Pre-Trained Models for Search and Recommendation: Introduction to the Special Issue—Part 1

1 Introduction

The emergence of pre-trained models, particularly **large language models (LLMs)** like GPT-4 [17] and Llama [6], has revolutionized the field of information retrieval, driving unprecedented advancements in search and recommendation systems. LLM-empowered AI search [1, 2, 10], based on **retrieval-augmented generation (RAG)** paradigms, shows the potential to transform traditional search systems. Recommender systems have also greatly benefited from pre-trained models, ranging from BERT [3] and T5 [19] to LLMs like GPT [17] and Llama [6], all of which have profoundly influenced the technological evolution across various recommendation tasks [14, 20, 28]. Generally speaking, by leveraging their remarkable representation, reasoning, and generalization capabilities, pre-trained models have introduced promising solutions to longstanding challenges in search and recommendation [11, 13, 21, 23]. Their applications span dense retrieval, neural ranking, user modeling, content generation, and evaluation, garnering significant attention from both academia and industry. However, despite promising progress, integrating pre-trained models effectively into search and recommendation tasks still faces many unresolved challenges, especially concerning foundational paradigm exploration, robustness in diverse scenarios, and trustworthiness.

This special issue aims to explore the dynamic interplay between pre-trained models and search and recommendation tasks, offering a platform to present significant innovations. For search systems, we emphasize advancements in leveraging pre-trained models across different stages of search systems and diverse real-world scenarios, as well as exploring new retrieval paradigms and evaluation frameworks. For recommendation systems, the focus includes leveraging pre-trained models to enhance user modeling, recommendation accuracy, personalized content generation, evaluation, and trustworthiness aspects. Through this special issue, we hope to catalyze groundbreaking research and deliver valuable insights in search and recommendation domains.

CCS Concepts: • **Information systems** → **Retrieval models and ranking**; **Recommender systems**;

Additional Key Words and Phrases: Pre-trained Models, Search, Recommendation, Large Language Models, Trustworthiness

ACM Reference format:
Wenjie Wang, Zheng Liu, Fuli Feng, Zhicheng Dou, Qingyao Ai, Grace Hui Yang, Defu Lian, Lu Hou, Aixin Sun, Hamed Zamani, Donald Metzler, and Maarten de Rijke. 2025. Pre-Trained Models for Search and Recommendation: Introduction to the Special Issue—Part 1. *ACM Trans. Inf. Syst.* 43, 2, Article 27 (February 2025), 6 pages.
https://doi.org/10.1145/3709134

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2025 Copyright held by the owner/author(s).
ACM 1558-2868/2025/2-ART27
https://doi.org/10.1145/3709134

2 Overview of Articles

The submission deadline for this special issue was March 31, 2024. A total of 60 valid submissions were received. We introduce 14 accepted papers in Part 1, with the remaining accepted papers to be included in the forthcoming Part 2. This issue explores various topics related to pre-trained models for search and recommendation, featuring seven papers on search and seven on recommendation.

2.1 Search

This search session introduces seven accepted submissions, comprising four research papers focused on enhancing pre-trained models for various search applications through novel algorithms and architectures. The remaining three papers explore RAG, introducing new mechanisms to improve the quality and cost-effectiveness of RAG systems.

As for pre-trained models for search, Guo et al. [7] identify the problem that traditional ensemble method cannot effectively leverage the diverse matching patterns during the training process. To address this problem, the authors propose a novel architecture based on mixture-of-expert. The new model utilizes shared semantic layers, specialized experts, and a competitive learning mechanism to enhance expertise, leading to state-of-the-art performances for both in-domain and out-of-domain retrieval tasks. Zhang et al. [27] focus on efficient code search. It introduces a novel approach that enhances semantic coherence directly at the code and query representation levels, thus improving alignment without extensive processing. The empirical study demonstrates its significant performance gains across a variety of settings, with the implementation made publicly available on GitHub. Ge et al. [5] address the challenge associated with the accurate reasoning of ambiguous user intents in multi-modal queries for composed image retrieval. It introduces the IUDC model, combining LLM-based triplet augmentation, dual semantic-visual matching channels, and probabilistic intent encoding to enhance intent reasoning and visual alignment. The approach achieves state-of-the-art performance, leveraging synthetic data and multi-modal fusion for superior retrieval accuracy. Finally, Parastoo et al. [8] focus on solving factoid entity questions by effectively leveraging textual relationships and semantic similarities in knowledge graphs. It uses a two-step process, Triple Retrieval and Answer Selection, where knowledge graph embeddings are employed to effectively align question and answer entities.

