, the model emits an ordered multi-intent set, resolved by a configurable disambiguation layer that applies deterministic business policies and is designed for extensibility to personalization signals. This decoupled design generalizes across domains, allowing any marketplace to supply its own grounding sources and resolution rules without modifying the core architecture. Evaluated on DoorDash's multi-vertical search platform, the system achieves +10.9pp over the ungrounded LLM baseline and +4.6pp over the legacy production system. On long-tail queries, incremental ablations attribute +8.3pp to catalog grounding, +3.2pp to agentic web search grounding, and +1.5pp to dual-intent disambiguation, yielding 90.7% accuracy (+13.0pp over baseline). The system is deployed in production, serving over 95% of daily search impressions, and establishes a generalizable paradigm for applications requiring foundation models grounded in proprietary context and real-time web knowledge to resolve ambiguous, context-sparse decision problems at scale.

P053 · **Joint Optimization of Relevance and Engagement in Multi-Task Ranking for E-Commerce with Efficient LLM Supervision** [Industry]

Luming Chen (DoorDash Inc.); Jiaqi Xi (DoorDash Inc.); Raghav Saboo (DoorDash Inc.); Martin Wang (DoorDash Inc.); Sudeep Das (DoorDash Inc.)

Optimizing industrial search ranking models solely for user engagement signals often introduces systematic biases, prioritizing popular or price-anchored items that may not satisfy semantic intent. We present a production-scale multi-task ranking system that integrates semantic relevance as a primary optimization objective, enabling explicit and controllable relevance--engagement trade-offs. Our architecture employs an ordinal relevance head that predicts cumulative probabilities over relevance thresholds, preserving the inherent ordering of labels. These outputs are integrated with engagement heads through a unified value model scoring function, enabling systematic balancing of semantic quality and short-term behavioral signals. To provide high-quality supervision for this multi-task framework, we utilize fine-tuned lightweight Large Language Models (LLMs) to generate three-level ordinal relevance labels: irrelevant, moderately relevant, and highly relevant. We address challenges regarding label distribution sensitivity and ensure high alignment with human annotations to enable efficient labeling for over 100 million query--item pairs. Evaluation across offline metrics, including NDCG@10, and online A/B experiments demonstrates that our approach significantly improves semantic alignment while preserving core engagement objectives.

P054 · **Building a Production Shopping Agent at Scale** [Industry]

Chen Luo (Amazon); Jason Choi (Amazon); Ziwei Dong (Amazon); Rahul Dua (Amazon); Cong Xu (Amazon); Xuejing Lei (Amazon); Yuchen Yan (Amazon); Xin Zhang (Amazon); Josef Valvoda (Amazon); Gaurang Sinkar (Amazon); Binit Jha (Amazon); Yi Liu (Amazon); Monica Cheng (Amazon)

Deploying a conversational shopping agent at production scale remains challenging despite rapid advances in large language models. Unlike research prototypes, production systems must satisfy strict requirements on accuracy, latency, reliability, and cost while serving millions of customers. We share our year-long journey of building a production shopping agent at scale and show that combining agentic reasoning with production-aware retrieval, tool orchestration, and system optimization enables a single shopping agent to support product discovery, shopping question answering, and agentic actions under real-world traffic. Our experience shows that LLM-based agents can be reliably operated at Amazon scale and provides practical design principles for industrial search and recommendation systems.

P055 · **Constraint Persistence in Conversational Product Search** [Industry]

Saran Kumar Krishnasamy (GigitAI); Inez Wihardjo (GigitAI)

Conversational product search enables users to iteratively refine their preferences through natural language dialogue, yet managing user constraints across multiple turns and category transitions remains a fundamental challenge. Traditional systems employ either "persist-all" approaches that retain every constraint often leading to over-constrained queries with zero results or "clear-all" approaches that reset context at each turn, forcing users to repeatedly re-state preferences. We propose Constraint-Aware Conversational Search (CACS), a framework with three components powered by large language model reasoning: (1) a Constraint Taxonomy Layer that classifies constraints into identity, contextual, and session types through semantic analysis; (2) a Category Transition Layer that determines constraint relevance during category switches using learned compatibility functions; and (3) a Lightweight Persistence Classifier distilled from LLM reasoning for production deployment. Central to our approach is Constraint Persistence Reasoning (CPR), a structured chain-of-thought prompting strategy that enables LLMs to reason about constraint lifecycles with 96% agreement with human annotators. Through knowledge distillation, we transfer CPR capabilities into a compact model achieving 94% accuracy with P99 latency under 120ms. Online A/B testing demonstrates 23% reduction in zero-result queries and 18% improvement in conversion rate. CACS has been deployed in production at GigitAI across beauty and grocery verticals.

P056 · **A Cost-Efficient AI Copilot for High-Volume E-commerce Customer Support** [Industry]

Aditya Jindal (Flipkart); Midthur Ayeshasiddiqa (Flipkart); Adhish Prasoon (Flipkart)

High-volume customer support operations face persistent challenges: inconsistent adherence to Standard Operating Procedures (SOPs), continuous human agent ramp-up cycles driven by attrition, and high Average Handling Time (AHT) due to repetitive context reconstruction across chat history and backend order views. We present AgentChat Copilot, a human-in-the-loop LLM system deployed at Flipkart, one of the largest e-commerce platforms in India. The system drafts SOP-grounded responses using structured order context and retrieved policy knowledge, while human agents remain responsible for reviewing/editing messages and performing all operational actions. AgentChat Copilot uses a modular pipeline: a Standalone User Query (SAQ) rewriter converts the latest customer utterance into a context-complete query; a small intent classifier routes to fetch intent-specific customer facts; a retrieval layer fetches the most relevant SOP/policy snippets; and a response generation model produces an editable draft consistent with brand tone and policy constraints. We built the system by leveraging supervision from Flipkart's existing support automation: training data was collected from interactions with the production chatbot and augmented with large-scale historical agent-customer chat logs. Deploying in India introduces additional challenges beyond conventional enterprise support: highly heterogeneous customer behavior across languages and

regions, noisy backend order state, and extreme peak-season scale requiring strict latency. Other than these, one of the important facets of this work comes from Deterministic Policy Grounding (DPG): backend-computed eligibility and timeline flags are passed as explicit variables, reducing dependence on the LLM for policy-critical reasoning. In production, AgentChat Copilot improved response consistency and agent concurrency, supported high peak loads, and delivered substantial AHT reductions of ~50% with significant bottom-line savings at scale.

P057 · **Synthetic Data Powers Product Retrieval for Long-tail Knowledge-Intensive Queries in E-commerce Search** [Industry]

Gui Ling (Taobao & Tmall Group of Alibaba); Weiyuan Li (Taobao & Tmall Group of Alibaba); Yue Jiang (Taobao & Tmall Group of Alibaba); Wenjun Peng (Taobao & Tmall Group of Alibaba); Xingxian Liu (Taobao & Tmall Group of Alibaba); Dongshuai Li (Taobao & Tmall Group of Alibaba); Fuyu Lv (Taobao & Tmall Group of Alibaba); Dan Ou (Taobao & Tmall Group of Alibaba); Haihong Tang (Taobao & Tmall Group of Alibaba)

Product retrieval is the backbone of e-commerce search: for each user query, it identifies a high-recall candidate set from billions of items, laying the foundation for high-quality ranking and user experience. Despite extensive optimization for mainstream queries, existing systems still struggle with long-tail queries, especially knowledge-intensive ones. These queries exhibit diverse linguistic patterns, often lack explicit purchase intent, and require domain-specific knowledge reasoning for accurate interpretation. They also suffer from a shortage of reliable behavioral logs, which makes such queries a persistent challenge for retrieval optimization. To address these issues, we propose an efficient data synthesis framework tailored to retrieval involving long-tail, knowledge-intensive queries. The key idea is to implicitly distill the capabilities of a powerful offline query-rewriting model into an efficient online retrieval system. Leveraging the strong language understanding of LLMs, we train a multi-candidate query rewriting model with multiple reward signals and capture its rewriting capability in well-curated query–product pairs through a powerful offline retrieval pipeline. This design mitigates distributional shift in rewritten queries, which might otherwise limit incremental recall or introduce irrelevant products. Experiments demonstrate that without any additional tricks, simply incorporating this synthetic data into retrieval model training leads to significant improvements. Online Side-By-Side (SBS) human evaluation results indicate a notable enhancement in user search experience.

P058 · **Learning to Trust: Dynamic Utilization of Retrieval-Augmented Generation for E-commerce Search Relevance** [Industry]

Tingqiao Xu (Shanghai Jiao Tong University); Shaowei Yao (Taobao & Tmall Group of Alibaba); Chenhe Dong (Taobao & Tmall Group of Alibaba); Yiming Jin (Taobao & Tmall Group of Alibaba); Zerui Huang (Taobao & Tmall Group of Alibaba); Dan Ou (Taobao & Tmall Group of Alibaba); Haihong Tang (Taobao & Tmall Group of Alibaba); Bo Zheng (Taobao & Tmall Group of Alibaba)

Accurately estimating query-item relevance is vital for e-commerce ranking and conversion. While Large Language Models (LLMs) excel at reasoning, they often lack specialized knowledge required for long-tail or fast-evolving queries, necessitating Retrieval-Augmented Generation (RAG). However, production environments face three critical challenges: (1) external context is inherently noisy and inconsistent; (2) extreme latency budgets prohibit multi-stage processing or refinement; and (3) the model must simultaneously assess relevance and context-trust within a unified inference pass. We propose DyKnow-RAG, a reinforcement learning framework that teaches LLMs to learn to trust through dynamic utilization of external knowledge. Built on Group Relative Policy Optimization (GRPO), DyKnow-RAG utilizes a dual-group rollout strategy (parametric-only vs. with-context) and a posterior-driven inter-group advantage scaling mechanism. This enables the model to optimize context utilization without human process labels or extra inference overhead. Our pipeline further integrates structured Chain-of-Thought (CoT) and an uncertainty-prioritized RL pool to stabilize training. Offline evaluations show significant Macro-F1 and Accuracy gains, particularly on noise-sensitive query slices. Importantly, DyKnow-RAG has been deployed in Taobao's production system, serving hundreds of millions of active users and billions of daily search requests. Controlled A/B tests demonstrate consistent lifts in key business metrics, including GSB and Item Goodrate, while maintaining a p99 latency under 400ms. This work provides a scalable and deployable paradigm for operationalizing noisy RAG under extreme efficiency constraints of large-scale industrial search.

P059 · **KARMA: Knowledge-Action Regularized Multimodal Alignment for Personalized Search at Taobao** [Industry]

Zhi Sun (Taobao & Tmall Group of Alibaba); Wenming Zhang (Taobao & Tmall Group of Alibaba); Yi Wei (Taobao & Tmall Group of Alibaba); Liren Yu (Taobao & Tmall Group of Alibaba); Zhixuan Zhang (Taobao & Tmall Group of Alibaba); Dan Ou (Taobao & Tmall Group of Alibaba); Haihong Tang (Taobao & Tmall Group of Alibaba)

Large Language Models (LLMs) are equipped with profound semantic knowledge, making them a natural choice for injecting semantic generalization into personalized search systems. However, in practice we find that directly fine-tuning LLMs on industrial personalized tasks (e.g. next item prediction) often yields suboptimal results. We attribute this bottleneck to a critical Knowledge--Action Gap: the inherent conflict between preserving pre-trained semantic knowledge and aligning with specific personalized actions by discriminative objectives. Empirically, action-only training objectives induce Semantic Collapse, such as attention "sinks". This degradation severely cripples the LLM's generalization, failing to bring improvements to personalized search systems. We propose KARMA (Knowledge--Action Regularized Multimodal Alignment), a unified framework that treats semantic reconstruction as a train-only regularizer. KARMA optimizes a next-interest embedding for retrieval (Action) while enforcing semantic decodability (Knowledge) through two complementary objectives: (i) history-conditioned semantic generation, which anchors optimization to the LLM's native next-token distribution, and (ii) embedding-conditioned semantic reconstruction, which constrains the interest embedding to remain semantically recoverable. On Taobao search system, KARMA mitigates semantic collapse (attention-sink analysis) and improves both action metrics and semantic fidelity. In ablations, semantic decodability yields up to +22.5 HR@200. With KARMA, we achieve +0.25 CTR AUC in ranking, +1.86 HR in pre-ranking and +2.51 HR in recalling. Deployed online with low inference overhead at ranking & pre-ranking stage, KARMA drives +0.9% increase in GMV.

P060 · **RecGPT-Mobile: On-Device Large Language Models for User Intent Understanding in Taobao Feed Recommendation** [Industry]

Bin Zhang (Taobao & Tmall Group of Alibaba); Weipeng Huang (Taobao & Tmall Group of Alibaba); Dimin Wang (Taobao & Tmall Group of Alibaba); Jialin Zhu (Taobao & Tmall Group of Alibaba); Yuning Jiang (Taobao & Tmall Group of Alibaba); Zhaode Wang (Taobao & Tmall Group of Alibaba); Chengfei Lv (Taobao & Tmall Group of Alibaba); Jian Wang (Taobao & Tmall Group of Alibaba); Qichao Ma (Taobao & Tmall Group of Alibaba); Li Chen (Taobao & Tmall Group of Alibaba); Junqing Wu (Taobao & Tmall Group of Alibaba); Yipeng Yu (Taobao & Tmall Group of Alibaba)

Predicting a user's next search query from recent interaction behaviors is a critical problem in modern e-commerce systems, particularly in scenarios where user intent evolves rapidly. Large Language Models (LLMs) offer strong semantic reasoning capabilities and have recently been adopted to enhance training data construction for next-query prediction. However, due to resource constraints on mobile devices, existing applications are deployed on cloud servers, resulting in high inference costs. In this paper, we propose RecGPT-Mobile, a framework that designs a lightweight LLM-based intent understanding agent to improve recommendation quality in mobile e-commerce scenarios. By deploying LLM directly on mobile devices, our approach can capture the evolving interests of users more quickly and adjust the recommendation results in real time. Extensive offline analyzes and online experiments demonstrate that our method significantly improves the accuracy of recommendation results, laying a practical path for LLM deployment in production-scale recommendation systems on mobile devices, as well as a scalable solution for integrating LLMs into real-world next-query prediction systems.

P061 · **Uniboost: Global Coordination with Value Alignment for Fair and Efficient Traffic Allocation** [Industry]

Ge Fan (Taobao & Tmall Group of Alibaba); Nan Zhao (Taobao & Tmall Group of Alibaba); Kai Meng (Taobao & Tmall Group of Alibaba); Cong Luo (Taobao & Tmall Group of Alibaba); Yang Fu (Taobao & Tmall Group of Alibaba); Huiping Chu (Taobao & Tmall Group of Alibaba); Jialin Liu (Taobao & Tmall Group of Alibaba); Yuning Jiang (Taobao & Tmall Group of Alibaba); Bo Zheng (Taobao & Tmall Group of Alibaba)

With the rapid evolution of internet services, recommendation systems have become indispensable. In particular, the blending (re-ranking) stage plays a pivotal role in allocating traffic across diverse business objectives. However, existing approaches often suffer from coupled allocation plans, score inflation, and a lack of interpretability. To address these challenges, we propose Uniboost, a unified traffic allocation framework. Uniboost introduces a posterior value alignment mechanism that calibrates abstract model scores to anchor metrics with explicit business semantics, significantly enhancing interpretability. Furthermore, it employs an independent linear boosting paradigm to decouple complex weighting schemes, enabling precise attribution of each plan's contribution. We validate the effectiveness of Uniboost through online A/B tests and in-depth data analysis, demonstrating three key findings: 1) Reducing the overall weight of weighted scores effectively mitigates unintended business interference, yielding a more efficient micro-level traffic allocation strategy; 2) Post-hoc analyses and aggregated dashboards provide intuitive, macro-level insights that guide the design of the overall traffic allocation mechanism; 3) The proposed "Effective Completion Score" serves as an easily obtainable post-metric that offers a reliable anchor for content recommendation pipelines. Collectively, our experiments show that Uniboost not only improves traffic allocation efficiency and recommendation performance at the micro level but also provides macro-level guidance for system iteration. Thus, this work provides an efficient and controllable traffic regulation solution for large-scale industrial recommendation systems.

P062 · **MedNQS: Medical State-Aware Dual-Stage Next-Turn Question Suggestion for Online Medical Consultations** [Industry]

Yue He (Ant Group); Dongsheng Bi (Ant Group); Minghui Yang (Ant Group); Jian Wang (Ant Group); Jinjie Gu (Ant Group); Junwei Liu (Ant Group)

High-quality follow-up questions are essential for efficient online medical consultations, yet they are difficult for non-professional users to ask. Next-turn Question Suggestion (NQS) addresses this by recommending candidate questions to guide multi-turn interactions. In production, NQS faces two practical challenges: (1) click through signals are noisy and do not reliably reflect question quality, and (2) the standard two-stage pipeline optimizes retrieval and ranking separately, leading to suboptimal end-to-end performance. We present MedNQS, a medical state-aware dual-stage NQS system for online medical consultations. MedNQS uses medical dialogue

state tracking to unify context across stages, retrieves candidates from both a curated FAQ set and LLM-generated questions with state-aware constraints, and reranks them with a lightweight transformer model trained on preference data. A closed-loop pipeline continuously collects real user preference signals from online interactions to refresh training data and update models. MedNQS has been deployed in a large-scale online medical consultation service. Offline evaluations show improved retrieval and ranking quality, and online A/B tests yield +23.5% click-through rate (CTR), +59.3% average session depth, and +19% session success rate.

P063 · **Unified Supervision for Walmart's Sponsored Search Retrieval via Joint Semantic Relevance and Behavioral Engagement Modeling** [Industry]

Shasvat Desai (Walmart Global Tech); Md Omar Faruk Rokon (Walmart Global Tech); Jhalak Nilesh Acharya (Walmart Global Tech); Isha Shah (Walmart Global Tech); Hong Yao (Walmart Global Tech); Utkarsh Porwal (Walmart Global Tech); Kuang-Chih Lee (Walmart Global Tech)

Modern search systems rely on a fast first-stage retriever to fetch relevant items from a massive catalog of items. Deployed search systems often use user engagement signals (clicks, etc.) to supervise bi-encoder retriever training at scale, because these signals are continuously logged from real traffic and require no additional annotation effort. However, engagement is an imperfect proxy for semantic relevance: items may receive interactions due to popularity, promotion, attractive visuals, titles, or price, despite weak query-item relevance. These limitations are further accentuated in Walmart's e-commerce sponsored search. User engagement on ad items is often structurally sparse because the frequency with which an ad is shown depends on factors beyond relevance—whether the advertiser is currently running that ad, the outcome of the auction for available ad slots, bid competitiveness, and advertiser budget. Thus, even highly relevant query–ad pairs can have limited engagement signals simply due to limited impressions. Moreover, e-commerce search pages typically allocate fewer slots for ads than for non-sponsored results, further limiting impressions and reducing engagement coverage over the candidate ad item set. We propose a bi-encoder training framework for Walmart sponsored search retrieval in e-commerce that uses semantic relevance as the primary supervision signal, with engagement used only as a preference signal among relevant items. Concretely, we construct a context-rich training target by combining (i) graded relevance labels from a cascade of cross-encoder teacher models, (ii) a multi-channel retrieval prior score derived from the rank positions and cross-channel agreement of retrieval systems running in production, and (iii) user engagement applied only to semantically relevant items to refine preferences. Our approach outperforms the current production system in both offline evaluation and online A/B tests, yielding consistent gains in average relevance and NDCG.

P064 · **BEATS: Bootstrapping E-commerce Attribute Taxonomies for Search through Iterative Human-AI Collaboration** [Industry]

Yung-Yu Shih (National Taiwan University); Shang-Yu Su (Rakuten Group Inc.); Tzu-I Ho (Taiwan Rakuten Ichiba Inc.); Dongzhe Wang (Rakuten Asia Pte. Ltd.); Yun-Nung Chen (National Taiwan University)

E-commerce platforms in emerging markets often operate with underdeveloped product catalogs that contain only category taxonomies but lack structured attribute schemas. This absence of fine-grained product attributes limits search capabilities---preventing faceted filtering, degrading query understanding, and weakening semantic representations used by search systems. We present BEATS, a human-in-the-loop LLM framework for bootstrapping product attribute taxonomies entirely from scratch. Our approach extends a multi-stage LLM generation pipeline with two critical production stages: (1) proactive quality checking by model developers to filter erroneous outputs, and (2) human annotation by domain-expert local staff to validate generated attributes. The framework operates iteratively---prompts at each generation stage are refined based on quality check observations and annotator feedback across successive rounds, progressively improving attribute quality. Once the attribute taxonomy is established, we employ LLMs to perform structured attribute tagging on individual product items, enriching their contextual representations. The enriched catalog directly benefits multiple components of the search system: enabling granular attribute-based filtering, providing structured features for ranking models, and improving semantic representations for dense retrieval. We validate the generated taxonomy by training dense retrieval models on attribute-enriched product data, demonstrating consistent improvements over baselines using original catalog information. Our system has been deployed at Rakuten Taiwan, enriching 9 major categories spanning 2,694 sub-categories with 67,277 generated attributes, and over 5.4 million products have been tagged with the generated attributes, with plans to enrich the entire product catalog.

P065 · **From Unstructured to Structured: LLM-Guided Attribute Graphs for Entity Search and Ranking** [Industry]

Yilun Zhu (Amazon.com, Inc.); Nikhita Vedula (Amazon.com, Inc.); Shervin Malmasi (Amazon.com, Inc.)

Entity search, i.e., finding the most similar entities to a query entity, faces unique challenges in e-commerce, where product similarity varies across categories and contexts. Traditional embedding-based approaches often struggle to capture nuanced context-specific attribute relevance. In this paper, we present a two-stage approach combining Large Language Model (LLM)-driven attribute graph construction with graph-aware LLM ranking. In the offline stage, we extract structured product attributes from unstructured text, and construct a reusable attribute graph with category-aware schemas. In the online stage, we rank retrieved candidates by reasoning over this structured representation rather than raw text, reducing per-product token usage by 57% while improving ranking precision. Experiments show that our approach outperforms multiple baselines under zero-shot scenarios, achieving a over 5% improvement in average precision without requiring training data, generalizes robustly across diverse product categories, and shows immense potential for real-world deployment.

P066 · **Beyond Semantic Similarity: Explicit Intent Modeling for Query–Product Matching** [Industry]

Amanuel Alambo (eBay Inc.); Sathappan Muthiah (eBay Inc.); Diego Sierra (eBay Inc.); Zhenzhong Zhang (eBay Inc.); Atiq Islam (eBay Inc.); Alex Cozzi (eBay Inc.)

Buyer intent in e-commerce is multi-faceted and is expressed through explicit attributes—such as brand, size, color, and material, rather than through general topical relevance. However, many state-of-the-art scalable query-product matching systems rely on aggregate representations, scoring a single query embedding against a single item embedding. While efficient, this aggregation frequently fails to satisfy individual attribute intent: items can be semantically related, yet violate key aspects specified in the query. In contrast, fine-grained interaction methods can better capture aspect-level constraints, but are typically too expensive due to increased run-time computation and storage costs. We propose an aspect-aware ranking framework that retrieves and resolves aspects in queries and performs fine-grained semantic affinity match against aspects in products to compute an aggregate query-product level aspect affinity score. The proposed approach integrates (i) query aspect resolution (canonicalization) using structured aspect data, (ii) a model to learn granular aspect affinity signal capturing individual aspect-level understanding; and iii) an efficient design for online serving, significantly cutting cost associated with inference speed and storage. This design preserves the scalability of two-tower retrieval while substantially improving explicit intent satisfaction. Experiments on large-scale e-commerce search data show that systematic modeling of aspect affinity on a high aspect-density query segment improves MRR by up to +0.70% over a strong production baseline. To our knowledge, this is among the first deployments of query–product aspect matching at broad aspect and category coverage in an industrial e-commerce search platform.

P067 · **SECRET: SEarch query Classification with label RETrieval** [Industry]

Anna Tigunova (Amazon); Ghadir Eraisha (Amazon); Ahmed Ragab (Amazon)

Search query classification is a crucial element of e-commerce systems, which has direct influence on search result ranking, navigation and visual elements of the website. Conventional query classifiers overfit on imbalanced and long-tail category distribution and are specifically error-prone in the real e-commerce settings with thousands of category labels. Moreover, such classifiers trained to make predictions over a fixed set of labels, cannot be easily extended to incorporate emerging categories. To overcome these challenges, we propose to use contrastive learning for category retrieval with comprehensive label descriptions. Our method SECRET allows us to utilize semantic similarities of queries and categories, instead of relying on simple statistical learning of instance-label co-occurrences. Such setup enhances prediction of rare categories for extreme classification and unlocks the ability to add new category labels in zero-shot settings. Our experiments on a large-scale e-commerce query classification dataset show that the proposed approach outperforms state-of-the-art baselines and overcomes model biases towards most frequent categories.

P068 · **Item-Targeting with Keyword Enhancement: A Hybrid Approach to Ads Targeting** [Industry]

Hua Zou (eBay); Dipanwita Saha (eBay); Ning Chen (eBay); Chen Yang (eBay); Xinxin Shu (eBay); Abraham Bagherjeiran (eBay)

Modern digital advertising systems rely on two primary targeting mechanisms: keyword targeting, where advertisers manually specify search terms, and item-based targeting, which matches ads to buyer queries through semantic similarity. While both approaches are effective in isolation, they leave a coverage gap, particularly for broad queries and new items, where item-based targeting often underperforms. This paper presents a novel hybrid targeting framework that bridges this gap by transferring successful query–item associations from keyword targeting into item-based campaigns via semantic embeddings. Our method leverages historical query–item pairs to create robust query representations and applies similarity-driven transfer to expand the candidate query set for item-targeted ads. To balance coverage and quality, we introduce adaptive thresholding and evaluate multiple query selection strategies. Experimental results, including large-scale A/B tests on eBay's advertising platform, demonstrate that our approach significantly expands impression coverage for broad queries while maintaining relevance. The best-performing configuration selected a small set of high-quality queries per item combined with an adaptive relevance threshold, achieving over 5% incremental coverage while preserving click-through rate (CTR) and conversions-over-impression (CVI) performance. These findings highlight the effectiveness of combining the breadth of keyword targeting with the precision of item-based targeting, offering a scalable solution to the coverage gap in ad marketplaces.

P069 · **Beyond Single Slot: Joint Optimization for Multi-Slot Guaranteed Display Advertising** [Industry]

Zhaoqi Zhang (Nanyang Technological University); Jiaming Deng (Meituan); Miao Xie (China Agricultural University); Linyou Cai (Meituan); Qianlong Xie (Meituan); Xingxing Wang (Meituan); Siqiang Luo (Nanyang Technological University); Gao Cong (Nanyang Technological University)

Guaranteed display advertising is crucial for platform monetization, yet existing methods often operate under a single-slot assumption, limiting their ability to optimize allocation across multi-slot page views. In this paper, we propose a novel joint optimization framework for multi-slot GD allocation, addressing key challenges such as slot-level redundancy, contract imbalance, and exposure concentration. Our approach formulates the allocation as an offline bipartite matching problem with a contract roulette mechanism for slot exclusivity and Page View constraints for impression control, and incorporates a scalable allocation optimization algorithm for efficient large-scale deployment. Extensive online tests on the Meituan advertising platform demonstrate that our method significantly improves merchant ROI, platform revenue efficiency, and contract fulfillment robustness. Specifically, online A/B tests show a 28.99% increase in Average Revenue Per User under 70% traffic, and DID analysis further indicates improved contract stability, demonstrating the strong applicability and effectiveness of our framework in real-world advertising deployments.

P070 · **Cross-Stage Signal Propagation for Negative Filtering in Multi-Stage Ad Delivery** [Industry]

Min Zhang (Meta Platforms, Inc.); Ganlin Song (Meta Platforms, Inc.); Dong Liang (Meta Platforms, Inc.); Lizhang Qin (Meta Platforms, Inc.); Chao Ma (Meta Platforms, Inc.)

Modern ad delivery systems rely on multi-stage ranking pipelines in which retrieval, early ranking, and final ranking progressively refine candidate sets. While this architecture scales well, limited cross-stage coordination often leads to redundancy: early stages forward candidates that are highly filtered downstream, and lightweight signals are underutilized for efficiency optimization. We propose Cross-Stage Signal Propagation, a unified framework for negative filtering in multi-stage systems. The framework supports two complementary signal flows: forward propagation from early ranking to final ranking, and backward propagation from early ranking to retrieval. Forward propagation introduces a consistency-driven filtering gate at the final ranking stage; by detecting signal stability, the system reuses cached outcomes to bypass final ranking for low-value ads. Backward propagation reuses early-stage feedback at retrieval to suppress persistently low-value ads before they enter the ranking pipeline. Deployed on Facebook Feed, the framework achieves significant efficiency gains beyond Negative Exclusion Filtering, including 10.68% CPU reduction in final ranking, 11.20% CPU reduction in early ranking, and 13.01% GPU reduction, while maintaining stable user and advertiser outcomes. This is expected, as the filtered candidates are consistently low-value, characterized by low expected return (e.g., low bid and/or low click-through rate) and low quality, and therefore rarely survive final-stage ranking or win auctions. These results demonstrate that treating ranking signals as reusable resources substantially improves the efficiency of large-scale ad delivery systems.