While for RAG, Li et al. [9] aim to improve the code generation tasks by enhancing the understanding of code structure and semantics. It proposes **Code Assistant via Retrieval-augmented Language Model (CONAN)**, which combines a code structure-aware retriever and a dual-view code representation mechanism for more effective code generation. Experimental results show that CONAN outperforms previous models, effectively assisting code generation by providing relevant code snippets and documentation while filtering out unnecessary information. Mao et al. [16] tackle the inefficiency and inaccuracy of black-box RAG systems, which struggle with irrelevant factual information and excessive token usage. The paper introduces FIT-RAG, a novel framework that improves factual retrieval with a bi-label document scorer and reduces token usage through a self-knowledge recognizer and sub-document-level token reduction. FIT-RAG demonstrates superior effectiveness and efficiency, significantly boosting Llama2-13B-Chat's accuracy across multiple datasets while halving token consumption. Lyu et al. [15] focus on improving the evaluation of RAG systems. It introduces a comprehensive benchmark based on CRUD actions, i.e., Create, Read, Update, and Delete, thus spanning diverse RAG application scenarios with dedicated datasets. By analyzing the impact of key RAG components like retrievers, context length, and knowledge base construction, the study provides actionable insights for optimizing RAG systems across various use cases.

2.2 Recommendation

For recommendation, one survey paper explores how to leverage LLMs for recommendation [12], two papers focus on learning fine-grained user intention representations [22, 24], two papers address noise issues by robust representation learning [4, 18], and two papers investigate using graphs to learn higher-order representations [25, 26].

Lin et al. [12] provide a comprehensive survey on the integration of LLMs into recommender systems. This paper explores where and how LLMs can be adapted within recommendation pipelines, from feature engineering, scoring, and ranking to user interaction. It categorizes the adaptation strategies based on whether LLM parameters are fine-tuned during training and whether conventional models are involved during inference. Lastly, it discusses the key challenges and future directions in this field.

Two papers focus on advancing the learning of fine-grained user intention representation for recommendation [22, 24]. Wang et al. [24] propose to disentangle user intention representation for sequential recommendation by the AutoDisenSeq model. AutoDisenSeq leverages neural architecture search to automate the design of attention mechanisms, tailoring the search space to disentangle user intentions effectively. The proposed AutoDisenSeq-LLM further incorporates LLMs to refine candidate recommendations, showing significant performance improvements over existing methods in diverse scenarios. Besides, Wang et al. [22] present a fine-grained pre-training approach to generate multiple user preference factors for fine-grained representation learning. This approach improves user representation learning by addressing the negative transfer problem and providing a more precise alignment of user preferences across domains.

Two papers address noise issues through robust representation learning. Di et al. [4] focus on denoising recommendation with generative models. By employing a diffusion augmentation strategy and a guided denoising process, this work ensures diversity in the latent data distribution and suppresses noise during the generation process. It demonstrates how generative models can address sparsity while preserving recommendation quality. Furthermore, Peng et al. [18] target the integration of textual semantics into GNN-based recommendation models. It integrates structural representations from GNNs with textual embeddings from LLMs. Through a denoising contrastive learning scheme, this work enhances the robustness of representations and captures intricate user–item interactions.

The last two papers explore the use of graphs to learn higher-order representations. The first tackles the limitations of high-order propagation in heterogeneous GNNs [26]. It enhances academic paper recommendations by introducing low-pass propagation through relation-aware GNNs, where the user–user and item–item relation graphs are constructed by side information like common authors, venues, and text embeddings from pre-trained models. The second paper addresses the issue of noisy data in contrastive learning-based recommendation models [25]. It introduces a dual graph augmentation framework that combines topological and semantic adaptations with structural optimization to create contrasting views. By reconstructing adjacency matrices and employing PageRank-based node masking, this method filters noise while preserving data semantics, leading to superior collaborative filtering performance.

3 Conclusion

In summary, this special issue contains a variety of studies on the application of pre-trained models to search and recommendation, covering both survey papers and technical studies. The topics include, but are not limited to, leveraging pre-trained models for fine-grained user representation learning, denoising learning, and higher-order representation learning in recommendation, as well

as multi-stage and multi-modal retrieval, and RAG optimization in search. Meanwhile, there remain many unexplored avenues in this direction. We will introduce more papers in Part 2.

Acknowledgments

We extend our gratitude to the researchers who contributed their work to this special issue and to the reviewers who dedicated considerable time and effort to offering insightful comments. Our sincere thanks also go to Prof. Min Zhang, the Editor-in-Chief of TOIS, and Clarissa Nemeth, the journal administrator, for their invaluable guidance and support.

Wenjie Wang
National University of Singapore, Singapore, Singapore

Zheng Liu
Beijing Academy of Artificial Intelligence, Beijing, China

Fuli Feng
University of Science and Technology of China, Hefei, China

Zhicheng Dou
Renmin University of China, Beijing, China

Qingyao Ai
Tsinghua University, Beijing, China

Grace Hui Yang
Georgetown University, Washington, District of Columbia, USA

Defu Lian
University of Science and Technology of China, Hefei, China

Lu Hou
Huawei Technologies Co Ltd, Shenzhen, China

Aixin Sun
Nanyang Technological University, Singapore, Singapore

Hamed Zamani
University of Massachusetts Amherst, Amherst, Massachusetts, USA

Donald Metzler
Google Inc., Mountain View, California, USA

Maarten de Rijke
University of Amsterdam, Amsterdam, The Netherlands

References

[1] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *Proceedings of the 36th International Conference on Neural Information Processing System (NIPS '22)*. Curran Associates Inc.