P071 · **SOLARIS:Speculative Offloading of Latent-based Representation for Inference Scaling** [Industry]

Zikun Liu (Inc.); Liang Luo (Inc.); Qianru Li (Inc.); Zhengyu Zhang (Inc.); Wei Ling (Inc.); Jingyi Shen (Inc.); Zeliang Chen (Inc.); Yaning Huang (Inc.); Jingxian Huang (Inc.); Abdallah Aboelela (Inc.); Chonglin Sun (Inc.); Feifan Gu (Inc.); Fenggang Wu (Inc.); Hang Qu (Inc.); Huayu Li (Inc.); Jill Pan (Inc.); Kaidi Pei (Inc.); Laming Chen (Inc.); Longhao Jin (Inc.); Qin Huang (Inc.); Tongyi Tang (Inc.); Varna Puvvada (Inc.); Wenlin Chen (Inc.); Xiaohan Wei (Inc.); Xu Cao (Inc.); Yantao Yao (Inc.); Yuan Jin (Inc.); Yunchen Pu (Inc.); Yuxin Chen (Inc.); Zijian Shen (Inc.); Zhengkai Zhang (Inc.); Dong Liang (Inc.); Ellie Wen (Inc.)

Recent advances in recommendation scaling laws have led to foundation models of unprecedented complexity. While these models offer superior performance, their computational demands make real-time serving impractical, often forcing practitioners to rely on knowledge distillation—compromising serving quality for efficiency. To address this challenge, we present SOLARIS (Speculative Offloading of Latent-bAsed Representation for Inference Scaling), a novel framework inspired by speculative decoding. SOLARIS proactively precomputes user-item interaction embeddings by predicting which user-item pairs are likely to appear in future requests, and asynchronously generating their foundation model representations ahead of time. This approach decouples the costly foundation model inference from the latency-critical serving path, enabling real-time knowledge transfer from models previously considered too expensive for online use. Deployed across Meta's advertising system serving billions of daily requests, SOLARIS achieves 0.67% revenue-driving top-line metrics gain, demonstrating its effectiveness at scale.

P072 · **DIVER: Unlocking Diversity in Ad Headline Generation with Large Language Models** [Industry]

Chang Wang (Xiaohongshu Inc.); Siyu Yan (East China Normal University); Depeng Yuan (Xiaohongshu Inc.); Yuqi Chen (Xiaohongshu Inc.); Yanhua Huang (Xiaohongshu Inc.); Yuanhang Zheng (Xiaohongshu Inc.); Shuhao Li (Xiaohongshu Inc.); Yinqi Zhang (Xiaohongshu Inc.); Kedi Chen (East China Normal University); Mingrui Zhu (Xiaohongshu Inc.); Ruiwen Xu (Xiaohongshu Inc.)

While Large Language Models (LLMs) possess remarkable generative capabilities, generating diversified and engaging ad headlines in industrial applications remains challenging. Conventional training paradigms often suffer from mode collapse, converging on dominant data patterns and yielding homogeneous outputs. Meanwhile, existing diversity-enhancing techniques like stochastic decoding frequently compromise semantic coherence and controllability. To break this trade-off, we propose DIVER, an automated training framework that internalizes diversity as an intrinsic model capability. DIVER employs an automatic data pipeline to synthesize high-quality, multi-faceted training pairs and utilizes multi-objective reinforcement learning to effectively co-optimize diversity with advertising metrics such as faithfulness and click-through rate (CTR). Unlike personalized approaches, our framework generates diverse content for general users without relying on heavy and costly user-behavior modeling, ensuring efficient inference for large-scale real-time systems. Real-world deployment on Xiaohongshu's Explore Feed demonstrates significant commercial impact, increasing advertiser value (ADVV) by 4.0% and CTR by 1.4%.

P073 · **LLM-Informed Bayesian Content Exploration in Ultra-Recency Recommendation** [Industry]

Qing Feng (Meta Platforms, Inc.); Ivan Ji (Meta Platforms, Inc.); Yiting Hua (Meta Platforms, Inc.); Shujian Bu (Meta Platforms, Inc.)

Cold-start content discovery is a challenge in recommender systems—one that intensifies on modern social platforms. Ultra-recency environments operate in an extreme cold-start regime where most recommended content is newly created and must be evaluated within hours. Rapid content turnover limits the usefulness of historical signals, while sparse early interactions and limited exposure budgets make quality estimation noisy. In this setting, efficient exploration can improve user experience while surfacing emerging creators and trends for long-term platform health. We propose an approach that uses large language models (LLMs) to provide exposure-independent semantic priors for cold-start exploration in such environments. We repurpose LLMs as Bayesian prior constructors: the LLM assesses semantic quality and engagement potential from content alone, before any user interaction occurs, producing an informed prior for early exploration. This prior is integrated into an Upper Confidence Bound (UCB) policy, where it guides early impression allocation and naturally decays as observed data accumulates. The framework combines semantic understanding with uncertainty-aware exploration to enable scalable and efficient cold-start discovery. We apply this item-centric approach within a reranking layer, where a personalized base ranker serves as a relevance gate and impression pacing controls exposure for short-lifecycle content. Through offline evaluation and online A/B experiments on a large-scale social platform, we demonstrate that our LLM-informed approach yields significant gains in exploration efficiency, validating that semantic priors provide meaningful guidance for cold-start recommendation.

P074 · **Pin-SCALE: Semantic Cascading and Alignment Learning for Engagement-Aware IDs in Cold-Start Recommendations** [Industry]

Jiaxing Qu (Pinterest); Junpeng Hou (Pinterest); Yijie Ding (Carnegie Mellon University); Jaewon Yang (Pinterest); Olafur Gudmundsson (Pinterest); Sai Xiao (Pinterest); Huizhong Duan (Pinterest)

Semantic IDs (SIDs) are hierarchical item identifiers learned via residual quantization, offering a promising solution to cold-start recommendation, yet their integration into discriminative recommendations remains challenging and lacks systematic investigation. We present Pin-SCALE, a framework for optimized end-to-end integration of SID into dense retrieval in cascading recommender systems. Pin-SCALE instantiates the principles through three technical contributions: (1) cascading pooling preserving residual quantization hierarchy that outperforms common pooling schema and attention;(2) engagement-aware tokenization that encodes collaborative signals to fully utilize both content and user engagement pattern; and (3) multi-view contrastive learning that harnesses query-item and item-item co-occurrence to align SID representations with engagement prediction. Through extensive offline analyses, we collected systematic guidance and proved the effectiveness of Pin-SCALE. Furthermore, the framework has been tested online on multiple surfaces at Pinterest including Closeup, Home Feed and Search, with different content types like organic and e-commerce. Pin-SCALE achieves substantial improvements in fresh engagements (+3.67% repins) as well as platform-level retention (+0.05% DAU). Our work has been shipped into production to serve billions of requests daily and drive discovery journey of our users, as well as provides principled guidelines for SID adoption in discriminative recommendation models at production scale.

P075 · **Scaling and Stabilizing Large-Scale Embedding-Based Retrieval** [Industry]

Zhen Yang (Walmart Global Tech); Juexin Lin (Walmart Global Tech); Hongwei Shang (Walmart Global Tech); Kaihao Li (Walmart Global Tech); Feng Liu (Walmart Global Tech); Satya Chembolu (Walmart Global Tech); Xunfan Cai (Walmart Global Tech); Xinyi Liu (Walmart Global Tech); Cun Mu (Walmart Global Tech); Tony Lee (Walmart Global Tech); Ciya Liao (Walmart Global Tech)

Embedding-based retrieval (EBR) is foundational to large-scale e-commerce search, yet its effectiveness is often constrained by the quality of training signals and the representational capacity of the encoder. Standard dual-encoders suffer from a training-inference gap: they are optimized on narrow candidate pools but must discriminate against hundreds of millions of items during inference. Furthermore, while transitioning to higher-capacity backbones can mitigate this gap, simply replacing a mature model can lead to inconsistent retrieval behavior and a loss of the domain-specific knowledge established in previous iterations. In this paper, we present a unified pipeline deployed at Walmart that addresses both signal quality and model evolution. Our contributions are two-fold: (1) Hybrid Hard Negative Mining: We integrate Online Cross-Batch Sampling to increase negative diversity by an order of magnitude and Hybrid Offline Mining, which combines cross-encoder predictions with metadata heuristics to identify nuanced mismatches. (2) Legacy-Aware Distillation: We transition from DistilBERT to a higher-capacity GTE-base encoder. To ensure a smooth and

superior transition, we introduce a Warm-Start Distillation technique that transfers domain-specific expertise from the legacy model to the new backbone. Validated through extensive offline experiments and online A/B testing, the proposed pipeline is deployed in live production, delivering a +7.34% improvement in NDCG@5 and a +0.50% lift in gross revenue.

P076 · **Learning to Summarize for Search Relevance with Reinforcement Learning** [Industry]

Nitin Yadav (Walmart Global Tech); Changsung Kang (Walmart Global Tech); Hongwei Shang (Walmart Global Tech)

E-commerce search ranking models face the challenging and critical problem of balancing strict real-time latency constraints with the need for high-quality relevance predictions. In production environments, ranking models often rely primarily on product titles, which frequently omit critical attributes required to satisfy diverse query intents. While full product descriptions provide richer information, their length and verbosity make them computationally impractical for real-time ranking, particularly when using cross-encoder architectures. To address this challenge, we propose ReLSum, a reinforcement learning framework that generates concise, relevance-optimized product summaries for search ranking. ReLSum directly aligns summarization with the ranking objective by using downstream relevance scores as reward signals. The framework conditions the Large Language Model (LLM) solely on product information, while queries are used only to compute rewards during training. This design enables summaries to be generated and cached offline, ensuring no additional inference latency at serving time. Experiments on large-scale production data show substantial improvements in offline NDCG and recall. In online A/B tests, ReLSum delivers statistically significant gains in user engagement metrics such as orders per visitor and units per completed order, with particularly strong improvements for tail queries.

P077 · **BERT-Based Cross-Encoder for Large-Scale Engagement Prediction and Re-ranking in Walmart Search Engine** [Industry]

Shuyi Chen (Walmart Global Tech); Philip Fu (Walmart Global Tech); Ajit Puthenputhussery (Walmart Global Tech); Changsung Kang (Walmart Global Tech); Ming Sun (Walmart Global Tech); Cun Mu (Walmart Global Tech); Sachin Yadav (Walmart Global Tech); Hongwei Shang (Walmart Global Tech)

Product search systems must not only return relevant items but also understand users' implicit preferences beyond their explicit queries. For instance, when searching for "steak", most users implicitly prefer beef steak over equally relevant alternatives like pork steak. Predicting such engagement preferences presents a more complex challenge than traditional relevance modeling, as it requires capturing nuanced query-item relationships that reflect both relevance and user intent. To capture these nuances, we extend semantic understanding to engagement prediction by learning directly from query-item text with engagement labels as supervision rather than relying on historical engagement statistics as input. Our approach effectively captures users' implicit preferences across diverse query types, from tail queries where historical signals are sparse to broad queries where understanding latent intent is critical. Extensive experiments on Walmart's production search data demonstrate significant improvements over production model with a strong relevance foundation: +1.71% add-to-cart lift in interleaving tests and +0.33% in overall search sessions with add-to-cart. Our model is deployed in the production environment of Walmart.com.

P078 · **jina-embeddings-v5-text: Compact and Robust Text Embedding Models using Task-Targeted Distillation** [Industry]

Mohammad Kalim Akram (Jina by Elastic); Saba Sturua (Jina by Elastic); Nastia Havriushenko (Jina by Elastic); Quentin Herreros (Jina by Elastic); Michael Günther (Jina by Elastic); Maximilian Werk (Jina by Elastic); Han Xiao (Jina by Elastic)

Many enterprise search applications require small, efficient embedding models with strong performance on varying downstream tasks. In this paper, we present a pair of new embedding models trained with a novel training regimen that combines model distillation with task-specific contrastive losses. The resulting models are compact, general-purpose text embedding models that match or surpass the performance of existing models of similar size. Furthermore, we study the impact of training using a regularization strategy designed to make embeddings more robust under binary quantization. Our approach yields highly compact models suitable for resource-constrained deployment. All weights are publicly released to foster further research.

P079 · **SkyDistill: Navigating Fuzzy Flight Search Ranking via Precise Intent Distillation** [Industry]

Maolei Huang (Alibaba); Shuhan Song (State Key Lab of Processors, Institute of Computing Technology, CAS; University of Chinese Academy of Sciences); Huawei Cao (State Key Lab of Processors, Institute of Computing Technology, CAS; University of Chinese Academy of Sciences)

Fuzzy flight search is a vital traffic entry for large-scale platforms like Fliggy, serving over 200k daily active users. Unlike traditional fixed-itinerary searches, fuzzy search involves highly ambiguous intentions and flexible constraints, leading to low conversion rates (1.6% UV-CVR) due to sparse intent signals and complex user trade-offs. Standard ranking models struggle with the domain discrepancy between precise and fuzzy queries, often resulting in a misalignment between offline metrics and online performance. To address these challenges, we propose the Multi-level Cross-scenario Knowledge Distillation (MCKD) framework. MCKD transfers "dark knowledge" from a high-capacity Teacher model (trained on precise search data) to a lightweight Student (fuzzy search) model. Our framework introduces three core innovations: 1. Feature-level Hint Learning to align latent semantic representations across heterogeneous feature spaces; 2. Uncertainty-aware Distillation to adaptively weight knowledge transfer based on teacher confidence, mitigating noise propagation; 3. Pairwise Ranking Distillation to explicitly preserve the ranking manifold and relative preference orders. Extensive industrial evaluations and online A/B tests demonstrate that MCKD significantly outperforms state-of-the-art baselines and successfully translates offline gains into substantial online conversion improvements, offering a scalable solution for intent-vague retrieval tasks.

P080 · **PCANet: Price Change Aware Framework for Mitigating Inconsistencies in Large-Scale Ranking Systems** [Industry]

Maolei Huang (Alibaba); Shuhan Song (State Key Lab of Processors, Institute of Computing Technology, CAS; University of Chinese Academy of Sciences); Huawei Cao (State Key Lab of Processors, Institute of Computing Technology, CAS; University of Chinese Academy of Sciences)

Price inconsistency between the flight listing page and subsequent booking stages is a critical yet underexplored challenge in large-scale Online Travel Platforms (OTPs). Due to caching latency and real-time inventory dynamics, particularly cabin-class exhaustion, users frequently encounter unexpected fare increases, leading to degraded trust and reduced conversion. While existing ranking and CTCVR models excel at relevance and conversion prediction, they largely ignore the impact of price volatility on user experience. In this work, we formally define the price change aware ranking problem and propose PCANet, Price Change Aware Framework for Mitigating Inconsistencies in Large-Scale Ranking Systems. PCANet integrates three key components: (1) Price Consistency Aligning (PCA), a pre-ranking module that calibrates cached prices using real-time inventory signals; And (2) Price-Sensitive Matching (PSM), a personalized attention mechanism that adapts ranking based on individual user sensitivity to price jumps; Extensive offline experiments on production data and large-scale online A/B tests on Fliggy demonstrate that PCANet significantly improves both ranking accuracy and price consistency, yielding substantial gains in user engagement and booking conversion. To the best of our knowledge, this is the first industrial-scale solution to address price inconsistency in flight ranking systems.

P081 · **STAR: Staged Training with Aligned Reinforcement Learning and Multi-Faceted Distillation for Interpretable E-commerce Relevance** [Industry]

Chenxu Wang (Alibaba Group); Jianzhi Shao (Alibaba Group); Chi Zhang (Alibaba Group); Tao Zhang (Ant Group)

E-commerce search relevance modeling faces a critical dilemma: traditional models falter with complex queries, while Large Language Models (LLMs), despite their superior reasoning, suffer from the prohibitive latency of auto-regressive Chain-of-Thought (CoT) generation, rendering them infeasible for production. Knowledge distillation offers a promising solution, yet current methods force an undesirable trade-off: sacrificing the very interpretability that makes LLMs powerful, or relying on expensive, unscalable human-annotated rationales. To address this, we propose STAR—Staged Training with Aligned Reinforcement Learning and Multi-Faceted Distillation, a progressive framework that follows a reasoning, ranking, and transfer pipeline to imbue dense models with both high performance and interpretability. First, STAR aligns a teacher LLM's reasoning with task objectives using a novel multi-granularity reward in Group Relative Policy Optimization (GRPO), leveraging only binary labels. Next, it refines the teacher's ability for calibrated scoring via token-level supervision, enabling efficient ranking through a single forward pass without any additional layers. Finally, this "white-box" knowledge is transferred to a compact student via multi-faceted distillation that preserves both reasoning logic and ranking behavior. Offline experiments demonstrate that our 0.6B student model rivals the performance of a strong 8B baseline, making it highly efficient and fully deployable. Real-world effectiveness is validated by significant online A/B test gains, including a +0.93% GoodRate lift and a +1.04% increase in GMV. STAR has been fully deployed to 100% of main search traffic on 1688.com.

P082 · **TRACE: Term-level Reasoning And Chain-of-thought Enhanced distillation for E-commerce Multi-modal Relevance Learning** [Industry]

Chenxu Wang (Alibaba Group); Chi Zhang (Alibaba Group); Fangyi Liang (Shandong University); Jianzhi Shao (Alibaba Group); Manyi Wang (Alibaba Group); Tao Zhang (Ant Group)

In large-scale e-commerce search, accurately modeling multi-modal relevance is paramount for matching user intent—especially given the growing influence of visual content on shopping decisions. However, existing methods often fail to perform fine-grained reasoning. For instance, they struggle when a product title is irrelevant due to marketing language, but its image is highly relevant to the query. Furthermore, they cannot effectively disambiguate which specific query terms are satisfied by the visual versus the textual modality. While Large Language Models (LLMs) excel at such reasoning, their high computational overhead makes direct online deployment infeasible. To bridge this gap, we propose TRACE (Term-level Reasoning And Chain-of-thought Enhanced distillation), a framework designed for deploying advanced reasoning capabilities at scale. TRACE operates in two stages. First, it enhances an LLM's multi-modal reasoning by employing Group Relative Policy Optimization (GRPO) guided by a term-level Chain-of-Thought (CoT) reward function, enabling it to generate detailed, step-by-step relevance judgments. Second, it efficiently transfers this fine-grained reasoning to a lightweight, deployable model using a novel term-level knowledge distillation strategy that inherits reasoning ability.

Offline evaluations show significant improvements across different datasets. More critically, online A/B tests on 1688.com resulted in a +1.04% GMV uplift, a +0.906% LTV increase, and a +0.523% improvement in UV_L2O, demonstrating its significant value in a real-world production environment.

P083 · **GenFacet: End-to-End Generative Faceted Search via Multi-Task Preference Alignment in E-Commerce** [Industry]

Zhouwei Zhai (JD.com); Min Yang (JD.com); Jin Li (JD.com)

Faceted search acts as a critical bridge for navigating massive e-commerce catalogs, yet traditional systems rely on static rule-based extraction or statistical ranking, struggling with emerging vocabulary, semantic gaps, and a disconnect between facet selection and underlying retrieval. In this paper, we introduce GenFacet, an industrial-grade, end-to-end generative framework deployed at JD.com. GenFacet reframes faceted search as two coupled generative tasks within a unified Large Language Model: Context-Aware Facet Generation, which dynamically synthesizes trend-responsive navigation options, and Intent-Driven Query Rewriting, which translates user interactions into precise search queries to close the retrieval loop. To bridge the gap between generative capabilities and search utility, we propose a novel multi-task training pipeline combining teacher-student distillation with GRPO. This aligns the model with complex user preferences by directly optimizing for downstream search satisfaction. Validated on China's largest self-operated e-commerce platform via rigorous offline evaluations and online A/B tests, GenFacet demonstrated substantial improvements. Specifically, online results reveal a relative increase of 42.0% in facet Click-Through Rate (CTR) and 2.0% in User Conversion Rate (UCVR). These outcomes provide strong evidence of the benefits of generative methods for improving query understanding and user engagement in large-scale information retrieval systems.

P084 · **Probe-then-Plan: Environment-Aware Planning for Industrial E-commerce Search** [Industry]

Mengxiang Chen (JD.com); Zhouwei Zhai (JD.com); Jin Li (JD.com)

Modern e-commerce search is evolving from simple keyword matching to resolving complex user intents. While large language models (LLMs) offer powerful reasoning capabilities, existing LLM-based search paradigms suffer from a fundamental blindness-latency dilemma: query rewriting methods are agnostic to retrieval tool capabilities and real-time inventory states, resulting in invalid plans; conversely, deep search agent approaches initially plan without environment awareness, then rely on iterative tool calls and reflection to perceive and correct failures, leading to seconds of latency, incompatible with the sub-second budget for the planning module in industrial e-commerce search. To resolve this conflict, we propose Environment-Aware Search Planning (EASP), a novel paradigm that reformulates search planning as a dynamic reasoning process grounded in environmental reality. EASP introduces a Probe-then-Plan mechanism: a lightweight Retrieval Probe first exposes the retrieval snapshot, enabling the Planner to diagnose execution gaps and generate grounded search plans. Our methodology unfolds in three stages: (1) Offline Data Synthesis: The Teacher Agent synthesizes diverse, execution-validated plans by diagnosing the retrieval environment exposed by the Retrieval Probe. (2) Planner Training and Alignment: The Planner is initialized via supervised fine-tuning (SFT) on the offline dataset to internalize the Teacher's diagnostic capabilities, followed by alignment with business outcomes (conversion rate) through reinforcement learning. (3) Adaptive Online Serving: A complexity-aware routing mechanism selectively activates the planning pipeline only for complex queries, ensuring optimal resource allocation.Extensive offline evaluations and online A/B testing on JD.com demonstrate that EASP significantly improves relevant recall and achieves substantial lifts in UCVR and GMV. EASP has been successfully deployed in JD.com's AI-Search system.

P085 · **Effective Offline LLM and DNN based Matching, Filtering and Ranking for Search Ads Retrieval in E-Commerce** [Industry]

Shuping Ji (Coupang); Jianguo Yao (Shanghai Jiao Tong University); Dong Zhang (Jinan Inspur Data Technology Co., Ltd)

For search Ads retrieval in e-Commerce, when a customer inputs a query, from tens of millions of Ads products, the system needs to quickly retrieve a limited number of candidates that can not only meet the customer's search intension but also have high conversion probability. This constructs challenges for both retrieval efficiency and retrieval quality on the perspective of relevance and engagement. To alleviate these challenges, traditional solutions usually adopt a cascading architecture including the recall, pre-ranking, ranking and re-ranking stages. These methods are helpful, however, due to the hard limit of online customer request processing latency, only a small ratio of relevant candidates can be retrieved and the quality of pre-ranking is limited. Different from the online cascading architecture, we propose an effective offline-based solution: for each query, we first use multi-path recall, such as BERT-based embedding to retrieve a much larger number of candidates. Then, a carefully designed LLM-based relevance model is used to filter out the irrelevant candidates. Finally, we use powerful DNN models to rank the left candidates, and only a small number of best candidates are passed to the online system for real-time usage. In the online system, we keep a single ranking stage and use the most powerful models only. Compared to the online cascading architecture, our proposed method can not only largely reduce the online system's overhead and latency, but also significantly improve the overall quality. Real world A/B test experiments on Coupang search Ads show that this solution can improve the revenue, GMV, and advertiser ROAS by 5.79%, 11.37%, and 6.69%, respectively. The relevance defect ratio can be reduced by 68%. Meanwhile, the customer request processing P99 latency can reduced by 21%, when the online serving cluster's hardware cost is reduced by 52%.

P086 · **Scalable Multi-vector Retrieval for Policy Enforcement on E-commerce Marketplaces** [Industry]

Akshit Sarpal (Walmart Global Tech); Srinivas Kotamraju (Walmart Global Tech); Raviteja Uppalapati (Walmart Global Tech)

This paper presents a hybrid enforcement system designed to retrieve relevant content policies for e-commerce product listings. Given the scale of millions of daily product submissions and over a thousand distinct content policies, an exhaustive evaluation of every item against every policy is computationally intractable. To address this, we introduce a retrieval-based architecture that identifies relevant policies for each newly listed product. We employ zero-shot Large Language Model (LLM) prompting to distill complex policies into concise, CLIP-aligned textual representations, creating a lightweight vector store for efficient retrieval. This layer functions as a high-recall router that effectively filters conformant content while accurately routing suspicious items to specific policy classes, significantly reducing the computational burden on the more expensive downstream verification layer. Experimental results on an internal dataset demonstrate that our method achieves a 3.8% improvement in recall at a fixed flag rate compared to standard dense retrieval baselines. Additionally, we demonstrate that use of hard negative samples for index expansion improves recall by an average of 26.3% for a representative set of policies.

P087 · **Negative Data Mining for Contrastive Learning in Dense Retrieval at IKEA.com** [Industry]

Eva Agapaki (IKEA Retail (Ingka Group)); Amritpal Singh Gill (IKEA Retail (Ingka Group))

Contrastive learning is a core component of modern retrieval systems, but its effectiveness heavily relies on the quality of negative examples used during training. In this work, we present a systematic approach to improving dense retrieval for IKEA product search through structured negative sampling strategies and scalable LLM-as-a-judge relevance evaluation. Building on IKEA Search Engine's late-interaction retrieval architectures, we introduce two key contributions: (1) structured negative sampling strategies that leverage product hierarchical taxonomy and product attributes to generate semantically challenging negatives, and (2) a comprehensive LLM-based evaluation methodology for generating training data. Rather than relying on sparse human annotations or random sampling, our LLM-based evaluation system allocates a score for all candidate products against each query.Our methodology achieves +2.6\% average category accuracy on offline real user query experiments on the Canada market. However, our A/B test on long-tail queries showed no statistically significant differences in user engagement metrics between the improved and baseline models ($p > 0.05$). We trace this gap to user search behavior: 67\% of popular searches exhibit zero-click rates above 50\%, indicating that a substantial proportion of search sessions result in no product engagement regardless of result ranking. These findings underscore the importance of hard negative mining but also the need for grounding training data and offline evals in real user search behavior---including query intent distribution and zero-click patterns---to bridge the gap between offline retrieval quality and online user engagement.

P088 · **LLM Agents Factory: Retrieval of Domain-Specific LLM Agents** [Industry]

Vitalii Belov (Sber AI); Artyom Sosedka (Sber AI); Andrey Sakhovskiy (Sber AI); Elizaveta Kovtun (Sber AI); Artyom Boyarskikh (Sber AI); Semen Budennyy (Sber AI)

Large language model (LLM) agents improve task performance by decomposing problems into role-specialized behaviors. However, their practical deployment is often limited by the computational cost and instability associated with the on-the-fly agent design for each user request. To address this, we present LLM Agents Factory, a retrieval-based framework that constructs domain-specific and Wikipedia-grounded agents on demand using a base of over 20K predetermined agent profiles. Our framework supports two modes: (1) agent profile retrieval via semantic search and (2) distillation into a compact model fine-tuned for direct agent generation. Experiments on MMLU, BIG-bench, and BIG-bench Hard in a single-agent scenario demonstrate that our retrieval-based agent construction surpasses non-agent baselines in accuracy while matching AutoGen generation quality with a 120B backbone at a substantially lower inference cost. Our work reveals that retrieval from a structured agent repository provides a cost-efficient, accurate, and controllable alternative to dynamic agent generation, responding to the strict demands of industrial applications. We provide the implementation code and the agent base in https://huggingface.co/frontier-ai/llm-agent-factory.

P089 · **Beyond Fluency: Toward Reliable Trajectories in Agentic IR** [Industry]

Anushree Sinha (Google); Abhishek Dharmaratnakar (Google); Debanshu Das (Google); Srivaths Ranganathan (Google)

Information Retrieval is shifting from passive document ranking toward autonomous agentic workflows that operate in multi-step Reason–Act–Observe loops \cite{ranganathan2026multi}. In such long-horizon trajectories, minor early errors can cascade, leading to functional misalignment between internal reasoning and external tool execution despite continued linguistic fluency.This position paper synthesizes failure modes observed in industrial agentic systems, categorizing errors across planning, retrieval, reasoning, and execution. We argue that safe deployment requires moving beyond endpoint accuracy toward trajectory integrity and causal attribution. To address compounding error and deceptive fluency, we propose verification gates at each interaction unit and advocate systematic abstention under calibrated uncertainty. Reliable Agentic IR systems must prioritize process correctness and

grounded execution over plausible but unverified completion.