[2] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. Retrieved from OpenReview.net

[3] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[4] Yicheng Di, Hongjian Shi, Xiaoming Wang, Ruhui Ma, and Yuan Liu. 2024. Federated recommender system based on diffusion augmentation and guided denoising. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 31, 1–36.

[5] Hongfei Ge, Yuanchun Jiang, Jianshan Sun, Kun Yuan, and Yezheng Liu. 2024. LLM-enhanced composed image retrieval: An intent uncertainty-aware linguistic-visual dual channel matching model. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 37, 1–30.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. arXiV:2407.21783. Retrieved from https://arxiv.org/abs/2407.21783

[7] Jiafeng Guo, Yinqiong Cai, Keping Bi, Yixing Fan, Wei Chen, Ruqing Zhang, and Xueqi Cheng. 2024. CAME: Competitively learning a mixture-of-experts model for first-stage retrieval. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 35, 1–25.

[8] Parastoo Jafarzadeh, Faezeh Ensan, Mahdiyar Ali Akbar Alavi, and Fattane Zarrinkalam. 2024. A knowledge graph embedding model for answering factoid entity questions. *ACM Trans. Inf. Syst.* (2024). DOI: https://doi.org/10.1145/3678003

[9] Xinze Li, Hanbin Wang, Zhenghao Liu, Shi Yu, Shuo Wang, Yukun Yan, Yukai Fu, Yu Gu, and Ge Yu. 2024. Building a coding assistant via the retrieval-augmented language model. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 39, 1–25.

[10] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[11] Yongqi Li, Zhen Zhang, Wenjie Wang, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Distillation enhanced generative retrieval. In *Findings of the Association for Computational Linguistics (ACL '24)*. Association for Computational Linguistics.

[12] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. 2024. How can recommender systems benefit from large language models: A survey. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 28, 1–47.

[13] Xinyu Lin, Wenjie Wang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Bridging items and language: A transition paradigm for large language model-based recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. ACM.

[14] Sichun Luo, Bowei He, Haohan Zhao, Wei Shao, Yanlin Qi, Yinya Huang, Aojun Zhou, Yuxuan Yao, Zongpeng Li, Yuanzhang Xiao, et al. 2024. RecRanker: Instruction tuning large language model as ranker for top-k recommendation. *ACM Trans. Inf. Syst.* (2024). DOI: https://doi.org/10.1145/3705728

[15] Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. *ACM Trans. Inf. Syst.* (2024). DOI: https://doi.org/10.1145/3701228

[16] Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2024. FIT-RAG: Black-Box RAG with factual information and token reduction. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 40, 1–27.

[17] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 Technical Report. arXiV:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[18] Yingtao Peng, Chen Gao, Yu Zhang, Tangpeng Dan, Xiaoyi Du, Hengliang Luo, Yong Li, and Xiaofeng Meng. 2024. Denoising alignment with large language model for recommendation. *ACM Trans. Inf. Syst.* (2024). DOI: https://doi.org/10.1145/3696662

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[20] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. In *Proceedings of 37th Conference on Neural Information Processing Systems*.

[21] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. In *Proceedings of 37th Conference on Neural Information Processing Systems*.

[22] Hao Wang, Mingjia Yin, Luankang Zhang, Sirui Zhao, and Enhong Chen. 2024. MF-GSLAE: A multi-factor user representation pre-training framework for dual-target cross-domain recommendation. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 30, 1–28.

[23] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. ACM.

[24] Xin Wang, Hong Chen, Zirui Pan, Yuwei Zhou, Chaoyu Guan, Lifeng Sun, and Wenwu Zhu. 2024. Automated disentangled sequential recommendation with large language models. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 29, 1–29.

[25] Lixiang Xu, Yusheng Liu, Tong Xu, Enhong Chen, and Yuanyan Tang. 2024. Graph augmentation empowered contrastive learning for recommendation. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 34, 1–27.

[26] Dan Zhang, Shaojie Zheng, Yifan Zhu, Huihui Yuan, Jibing Gong, and Jie Tang. 2024. MCAP: Low-pass GNNs with Matrix completion for academic recommendations. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 33, 1–39.

[27] Xu Zhang, Zexu Lin, Xiaoyu Hu, Jianlei Wang, Wenpeng Lu, and Deyu Zhou. 2024. SECON: Maintaining semantic consistency in data augmentation for code search. *ACM Trans. Inf. Syst.* 43, 2 (2025), Article 36, 1–26.

[28] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language models as recommender systems: Evaluations and limitations. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*.