P090 · **Fast and Faithful: Real-Time Verification for Long-Document Retrieval-Augmented Generation Systems** [Industry]

Xunzhuo Liu (vLLM Semantic Router Project); Bowei He (MBZUAI); Xue Liu (MBZUAI); Haichen Zhang (AMD); Huamin Chen (Microsoft)

Retrieval-augmented generation (RAG) is increasingly deployed in enterprise search and document-centric assistants, where responses must be grounded in long and complex source materials. In practice, verifying that generated answers faithfully reflect retrieved documents is difficult: large language models can check long contexts but are too slow and costly for interactive services, while lightweight classifiers operate within strict context limits and frequently miss evidence outside truncated passages. We present the design of a real-time verification component integrated into a production RAG pipeline that enables full-document grounding under latency constraints. The system processes documents up to 32K tokens and employs adaptive inference strategies to balance response time and verification coverage across workloads. We describe the architectural decisions, operational trade-offs, and evaluation methodology used to deploy the verifier, and show that full-context verification substantially improves detection of unsupported responses compared with truncated validation. Our experience highlights when long-context verification is necessary, why chunk-based checking often fails in real documents, and how latency budgets shape model design. These findings provide practical guidance for practitioners building reliable large-scale retrieval-augmented applications.

P091 · **What's in a Name? Product Normalization for Enterprise RAG at NetApp** [Industry]

Grant Glass (NetApp)

Retrieval-Augmented Generation (RAG) systems for enterprise technical documentation face a persistent vocabulary challenge: users refer to the same product by many different names, abbreviations, and legacy terms, while the underlying search index uses a controlled product taxonomy. This mismatch degrades retrieval precision and recall, ultimately reducing answer quality. I present DOC, the production conversational assistant deployed on docs.netapp.com, and describe two complementary techniques that leverage curated product-name knowledge artifacts to close this vocabulary gap. First, a query-time product-name normalizer rewrites user queries by mapping informal or ambiguous product mentions to their canonical names and appending the latest version identifiers. Second, a product-scope filter dynamically constructs structured filter expressions from a product catalogue, ensuring that retrieval is restricted to the set of documented products. Together, these lightweight, maintainable components yield measurable improvements in retrieval relevance without retraining any model. On a stratified sample of 300 real production queries from docs.netapp.com, with slug-derived ground-truth relevance judgments, curated normalization improves Mean Reciprocal Rank by +42%, +31%, and +24% over raw queries on BM25 and on two Azure dense retrievers (text-embedding-3-small and text-embedding-3-large), all with non-overlapping 95% bootstrap CIs. A factorial ablation reveals that the Product-Scope Filter, when driven dynamically by the same detected products, contributes a complementary +25-78% MRR gain on top of every retriever/normalizer pair, and that the curated normalizer still out-performs a hyperparameter-tuned NPMI log-mining baseline trained on 93, 028 production query-URL pairs across both filter states. I share system design details, experimental results, and operational lessons from two years of production deployment spanning 60+ enterprise storage and cloud products.

P092 · **LLM-Click Agreement: Harmonizing Implicit Feedback and Semantic Judgments for Enterprise Search** [Industry]

Avinash Kumar (Microsoft Development Center); Hardi Rathod (Microsoft Development Center); Rohan Mallick (Microsoft Development Center); Swarnangsu Acharyya (Microsoft Development Center)

Message search and retrieval on enterprise collaboration platforms is challenging in "eye-off" environments, where explicit human relevance labels cannot be collected and engagement signals such as clicks and dwell times are noisy and behaviorally biased. Large Language Models (LLMs) offer an alternative source of semantic supervision, but models trained solely on LLM-derived labels often regress on engagement-based metrics. We present LLM-Click Agreement Labeling, an industrial-scale supervision strategy that retains only those query--message pairs where click-based labels and LLM-generated labels agree. This selective filtering reduces supervision noise and maintains a balance between user interaction patterns and semantic relevance. In a human-annotated pilot, agreement-based labels improved accuracy by +18% relative to click-only supervision. A worldwide A/B deployment further showed that the approach preserves traditional search quality while delivering statistically significant gains in conversational grounding (e.g., +1.4% CiteDCG, +5.7% Good Citation Count). These results highlight that improving supervision quality, rather than modifying model architecture, is the most effective lever for advancing retrieval performance in large-scale enterprise systems.

P093 · **As It Was: Aligning LLM Search Evaluation with Historical User Preferences** [Industry]

Ali Vardasbi (Spotify); Gustavo Penha (Spotify); Enrico Palumbo (Spotify); Claudia Hauff (Spotify); Hugues Bouchard (Spotify); Mounia Lalmas (Spotify)

Large-scale search systems evolve faster than human quality assurance scales, especially for long-tail intents and multilingual queries.
LLM-as-a-judge approaches are a scalable alternative for evaluating the relevance of search engine result pages (SERPs), but judgments based solely on semantic similarity or world knowledge can drift from actual user preferences, particularly for ambiguous queries. We introduce a behavior-grounded LLM judge that augments each SERP item with a lightweight, auditable behavioral prior in the form of a Query--Relevance--Impressions (QRI) card. Each card summarizes how users have historically interacted with similar queries and results, providing compact empirical evidence that the judge can cite to resolve ambiguity and make more consistent relevance judgments, while still relying on semantic reasoning. In a large-scale music search evaluation at Spotify, using relevance estimates derived from historical user interactions across 6,000 recomposed SERPs, the behavior-grounded judge achieves stronger alignment with user preferences, improving Spearman rank correlation by approximately +5% overall and yielding a +91% relative improvement on disagreement cases. On a multilingual human-judged dataset spanning five languages, grounding further increases correlation with human relevance judgments by +15%. Importantly, when evaluated against outcomes from a live A/B test, the grounded judge shows consistently higher alignment with the observed winning model. While absolute alignment remains moderate, these findings demonstrate that lightweight behavioral grounding can improve the reliability and practical usefulness of LLM-based evaluation in real-world search systems.

P094 · **Scaling Search Relevance: Augmenting App Store Ranking with LLM-Generated Judgments** [Industry]

Evangelia Christakopoulou (Apple); Vivekkumar Patel (Apple); Hemanth Velaga (Apple); Sandip Gaikwad (Apple)

Large-scale commercial search systems optimize for relevance to drive successful sessions that help users find what they are looking for. To maximize relevance, we leverage two complementary objectives: behavioral relevance (results users tend to click or download) and textual relevance (a result's semantic fit to the query). A persistent challenge is the scarcity of expert-provided textual relevance labels relative to abundant behavioral relevance labels. We first address this by systematically evaluating LLM configurations, finding that a specialized, fine-tuned model significantly outperforms a much larger pre-trained one in providing highly relevant labels. Using this model as a force multiplier for human annotation, we generate millions of textual relevance labels to overcome the data scarcity. We show that augmenting our production ranker with these textual relevance labels leads to a significant outward shift of the Pareto frontier: offline NDCG improves for behavioral relevance while simultaneously increasing for textual relevance. These offline gains were validated by a worldwide A/B test on the App Store ranker, which demonstrated a statistically significant +0.24% increase in conversion rate, with the most substantial performance gains occurring in tail queries, where the new textual relevance labels provide a robust signal in the absence of reliable behavioral relevance labels.

P095 · **Fast and Feasible: Permutation-based Constrained Reranking for Revenue Maximization** [Industry]

Svetlana Shirokovskikh (Avito); Anastasiia Soboleva (Avito); Ekaterina Solodneva (Lomonosov Moscow State University); Aleksandr Katrutsa (Avito); Roman Loginov (Avito); Egor Samosvat (Avito)

Search and recommender systems have produced highly relevant search results. A natural next step in the development of such systems in e-commerce is to rerank these results to increase the platform's revenue from paid promotion products. However, maximizing revenue alone may degrade the user experience by reducing relevance or increasing fraud risk. To avoid this, we state the reranking problem as an integer linear program (ILP) that maximizes revenue subject to per-query constraints on other metrics, e.g., relevance. Since solving ILP exactly for every query is slow for deployment to the online service, we propose a lightweight permutation-based reranking approximation algorithm PermR. At each step, the algorithm selects a pair of neighboring items and swaps them to either improve the objective or repair a violated constraint. We evaluate PermR across multiple categories of a large classified platform in offline and online settings. PermR achieves about 63% of the ILP revenue improvement, within production latency limits, preserving all constraints. In a 14-day online A/B test over 56 million search queries, PermR increased revenue by 2%.

P096 · **Domain-Specific Reranking: When is Adaptation Worth the Cost?** [Industry]

Tuukka Karvonen (KPMG Advisory Holdings Co., Ltd.); Alessio Staffini (KPMG Advisory Holdings Co., Ltd.); Kenichi Maeda (KPMG Advisory Holdings Co., Ltd.); Menaka Arudchelvan (KPMG Advisory Holdings Co., Ltd.); Tomomi Ozawa (KPMG Advisory Holdings Co., Ltd.); Nami Tokunaga (KPMG Advisory Holdings Co., Ltd.); Noriaki Hirokawa (KPMG Advisory Holdings Co., Ltd.)

Two-stage retrieval pipelines, where a fast retriever identifies candidates and then a more powerful reranker scrutinizes them, have become standard across retrieval-based applications, including open-domain question answering and retrieval-augmented generation systems. In specialized domains or low-resource languages, domain adaptation is often required to ensure retrieval quality. In this paper, we explore domain adaptation of rerankers using unlabeled and labeled, real and synthetic datasets across legal, biomedical, and auditing domains. In the domains we study, we find that when pretraining data is limited, task-specific fine-tuning—using only a few hundred labeled examples in the domain—can match or surpass the gains of domain-adaptive pretraining, suggesting that domain-adaptive pretraining may not always be necessary for strong domain performance. We further evaluate this on two industrial Japanese datasets, where we observe practical improvements with purely synthetic training data.

P097 · **Efficient LLM Adaptation for Opinion Knowledge Graph Construction: Lessons from the Telecom Industry** [Industry]

Nai-Chi Yang (Academia Sinica); Yu-Ming Hsieh (Academia Sinica); Wei-Yun Ma (Academia Sinica); Kuo-Wei Chang (Chunghwa Telecom)

Deploying Large Language Models (LLMs) to build real-time public opinion analysis platforms is critical for enterprises to drive strategic decisions, monitor competitor dynamics, and manage brand reputation crises. However, adapting these models for vertical domains, such as extracting structured insights from customer feedback, requires balancing performance with prohibitive training costs, which is particularly challenging for low-resource languages like Traditional Chinese. In this work, partnered with Chunghwa Telecom, the largest telecommunications operator in Taiwan, we present a systematic study on adapting open-source LLMs (LLaMA-3, Qwen and Mistral) for structured opinion mining. Our goal is to transform noisy, multi-turn forum discussions into Opinion Knowledge Graphs (OKGs) to support downstream retrieval, competitor analysis, and real-time customer service monitoring. We investigate critical engineering trade-offs in the instruction-tuning pipeline, focusing on data efficiency and training strategies. Our experiments yield several counterintuitive findings valuable for industrial practitioners: (1) Cost-Efficiency: Contrary to common practices, we find that Continual Pretraining (CPT) offers limited benefits when starting with strong instruction-tuned baselines, suggesting practitioners can skip this resource-intensive step. (2) Data Strategy: We demonstrate that padding strategies significantly outperform packing in terms of data utilization and convergence speed for this task, effectively compensating for limited training data. (3) Optimization: We show that reducing batch size serves as an effective alternative to increase parameter update frequency without overfitting. This paper provides a cost-effective, scalable recipe for deploying GenAI-based information extraction systems in resource-constrained industrial settings. Our code is available at https://github.com/ckiplab/telecom-okg.

P098 · **Bridging the Topology-Semantic Gap: A Benchmark and Framework for Power Grid Work Ticket Generation** [Industry]

Jiangbing Mao (Xiamen University); Tianke Xiang (Xiamen University); Yantong Zhu (Xiamen University); Wenliang Liu (State Grid Fujian Electric Power Co., Ltd. Xiamen Power Supply Company); Qinggang Zhang (The Hong Kong Polytechnic University); Zhihong Zhang (Xiamen University)

\begin{abstract}Retrieval-Augmented Generation (RAG) is a promising paradigm for domain-specific knowledge integration. However, applying generic RAG frameworks to power grid operation and maintenance (O\&M) faces critical challenges due to the strict physical constraints of electrical networks. We identify a fundamental \textit{Topological Association Modeling Deficiency} in existing methods, manifesting as Retrieval Linkage Degradation (where semantic ambiguity severs links between abstract intents and specific technical entities) and Topological Reasoning Deficiency (failing to adhere to rigid physical interlocking rules, causing fatal safety violations). To address these bottlenecks, we introduce the \textbf{Power Grid Work Ticket (PGWT)} dataset to systematically evaluate topological reasoning under real-world noise and sparsity. Furthermore, we propose \textbf{Topology-Semantic Aligned Retrieval (TSA-Retrieval)}. This framework harmonizes semantic spaces with topological structures by fusing symbolic pattern matching with subgraph structural encoding to ensure physical validity, while explicitly injecting topology-aware reasoning chains to guide constraint-compliant generation. Extensive experiments demonstrate that TSA-Retrieval significantly outperforms state-of-the-art LLMs and generic RAG baselines, offering a robust and

P099 · **JARVIS: An Evidence-Grounded Retrieval System for Interpretable Deceptive Reviews Adjudication** [Industry]

Nan Lu (Beijing Jiaotong University); Leyang Li (JD.com); Yurong Hu (JD.com); Rui Lin (JD.com); Shaoyi Xu (Beijing Jiaotong University)

Deceptive reviews are fabricated feedback designed to artificially manipulate the perceived quality of products. Within modern e-commerce ecosystems, these reviews remain a critical governance challenge. Despite advances in review-level and graph-level detection methods, two pivotal limitations remain: inadequate generalization and lack of interpretability. To address these challenges, we propose JARVIS, a framework providing Judgment via Augmented Retrieval and eVIdence graph Structures. Starting from the review to be evaluated, it retrieves semantically similar evidence via hybrid dense–sparse multimodal retrieval, expands relational signals through shared entities, and constructs a heterogeneous evidence graph. The large language model then performs evidence-grounded adjudication to produce interpretable risk assessments. Offline experiments demonstrate that JARVIS enhances performance on our constructed review dataset, achieving a precision increase from 0.953 to 0.988 and a recall boost from 0.830 to 0.901. In the production environment, our framework achieves a 27% increase in the recall volume and reduces manual inspection time by 75%. Furthermore, the adoption rate of the model-generated analysis reaches 96.4%.

P100 · **Code-Based English Models Reveal Surprising Performance on Chinese QA Pair Extraction Task** [Industry]

Jiajun Yu (Zhejiang University); Linghan Zheng (Ant Group); Hui Liu (Ant Group); Jiayuan Dong (Ant Group); Yue Shen (Ant Group); Zhiwei Liu (Ant Group); Yaozhen Liang (Zhejiang University); Yong Li (Ant Group); Haishuai Wang (Zhejiang University)

This paper explores advancements in automated Question-Answer (QA) extraction using large language models (LLMs), addressing challenges in transforming unstructured text into high-quality, retrievable QA pairs. Traditional approaches, whether through segmented question and answer generation or end-to-end extraction, often struggle with efficiency, dataset limitations, and performance consistency. Leveraging recent progress in LLMs, we constructed a large-scale Chinese QA extraction dataset with 143,846 documents and evaluated multiple fine-tuned models on public and private datasets. Surprisingly, code-based English LLMs outperformed Chinese-specialized models on Chinese text with a lower hallucination rate. Building upon this finding, we enhanced the best-performing code-based model with an expanded Chinese vocabulary, creating Code Llama-M, which achieved better results. Integrating Code Llama-M into our internal assistant, Luo Ying, demonstrated notable user satisfaction gains, affirming its practical impact. Key contributions include: (i) creation of a robust Chinese QA extraction instruction dataset; (ii) evidence of cross-lingual efficacy of code-based LLMs for Chinese QA tasks, further enhanced through Code Llama-M's expanded Chinese vocabulary; and (iii) successful application of the fine-tuned LLM in a live assistant system, enhancing user experience.

P101 · **From Doxa to Logos in Scientific Peer Review** [Industry]

Negar Arabzadeh (Berkeley); Sajad Ebrahimi (University of Toronto); Alireza Daghighfarsoodeh (Reviewerly); Soroush Sadeghian (Reviewerly); Seyed Mohammad Hosseini (Reviewerly); Hai Son Le (Reviewerly); Mahdi Bashari (Reviewerly); Ebrahim Bagheri (University of Toronto)

Peer review is central to scientific decision-making, yet it is rarely evaluated or audited at scale. Growing submission volumes and the increasing use of large language models (LLMs) in drafting reviews have introduced new challenges for transparency, accountability, and quality control. Today, peer reviews are often produced through hybrid human--AI workflows, where a reviewer may develop the core evaluative ideas while using an LLM to refine wording, restructure arguments, or improve fluency. This shift raises new questions beyond authorship detection alone: Are reviews constructive? Are reviewer claims grounded in the submitted paper? How can we quantify collaboration between human reasoning and AI-assisted writing, and distinguish whether the intellectual contribution or the surface text originates from humans or models?In this industry talk, we present Reviewerly's retrieval-centered infrastructure for auditing peer review at scale. We describe three deployed systems: \textit{Peeriscope}, which evaluates review quality across multiple interpretable dimensions; \textit{Peerispect}, which uses retrieval-augmented generation (RAG) to verify whether reviewer claims are supported by evidence in the manuscript; and \textit{PeerPrism}, which analyzes hybrid human--AI authorship by disentangling the origin of ideas from the origin of text in peer reviews, enabling measurement of how human reasoning and AI-generated writing interact within a review. We will share architectural design decisions, lessons learned from real-world deployment, and practical trade-offs among model complexity, interpretability, computational efficiency, and predictive accuracy. The talk will include short live demonstrations\footnote{\url{https://app.reviewer.ly/app/peerispect}} \footnote{\url{https://app.reviewer.ly/app/peeriscope}} of our systems to illustrate how the combination of retrieval systems and LLM-based pipelines can improve peer review quality and strengthen transparency and research integrity across high-volume decision-making environments.

P102 · **Policy-Guided RAG: Enforcing Verbatim and Controlled Synthesis** [Industry]

Mina Naghash Asadi (Services Australia); Leila Tavakoli (Services Australia); Mustafa Bilgrami (Services Australia)

Retrieval-Augmented Generation (RAG) typically assumes retrieved text can be freely paraphrased, but this fails in regulated domains where some passages must be quoted verbatim. We propose a policy-guided transformation pattern for RAG, where retrieved segments carry sensitivity metadata and a lightweight router selects a transformation mode (Verbatim, Combined, or Synthesis). Quote-aware templates enforce constraints, and answer-side auditing verifies the final output using response--sensitive matching and a two-pass check for unquoted sensitive content. Experiments on an organisational corpus with simulated constraints show clear compliance--utility trade-offs. Compared to a Standard RAG baseline under the same retrieved context, policy-guided Combined substantially reduces verbatim exposure risk (VVR) across model families and generally reduces paraphrase-style sensitive reuse (PSR), while near-verbatim proximity can persist when sensitive and non-sensitive evidence are mixed. Under Verbatim, exposure decreases further, but some models pay a coverage cost via abstention; Synthesis preserves utility when sensitive evidence is removed. We summarise these outcomes (with emphasis on Combined) in a compact metric panel and distil engineering lessons for deploying auditable transformation governance in real RAG systems. We release the full evaluation protocol (routing rules, prompt templates, and compliance metrics) to enable replication on other corpora with similar policy constraints.

P103 · **Financial Transaction Retrieval and Contextual Evidence for Knowledge-Grounded Reasoning** [Industry]

Artem Sakhno (Sber AI Lab); Daniil Tomilov (Sber AI Lab); Yuliana Shakhvalieva (Sber AI); Inessa Fedorova (Sber AI); Daria Ruzanova (Sber AI); Omar Zoloev (Sber AI Lab); Andrey Savchenko (Sber AI Lab); Maksim Makarenko (Sber AI Lab)

Nowadays, the success of financial organizations heavily depends on their ability to process digital traces generated by their clients, e.g., transaction histories, gathered from various sources to improve user modeling pipelines. As general-purpose LLMs struggle with time-distributed tabular data, production stacks still depend on specialized tabular and sequence models that are limited in transferability and require labeled data. To address this, we introduce FinTRACE, a retrieval-first architecture that converts raw

transactions into reusable feature representations, applies rule-based detectors, and stores the resulting signals in a behavioral knowledge base with graded associations to the objectives of downstream tasks. Across public and industrial benchmarks, FinTRACE substantially improves low-supervision transaction analytics, doubling zero-shot MCC on churn prediction performance from 0.19 to 0.38 and improving 16-shot MCC from 0.25 to 0.40. We further use FinTRACE to ground LLMs via instruction tuning on retrieved behavioral patterns, achieving state-of-the-art LLM results on transaction analytics problems.

P104 · **RAPO: Reason-Aware Preference Optimization for Factual Table Question Answering over Structured Data** [Industry]

Phaniram Sayapaneni (Walmart Global Tech); Musen Wen (Walmart Global Tech); Sunil Goda (Walmart Global Tech)

Large Language Models (LLMs) are increasingly deployed in interactive information access systems over structured data, enabling natural language querying of large-scale enterprise tables in domains such as digital advertising. However, ensuring factually accurate and hallucination-free responses remains a critical challenge for reliable Table Question Answering (Table QA), particularly in production analytics workflows. While Direct Preference Optimization (DPO) has emerged as a popular approach for aligning LLMs using pairwise feedback, it overlooks the underlying rationale behind preferences, limiting the effectiveness of alignment. In this work, we introduce Reason-Aware Preference Optimization (RAPO), a novel alignment framework that incorporates reason-augmented preference signals to better guide LLM learning and improve the reliability of structured information retrieval. Evaluated on real-world advertising datasets, RAPO demonstrates substantial improvements over DPO in factual accuracy, enabling more trustworthy retrieval of business-critical insights.

P105 · **Meituan Merchant Business Diagnosis via Policy-Guided Dual-Process User Simulation** [Industry]

Ziyang Chen (Meituan Inc.); Renbing Chen (Meituan Inc.); Daowei Li (LongCat Interaction Team, Meituan Inc.); Jinzhi Liao (Independent Researcher); Jiashen Sun (Meituan Inc.); Ke Zeng (Meituan Inc.); Xiang Zhao (Independent Researcher)

Simulating group-level user behavior enables scalable counterfactual evaluation of merchant strategies without costly online experiments. However, building a trustworthy simulator faces two structural challenges. First, information incompleteness causes reasoning-based simulators to over-rationalize when unobserved factors such as offline context and implicit habits are missing. Second, mechanism duality requires capturing both interpretable preferences and implicit statistical regularities, which no single paradigm achieves alone. We propose Policy-Guided Hybrid Simulation (PGHS), a dual-process framework that mines transferable decision policies from behavioral trajectories and uses them as a shared alignment layer. This layer anchors an LLM-based reasoning branch that prevents over-rationalization and an ML-based fitting branch that absorbs implicit regularities. Group-level predictions from both branches are fused for complementary correction. We deploy PGHS on Meituan with 101 merchants and over 26,000 trajectories. PGHS achieves a group simulation error of 8.80%, improving over the best reasoning-based and fitting-based baselines by 45.8% and 40.9% respectively.

P106 · **Efficient Multi-Cohort Inference for Long-Term Effects and Lifetime Value in A/B Testing with User Learning** [Industry]

Dario Simionato (Huawei Ireland Research Centre); Andrea Tonon (Huawei Ireland Research Centre); Mingxue Wang (Huawei Ireland Research Centre); Weiguo Wang (Huawei Nanjing Research & Development Center); Tong Gui (Huawei Nanjing Research & Development Center); Xiaoyue Li (Huawei Nanjing Research & Development Center)

In streaming platforms churn is extremely costly, yet A/B tests are typically evaluated using outcomes observed within a limited experimental horizon. Even when both short- and predicted long-term engagement metrics are considered, they may fail to capture how a treatment affects users' retention. Consequently, an intervention may appear beneficial in the short term and neutral in the long term while still generating lower total value than the control due to users churn. To address this limitation, we introduce a method that estimates long-term treatment effects (LTE) and residual lifetime value change (Δ ERLV) in short multi-cohort A/B tests under user learning. To estimate time-varying treatment effects efficiently, we introduce an inverse-variance weighted estimator that combines multiple cohorts estimates, reducing variance relative to standard approaches in the literature. The estimated treatment trajectory is then modeled as a parametric decay to recover both the asymptotic treatment effect and the cumulative value generated over time. Our framework enables simultaneous evaluation of steady-state impact and residual user value within a single experiment. Empirical results show improved precision in estimating LTE and Δ ERLV and identify scenarios in which relying on either short-term or long-term metrics alone would lead to incorrect product decisions.

P107 · **Precision at Scale: An End-to-End Graph-based Framework for Mitigating Network Interference in TikTok A/B Tests** [Industry]

Yuhan Li (ByteDance Inc); Jiayun Ni (ByteDance Inc); Ao Li (ByteDance Inc); Yue Wang (ByteDance Inc); Seng You Paul Chee (ByteDance Inc); Yongmin Hu (ByteDance Inc); Runze Yu (ByteDance Inc)

Large-scale A/B tests on social platforms suffer from inherent network interference, violating the Stable Unit Treatment Value Assumption (SUTVA) and distorting measured results. Industrial practices for mitigating network interference face a foundational trade-off. For scalability, they often rely on clustering static graphs, which serve as imperfect proxies for true interference pathways. Conversely, theoretically sound methods remain computationally intractable at production scale. This paper presents a production-ready framework deployed at TikTok, which integrates three core contributions to address these challenges: 1) Learned Interference Graph (LIG): Estimates interference probabilities using dynamic interaction patterns for more context-aware modeling. 2) Scalable Community Partitioning (SCP): A Spark-optimized ParLeiden implementation that performs billion-node graph clustering daily and generalizes effectively across diverse interaction types, achieving a purity score of 0.898 for group chat interactions. 3) Sensitivity-Enhanced Estimation (SEE): A multivariate system leveraging Controlled-experiment Using Pre-Experiment Data (CUPED) to mitigate variance inflation from cluster-based randomization. In live production tests, our framework reduces interference rates by 68.8%, correcting a biased treatment effect estimate from +1.44% to a statistically significant +2.08%. It also enables previously undetectable cross-ecosystem measurements, revealing a +0.2% lift in creator upload volumes driven by user-side treatments.

P108 · **LiBRA: Enhancing Live-streaming Recommendation with Bigraph-based Capacity-aware Ranking Approach** [Industry]

Youran Zhang (Bytedance Inc.); Xiaotong Guo (Bytedance Inc.); Jiayi Lu (ByteDance Inc.); Runze Yu (ByteDance Inc.); Zhuang Cai (ByteDance Inc.); Qinglei Wang (ByteDance Inc.)

Live-streaming recommendation systems effectively match viewers' interests with relevant streamer content by ranking streamer candidates based on predicted relevance scores. However, existing works aggressively push popular streamers' content to match viewers' past interests, resulting in uneven traffic allocation. This creates a dual dilemma: viewers tend to concentrate on popular streamers, leading to reduced quality of experience due to streamers' limited capacity, while non-popular streamers suffer from traffic starvation. This paper formulates the viewer-streamer match problem as a capacity-aware optimal subgraph problem on the viewer-streamer bipartite graph, and proposes a novel capacity-aware ranking approach for live-streaming recommendation systems with explicit consideration of streamers' limited service capacity. Initially, we perform reinforcement learning to estimate the streamer capacity. Then, a novel graph algorithm FIMWS is introduced to solve the capacity-aware viewer-streamer match problem from a globally optimal perspective, with its theoretical optimality formally proven. The outcome of the proposed graph algorithm is finally distilled via deep models for point-wise online ranking. Currently, this capacity-aware ranking approach has been successfully deployed in TikTok Live. Extensive experiments on TikTok Live show that the proposed approach can significantly improve bilateral experience compared with other alternatives: +0.9% watch duration for viewers and +2% live duration for non-popular streamers.

P109 · **Variance Reduction for Heavy-Tailed Monetization Metrics in Ranking Experiments via Post-Stratification** [Industry]

Neeti Pokharna (ShareChat); Olivier Jeunen (Aampe); Yatharth Saraf (ShareChat); Aleksei Ustimenko (Simulacra AI)

Online evaluation of ranking and retrieval systems often relies on downstream monetization metrics such as app revenue or creator earnings. These metrics are typically heavy-tailed, with a small fraction of users dominating both mean and variance, leading to low statistical power and unreliable conclusions in A/B experiments---especially under limited traffic. We present a practical framework for variance reduction in online experiments by combining post-stratification with CUPED. Our approach leverages pre-experiment covariates to improve the sensitivity of monetization experiments without requiring additional traffic. Deployed at ShareChat across ranking-driven monetization experiments, the method substantially reduces variance and improves decision stability, achieving equivalent statistical confidence with ~45% less traffic than standard metrics. We further discuss practical design choices, guardrails, and limitations, providing guidance on when post- stratification is appropriate for real-world information retrieval and Recommendation systems.

P110 · **SCOPE: Scalable Cross-Task Orthogonal Progressive Experts for Multi-Task Learning in Recommendations** [Industry]

Zixian Yang (Ant Group); Wei Xu (Ant Group); Li Li (Ant Group); Zhaokai Huang (Ant Group); You Li (Ant Group); Jianbin Lin (Ant Group); Wenliang Zhong (Ant Group); Can Ye (Ant Group)

Online short-video recommendation systems, such as Alipay's Tab3, must model heterogeneous user feedback ranging from implicit playback signals to explicit interactions and domain-specific actions. This diversity poses significant challenges for Multi-Task Learning (MTL). Existing Mixture-of-Experts (MoE) approaches often suffer from knowledge entanglement due to indirect sharing, knowledge redundancy among correlated tasks, and scalability bottlenecks caused by super-linear computational growth as the number of tasks increases. To address this, we propose SCOPE (Scalable Cross-Task Orthogonal Progressive Experts). This framework features: (1) a hierarchical architecture utilizing omni-specific experts for efficient knowledge transfer; (2) a Transformer-based module to capture inter-task dependencies; and (3) dual regularization that ensures expert diversity and task alignment. Validated on public and industrial datasets, SCOPE outperforms state-of-the-art methods. Notably, it has been deployed on Alipay's Tab3, achieving a +0.61% Group Area Under the ROC Curve (GAUC) gain and +2.47% video views while reducing Floating-Point Operations (FLOPs) by 43%.

P111 · **Generative Enhanced Modeling: A Collaborative Framework for Enhancing User Representations via Semantic ID** [Industry]

Li Li (Ant Group); Wei Xu (Ant Group); Yu Cheng (Ant Group); Jianbin Lin (Ant Group); Can Ye (Ant Group)

User interest modeling is foundational to recommender systems. However, sparse and noisy behaviors make traditional item-level sequence models brittle, especially for new and low-activity users. Furthermore, relying solely on a user's own history limits exploration and reinforces the "filter bubbles". To address this, we propose GEM (Generative Enhanced Modeling). GEM shifts the paradigm from self-behavior induction to collective experience migration. Specifically, it constructs LLM-based semantic IDs and embeddings. Grounded in information theory, GEM performs multi-stage denoising at both the user and item levels. This design effectively suppresses reward-driven noise while preserving target-aware signals. We deployed GEM on the Alipay Tab3 video feed. Offline evaluations show significant GAUC gains. Online A/B tests demonstrate a 0.9% lift in watch time alongside stable video views and improved exposure diversity. These results confirm that GEM enhances recommendation quality and successfully broadens user interests.

P112 · **Bridging the Gap: An End-to-End Framework for Decoupled Alignment in Dynamic Semantic ID Generation** [Industry]

Yu Cheng (Ant Group); Wei Xu (Ant Group); Li Li (Ant Group); Jianbin Lin (Ant Group); Can Ye (Ant Group)

Semantic IDs derived from Multi-modal Large Language Models (MLLMs) integrate rich content semantics into recommendation systems but often lack the collaborative signals crucial for capturing user behavior. Additionally, static generation fails to adapt to evolving data distributions, causing codebook drift. To address these limitations, we propose a dynamic End-to-End Semantic ID Generation Framework based on Decoupled Representation Alignment. Our method aligns shared and private components from both MLLM content and Collaborative Filtering (CF) embeddings, integrating them via a hierarchical adaptive fusion module into a Residual Quantized Variational Autoencoder (RQ-VAE). This joint optimization promotes both semantic granularity and collaborative awareness while supporting dynamic codebook updates. Extensive offline experiments and online A/B tests on Alipay's Tab3 short-video scenario demonstrate the superiority of our method, which is now deployed to serve all users.

P113 · **Improving Dating Outcomes by Predicting for Conversations at Hinge** [Industry]

Frankie Lu (Hinge); Dan Turkel (Hinge); Sihan Chen (Hinge); Behrooz Afghahi (Hinge)

The ultimate measure of success for a dating platform is helping people go on in-person dates. In practice, ranking systems for such platforms rely on observable proxies, such as likes and matches, to predict whether an in-person date will occur. This paper introduces conversations, when both users have sent at least one message, as a stronger signal of dating intent. We model the dating journey as a probabilistic chain, decomposing the likelihood of a conversation into three sequential stages: the probability of a like, a match given a like, and a conversation given a match. By targeting conversations as a late-stage observable proxy for dating intent, this approach captures a higher-fidelity signal than using likes and matches alone. We present the production deployment of this system at Hinge via a multi-task two-tower model that jointly predicts three objectives with a shared embedding layer. User interaction sequences are encoded with a transformer architecture to capture temporal patterns. Our A/B test results demonstrate a +2.8% improvement in number of conversations and a +1.4% increase in exchange coverage.

P114 · **Gated Bidirectional Linear Attention for Generative Retrieval** [Industry]

Artem Matveev (Yandex); Vladislav Tytskiy (Yandex); Sergei Makeev (Yandex); Sergei Liamaev (Yandex)

In recommender systems, generative retrieval typically uses an encoder–decoder setup: an encoder processes a user interaction history, and an autoregressive decoder then generates recommended items. In large-scale streaming services, active users accumulate very long histories over time. As histories grow, the encoder becomes a major latency bottleneck because softmax attention scales quadratically with sequence length. In our experiments, using bidirectional attention in the encoder substantially improves quality. However, most sub-quadratic attention methods focus on causal attention. We propose Gated Bidirectional Linear Attention (GBLA), a linear-time bidirectional attention layer that extends kernelized linear attention with three lightweight components: local causal mixing (Conv1D), sequence-level key gating for soft forgetting, and a gated RMSNorm output. On a large-scale Yandex Music dataset, a hybrid encoder that interleaves self-attention (SA) and GBLA in a 1:2 ratio (one SA block followed by two GBLA blocks) matches bidirectional self-attention quality. On H100 GPUs, GBLA reaches up to an \(8.2\times\) single-layer speedup at a history length of 32768, compared to FlashAttention-v3. Finally, we show that the same hybrid design generalizes beyond our proprietary setting, consistently preserving self-attention retrieval quality on public Amazon benchmarks.

P115 · **Beyond Item IDs: Scaling Short-Form-Video Recommendation via Semantic-Native Long Sequence Modeling** [Industry]

Ruixiao Sun (Google); Diego Uribe Mora (Google); Zhimeng Jiang (Google); Yuanzhen Lin (Google); Jiarui Wang (Google); Yuening Li (Google); Danfeng Guo (Google); Zhizhong Chen (Google); Chuan He (Google); Liang Liu (Google)

Capturing user interests across extensive watch histories is critical for short-form video recommendation, yet scaling sequence length is limited by two bottlenecks: the semantic sparsity of atomic Video IDs and the quadratic computational complexity of Transformers. Traditional orthogonal Video IDs fail to capture content relationships and demand large embedding tables, while the quadratic complexity of self-attention restricts the maximum sequence length under strict industrial latency and resource constraints. In this work, we present a production-deployed framework for modeling ultra-long user behavior sequences at a billion-user scale. We first address the representation bottleneck by adopting content-native Semantic IDs. By utilizing depth-truncated, coarse-grained Semantic IDs, we shrink the embedding table size from corpus cardinality. This compact representation naturally generalizes to cold-start content through shared semantic prefixes. Second, to overcome the sequence scaling barrier, we introduce a Global-Aware Compression Transformer that leverages non-parametric temporal folding and unified global query integration to effectively condense the sequence, alleviating both the memory and computational bottlenecks of standard self-attention. Offline profiling on our computing infrastructure demonstrates an order-of-magnitude reduction in peak memory footprint and a drastic decrease in computational overhead. This efficiency gain enables supporting longer sequence lengths at an affordable cost in production, yielding substantial online gains in satisfied user engagement and satisfied content consumption in large-scale online A/B tests.

P116 · **Cross-Category Basket Recommendation via Transactional Debiasing and Graph Propagation** [Industry]

Serhat Mirkan Acar (Trendyol Group)

Selecting products for cross-sell basket recommendation is challenging due to massive candidate sets, extreme category imbalance, and co-occurrences driven by exposure rather than true complementarity. Complement discovery must therefore remove popularity artifacts while remaining robust under sparse transactions. We address this setting using large-scale purchase logs as primary evidence.Our approach learns complementary structure from purchase sequences via a debiased transactional scorer and a graph-based completion mechanism. The scorer aggregates co-purchase evidence with time- and season-aware weighting while penalizing globally frequent items and over-represented categories, enabling cross-category retrieval without manual constraints. To improve coverage for infrequent and newly introduced items, we propagate evidence over an item-similarity graph using regularized message passing.We deploy the system in production and evaluate it at industrial scale through offline analyses and an online A/B test on a checkout-adjacent basket recommendation surface. Offline results show consistent improvements in future co-purchase prediction and robustness under strong skew. In production, the online experiment demonstrates statistically significant gains in basket engagement and conversion, with positive movement in aggregate platform-level business metrics. Overall, the results highlight a scalable solution for cross-category basket recommendation by jointly modeling temporal dynamics, seasonal effects, and behavioral bias.Unlike prior basket recommenders that rely on category heuristics or downstream reranking, we construct bias-corrected item--item relations directly at evidence formation time.

P117 · **How Trendyol Enables Trustworthy A/B Testing at E-Commerce Scale Through Integrated Statistical Methodologies** [Industry]

Serhat Mirkan Acar (Trendyol Group); Damla H■zl■ Sert (Trendyol Group); Kaya Tatar (Trendyol Group)

A/B testing is the primary decision-making mechanism in large-scale e-commerce platforms. However, reliable experimentation requires more than random user assignment: heterogeneous user populations, low-sensitivity metrics, and interference from concurrent experiments can lead to misleading conclusions and costly product decisions. In this paper, we describe the statistical methodology layer developed by Trendyol's data science teams to improve both the trustworthiness and iteration speed of online experiments at scale. We first apply persona-based stratified assignment to balance behavioral distributions across experimental groups and reduce variance induced by population heterogeneity. During test execution, we continuously monitor statistical power at the metric level and expose power alongside statistical significance, enabling practitioners to detect underpowered outcomes and reduce false positives even when metrics appear significant. Finally, we enforce orthogonal assignment principles across concurrent experiments to minimize cross-experiment interference and enable safe parallel experimentation. Rather than focusing on experimentation infrastructure, we highlight compact and operational statistical design choices and formalized decision rules that strengthen experiment reliability in production environments.

P118 · **Behavioral Feature Boosting via Substitute Relationships for E-commerce Search** [Industry]

Chaosheng Dong (Amazon); Michinari Momma (Amazon); Yijia Wang (Amazon); Yan Gao (Amazon); Yi Sun (Amazon)

On E-commerce platforms, new products often suffer from the cold-start problem: limited interaction data reduces their search visibility and hurts relevance ranking. To address this, we propose a simple yet effective behavior feature boosting method that leverages substitute relationships among products (BFS). BFS identifies substitutes—products that satisfy similar user needs—and aggregates their behavioral signals (e.g., clicks,

add-to-carts, purchases, and ratings) to provide a warm start for new items. Incorporating these enriched signals into ranking models mitigates cold-start effects and improves relevance and competitiveness. Experiments on a large E-commerce platform, both offline and online, show that BFS significantly improves search relevance and product discovery for cold-start products. BFS is scalable and practical, improving user experience while increasing exposure for newly launched items in E-commerce search. The BFS-enhanced ranking model has been launched in production and has served customers since 2025.

P119 · **Preliminary Study of an Evaluation Benchmark for Vision–Language Models in Fashion E-Commerce** [Industry]

Ryotaro Shimizu (ZOZO Research); Sai Htaungkham (ZOZO Research); Shion Sakurai (ZOZO, Inc.); Yuki Shimizu (ZOZO, Inc.)

We report an evaluation benchmark for assessing the operational suitability of Vision-Language Models (VLMs) in fashion e-commerce. General-purpose benchmarks do not adequately cover fashion-specific attributes or the structured extraction tasks common in e-commerce workflows. We define five tasks across two image streams---outfit and single-item product images---and compare six commercial and two open-source models with multiple prompt variants, including a canonical prompt and model-proposed prompts. Experiments show that the best-performing model varies by task, error patterns are more model-dependent than prompt-dependent, and model updates can improve some tasks while degrading others. These results indicate that task-specific evaluation, prompt robustness checks, and continuous monitoring are practical requirements for deploying VLMs in production fashion systems.

P120 · **Enhancing Enterprise Assistant Responses with Rich Multimodal Artifacts from Product Documentation** [Industry]

Sohan Patnaik (Adobe); Sai Sree Harsha (Adobe); Kowndinya Renduchintala (Adobe); Milan Aggarwal (Adobe); Sumit Bhatia (Adobe); Yunyao Li (Adobe)

Modern enterprise documentation is rich with images (screenshots, flowcharts, system diagrams, etc.) and instructional videos (product tutorials), yet typical conversational systems often provide text-only answers that fail to capture procedural or interface-level nuances. We present our in-production retrieval framework designed to enrich conversational responses by effectively surfacing these multimodal artifacts. Our solution improves multimodal retrieval over enterprise documentation by combining two complementary signals: (1) direct multimodal retrieval over captions and transcripts, and (2) retrieval-conditioned multimodal expansion, which identifies multimodal artifacts linked to semantically relevant documentation passages. The retrieved multimodal artifacts are surfaced directly in the final response to the user query, enabling richer and more actionable answers grounded in documentation. Evaluation on two enterprise-scale multimodal documentation benchmarks derived from real-world queries shows that \approachName\ consistently outperforms text-only and multimodal retrieval baselines, improving Hit@10 by over 22% and Recall@10 by over 20%. These results demonstrate that jointly leveraging direct multimodal retrieval and text-guided expansion improves the reliability and completeness of multimodal retrieval, enabling conversational systems to effectively surface relevant images and instructional videos from documentation, thereby significantly enhancing enterprise QA systems.

P121 · **What Gets Cited: Competitive GEO in AI Answer Engines** [Industry]

Rahul Vishwakarma (Sprinklr); Shushant Kumar (Sprinklr); Ratnesh Jamidar (Sprinklr)

AI answer engines generate answers from retrieved pages but cite only a few sources. This makes visibility depend not just on ranking, but on being cited. We study competitive Generative Engine Optimization (GEO): when two retrieved candidates compete, what makes one more likely to be cited first? We build a controlled two-document retrieval-augmented generation (RAG) testbed that injects exactly two candidate sources into the model context and measures which source is referenced by the first citation marker in the output. Across six LLMs we execute 252,000 trials, repeated paired comparisons under one factorial program over 18 content factors. In each trial the two sources differ in exactly one factor; we use brand anonymization and counterbalanced source order to separate content effects from position bias. Mixed-effects models show that topical relevance and list position are the biggest drivers of being cited first. Including explicit price information and a recent timestamp also helps consistently. Completeness and trust cues add smaller gains, while formatting-only edits have little impact. We release a reproducible evaluation protocol and a prioritized GEO checklist for practitioners, and we exercised it in an early internal pilot at Sprinklr, where teams reported positive qualitative feedback on workflow usability.

P122 · **From Queries to Playlists: An LLM-Driven Architecture for Semantic Music Search at Scale** [Industry]

Rinat Mullakhmetov (Zvuk); Fedor Buzaev (Zvuk); Roman Bogachev (Zvuk); Ilya Sedunov (Zvuk); Oleg Pavlovich (Zvuk); Kamil Mazitov (Zvuk); Vladimir Kravtsov (Zvuk); Elena Tutubalina (HSE University); Daria Pugacheva (HSE University); Ivan Sukharev (Zvuk)

A core task for music streaming platforms is retrieving and ranking tracks in response to user queries over multi-million-track catalogs. Existing approaches either rely on tag-based, entity-centric retrieval and recommendation, which struggle with implicit and subjective queries that fall outside a predefined tag vocabulary, or on recent LLM-based generative methods that circumvent this limitation but are prone to hallucinations and factual errors. We introduce a semantic playlist generation service that retrieves and ranks tracks based on meaning rather than keyword overlap, while avoiding hallucination-related failures. Each track is represented as structured text combining metadata, lyrics, and descriptive attributes, and both user queries and track representations are encoded into a shared embedding space using an LLM. A cross-encoder reranker built on the same backbone refines candidate ranking, and its signals are distilled into the embedder to reduce serving cost. In offline and production evaluations, our semantic vector-search pipeline achieves the highest playlist quality, improving Precision@10 from 64% with faceted search and 74% with direct LLM generation to 81%, while remaining compatible with low-latency, large-scale deployment. In an online A/B test on smart-speaker traffic, routing a share of playlist requests to our system yields a consistent double-digit relative uplift in Average Time Spent, indicating that meaning-aware retrieval substantially enhances user engagement and supports broader production rollout.

P123 · **All the News That Fits in Bits: Learned Rotation-Aware Binary Projections for Efficient News Retrieval at NDTV** [Industry]

Ritwick Ghosh (NDTV)

Embedding-based retrieval powers recommendation and discovery features across modern news platforms, yet serving dense float vectors at scale introduces substantial latency and memory costs. We present a two-stage retrieval pipeline deployed at NDTV, one of India's largest news organizations, that replaces exhaustive float search with a fast binary shortlist followed by a lightweight float rerank. At the core of our approach is a learned binary projection trained with a contrastive objective to preserve the neighbor structure of the original embedding space. We evaluate the pipeline on a corpus of 47,300 news articles across three embedding models (Gemma, Nomic, Qwen) and two production-motivated retrieval scenarios: a \emph{top-20 recommendation} task requiring $\geq 98\%$ recall, and a \emph{top-5 precision-critical} task demanding near-perfect accuracy. Our results show that even moderate bit-widths (1024 bits) recover $\geq 95\%$ of float-exact neighbors, while our recommended operating point of 3072 bits achieves 98% top-20 recall through a 50-candidate shortlist and near-perfect ($\geq 99.9\%$) top-5 recall through a 200-candidate shortlist. Notably, RaBitQ-style scalar corrections---widely adopted in the literature---consistently failed to improve and often degraded retrieval quality for both learned and raw binary codes.At million-scale (1M documents), the binary shortlist delivers 4--$6\times$ speedup at 3072 bits and up to $70\times$ at lower bit-widths on a single 16-vCPU cloud instance with no GPU, reducing per-batch latency from 2.5--3.5\,s to under 600\,ms. Prior to deployment, an editorial acceptance evaluation by five senior NDTV editors confirmed that the binary pipeline's recommendations are indistinguishable from float-exact results across diverse news topics. We further introduce \emph{rotation-aware binary training}, a new training scheme that bakes a random orthogonal transform into the learned projection, consistently improving recall at higher bit-widths while narrowing the train--eval generalization gap. These findings generalize across all three embedding families with no model-specific tuning.

P124 · **HybridSparse: An End-to-End Hybrid Framework for Efficient Large-Scale Retrieval** [Industry]

Haotong Bao (Shanghai Jiao Tong University); Jianjin Zhang (Microsoft); Weihao Han (Microsoft); Xue Wu (Microsoft); Qi Chen (Microsoft); Dongzhe Jiang (Microsoft); Zhengxin Zeng (Microsoft); Mingzheng Li (Microsoft); Hao Sun (Microsoft); Weiwei Deng (Microsoft); Feng Sun (Microsoft); Qi Zhang (Microsoft)

Large-scale retrieval systems must operate under strict latency constraints while maintaining high recall. Sparse retrieval offers efficiency and interpretability, whereas dense retrieval provides stronger semantic matching. Although hybrid approaches combine both signals, their interaction is often limited, especially under intersection-based retrieval. We introduce HybridSparse, an end-to-end hybrid retrieval framework that strengthens sparse--dense interaction across modeling, training, and serving. It adopts a unified encoder with a shared backbone and jointly optimizes lexical and semantic representations through co-training. To further improve alignment, we incorporate hybrid score regularization and consistency distillation, enabling more stable and effective hybrid scoring. Experiments on public benchmarks demonstrate consistent improvements over strong sparse, dense, and hybrid baselines. In large-scale production deployment for Bing advertisement retrieval, HybridSparse delivers a +1.30% RPM gain, highlighting its practical impact.

P125 · **Surface-Form Neural Sparse Retrieval: Robust Fuzzy Matching for Industrial Music Search** [Industry]

Paul Greyson (Amazon); Zhichao Geng (Amazon); Wei Zhang (Amazon); Yang Yang (Amazon)

Music search at the scale of Amazon Music presents a unique challenge: queries frequently deviate from indexed metadata due to misspellings, transpositions, and phonetic variations, yet the retrieval system must operate under strict millisecond-level latency constraints. Our existing learning-to-retrieve system, the High Confidence Index (HCI), learns query-entity associations from customer behavior, relying on continual "exploration" to choose candidates. Traditional n-gram matching enables this exploration but suffers from poor semantic robustness and high noise, limiting the system's ability to learn from long-tail queries. In this work, we present a robust neural sparse retrieval system designed to maximize exploration efficiency. We adapt a state-of-the-art inference-free sparse retrieval architecture to the music domain, combining it with an effective domain-specific granular subword tokenization strategy. Our approach utilizes short-length token constraints (max 3 chars) to enforce the learning of

surface-form robustness over lexical memorization. By pre-computing the neural embeddings and term expansions during the offline indexing phase, online processing is reduced to minimal tokenization and IDF weighting, achieving effectively zero latency overhead for query encoding. Evaluations on a 6M-document production corpus show an aggregate 91.4% recall@10 (vs. 57.7% for trigrams) at comparable throughput. Simulation of the HCI feedback loop demonstrates improved exploration efficiency, with +0.8% higher stabilized recall than production trigrams. Ablation studies indicate that our sparse training methodology drives the performance gains, while domain-specific pretraining provides a cost-effective alternative to large-scale general-purpose pretraining.

P126 · User Activity Modeling under Inflated Distribution [Industry]

Xiang Li (ByteDance); Zhao-Yu Zhang (ByteDance); Chunyuan Zheng (ByteDance); Qingying Chen (ByteDance); Huiyou Jiang (ByteDance); Haoxuan Li (Peking University); Zhouchen Lin (Peking University)

User activity modeling serves as an effective approach to reflect Daily Active Users (DAU) levels by forecasting users' prospective behavioral engagement. In industrial practice, particularly in user active days prediction tasks, the data demonstrates extreme distributional imbalance. Specifically, completely inactive and fully active users constitute the vast majority, resulting in a highly skewed phenomenon known as a two-sided inflated distribution. Existing methods either rely on the assumption of a normal distribution, neglecting the optimization objectives to focus purely on designing complex model architectures for user feature extraction, which renders them ineffective in non-normally distributed real-world industrial scenarios, or they have explored modeling non-normal distributions, but still struggle to directly operate on the discrete data space in user activity modeling and mitigate the two-sided inflation prevalent in industrial data. Based on this finding, we propose a simple yet effective user activity modeling method under two-sided inflated distribution. Specifically, we decompose the problem into two objectives: first, modeling whether the target is at an inflated point; and second, modeling the distribution when the target is not at an endpoint. Finally, we apply this modeling approach to uplift modeling. Extensive offline evaluations and online A/B tests on Douyin platforms, involving over 1 billion users, demonstrate the effectiveness of the proposed method.

P127 · SkillForge: Forging Domain-Specific, Self-Evolving Agent Skills in Cloud Technical Support [Industry]

Xingyan Liu (Alibaba Cloud Computing); Xiyue Luo (Alibaba Cloud Computing); Linyu Li (Alibaba Cloud Computing); Ganghong Huang (Alibaba Cloud Computing); Jianfeng Liu (Alibaba Cloud Computing); Honglin Qiao (Alibaba Cloud Computing)

Deploying LLM-powered agents in enterprise scenarios such as cloud technical support demands high-quality, domain-specific skills. However, existing skill creators lack domain grounding, producing skills poorly aligned with real-world task requirements. Moreover, once deployed, there is no systematic mechanism to trace execution failures back to skill deficiencies and drive targeted refinements, leaving skill quality stagnant despite accumulating operational evidence. We introduce SkillForge, a self-evolving framework that closes an end-to-end creation–evaluation–refinement loop. To produce well-aligned initial skills, a Domain-Contextualized Skill Creator grounds skill synthesis in knowledge bases and historical support tickets. To enable continuous self-optimization, a three-stage pipeline---Failure Analyzer, Skill Diagnostician, and Skill Optimizer---automatically diagnoses execution failures in batch, pinpoints the underlying skill deficiencies, and rewrites the skill to eliminate them. This cycle runs iteratively, allowing skills to self-improve with every round of deployment feedback. Evaluated on five real-world cloud support scenarios spanning 1,883 tickets and 3,737 tasks, experiments show that: (1) the Domain-Contextualized Skill Creator produces substantially better initial skills than the generic skill creator, as measured by consistency with expert-authored reference responses from historical tickets; and (2) the self-evolution loop progressively improves skill quality from diverse starting points (including expert-authored, domain-created, and generic skills) across successive rounds, demonstrating that automated evolution can surpass manually curated expert knowledge.

P128 · Unified Semantic Modeling Framework for Large-Scale Job Understanding at LinkedIn [Industry]

Dan Xu (LinkedIn Corporation); Baofen Zheng (LinkedIn Corporation); Jianqiang Shen (LinkedIn Corporation); Qi Xiao (LinkedIn Corporation); Benjamin Hoan Le (LinkedIn Corporation); Wen Pu (LinkedIn Corporation); Saurabh Gupta (LinkedIn Corporation); Ran Zhou (LinkedIn Corporation); Neha Saraf (LinkedIn Corporation); Alice Leung (LinkedIn Corporation); Qianqi Shen (LinkedIn Corporation); Liangjie Hong (LinkedIn Corporation); Jingwei Wu (LinkedIn Corporation); Wenjing Zhang (LinkedIn Corporation)

Job understanding is critical to LinkedIn's mission of connecting talent with opportunity. This task involves transforming unstructured and noisy job postings into standardized or derived job attributes that power numerous LinkedIn products. However, building a scalable, cost-efficient, and high-performing job understanding system remains challenging. In this paper, we present a unified semantic modeling framework powered by a small language model (SLM) to address the challenges. We begin by fine-tuning an open-source SLM using a suite of carefully curated synthetic tasks augmented with reasoning traces. These tasks jointly target taxonomy-guided classification and taxonomy-agnostic entity extraction. This allows the resulting model to acquire robust zero-shot generalization for job understanding in structured and unstructured contexts. Building upon this foundation, we introduce a multi-adapter architecture with attribute grouping to facilitate efficient task-specific adaptation while streamlining model management across diverse downstream attributes. Offline evaluations and online A/B tests demonstrate significant performance improvement while reducing operational complexity. Our work provides practical insights into building industry-scale text understanding systems.

P129 · Policy-Grounded Dynamic Facet Suggestions for Job Search [Industry]

Dan Xu (LinkedIn Corporation); Baofen Zheng (LinkedIn Corporation); Qianqi Shen (LinkedIn Corporation); Jianqiang Shen (LinkedIn Corporation); Wenqiong Liu (LinkedIn Corporation); Chunnan Yao (LinkedIn Corporation); Ping Liu (LinkedIn Corporation); Rajat Arora (LinkedIn Corporation); Kevin Kao (LinkedIn Corporation); Hsiang Lin (LinkedIn Corporation); Wanjun Jiang (LinkedIn Corporation); Yusuke Takebuchi (LinkedIn Corporation); Jingwei Wu (LinkedIn Corporation); Wenjing Zhang (LinkedIn Corporation)

Job seekers often initiate search with short, underspecified queries. At LinkedIn, over 80% of job-related queries contain three or fewer keywords, making accurate user intent inference and relevant job retrieval particularly challenging. We present dynamic facet suggestion (DFS), an interactive query-refinement mechanism that facilitates intent disambiguation by surfacing personalized semantic attributes conditioned on the joint user-query context in real time. We propose a policy-grounded, retrieval-augmented ranking framework for facet suggestion, comprising offline taxonomy curation, embedding-based retrieval of top-K candidates, and a distilled small language model (SLM) based candidate scoring. The system is optimized for real-time serving via point-wise single-token scoring and batching/prefix caching. Offline evaluation demonstrates high precision for generated suggestions, and online A/B tests show significant lifts in suggestion engagement and job search outcomes.