Navigating Large Language Models for Recommendation: From Architecture to Learning Paradigms and Deployment

Xinyu Lin*
xylin1028@gmail.com
National University of Singapore
Singapore

Keqin Bao
baokq@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Jizhi Zhang
cdzhangjizhi@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Yang Zhang
zyang1580@gmail.com
National University of Singapore
Singapore

Wenjie Wang
wenjiewang96@gmail.com
University of Science and Technology of China
Hefei, China

Fuli Feng
fulifeng93@gmail.com
University of Science and Technology of China
Hefei, China

Abstract

Large Language Models (LLMs) are reshaping the landscape of recommender systems, giving rise to the emerging field of LLM4Rec that attracts both academia and industry. Unlike earlier approaches that simply borrowed model architectures or learning paradigms from language models, recent advances have led to a dedicated and evolving technical stack for LLM4Rec, spanning architecture design, pre-training and post-training strategies, inference techniques, and real-world deployment. This tutorial offers a systematic and in-depth overview of LLM4Rec through the lens of this technical stack. We will examine how LLMs are being adapted to recommendation tasks across different stages, empowering them with capabilities such reasoning, planning, and in-context learning. Moreover, we will highlight practical challenges including complex user modeling, trustworthiness, and evaluation. Distilling insights from recent research and identifying open problems, this tutorial aims to equip participants with a comprehensive understanding of LLM4Rec and inspire continued innovation in this rapidly evolving field.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Large Language Models, Recommender Systems, Generative Models

ACM Reference Format:
Xinyu Lin*, Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, and Fuli Feng. 2025. Navigating Large Language Models for Recommendation: From Architecture to Learning Paradigms and Deployment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3731696

*Main contact author.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731696

1 Format, Intended Audience, and Previous Tutorial

• **Format.** This is a half-day (3 hours plus breaks) lecture-style tutorial, which is conducted on-site. There will be at least five presenters planning to physically attend the conference. The remaining individuals will be responsible for collecting and organizing the relevant materials

• **Intended audience.** This tutorial is intended for researchers and professionals from both academia and industry in the fields of information retrieval, recommender systems, natural language processing who are interested in understanding and utilizing LLMs for recommendation. Attendees are expected to have a basic understanding of recommender systems and machine learning. The tutorial will be accessible to those new to LLM4Rec, while also offering in-depth insights for experienced researchers aiming to explore the latest developments and future directions in this emerging area.

• **Previous edition.** We offer this tutorial at SIGIR-AP'23 [4], WWW'24 [50], and SIGIR'24 [6]. Nonetheless, previous tutorials primarily focusing early explorations of LLM4Rec through the lens of how general LLM abilities can be utilized for recommendation. In contrast, this tutorial is grounded in the latest wave of LLM4Rec research. We systematically collect and analyze recent work, identifying a structured technical stack uniquely tailored for LLM4Rec—spanning architecture, pre-training, post-training, inference, and deployment. Building on this stack, we not only organize prior efforts in a unified framework but also introduce the most recent advances, such as long Chain-of-Thought (CoT) preference reasoning and complex user behavior modeling. Through this tutorial, we aim to benefit researchers and industry professionals with a structured overview of current progress and future directions from a systematic technical perspective.

2 Presenters

Xinyu Lin[1] is a Ph.D. candidate at the University of Singapore, under the supervision of Prof. Tat-seng Chua. Her research interests lie in LLM-based recommendation, and her work has been published in top conferences and journals such as SIGIR and TOIS. She presents tutorials on the LLM-based recommendation at SIGIR'24.

[1]https://scholar.google.com/citations?user=0O_bs3UAAAAJ&hl=en.

Moreover, she has served as reviewer and PC member for the top conferences such as SIGIR and WWW.

Keqin Bao[2] is a Ph.D. student at University of Science and Technology of China (USTC), supervised by Prof. Fuli Feng and Prof. Xiangnan He. His research interest lies in the recommender system and LLMs. He has several publications in top conferences such as RecSys, EMNLP and WWW. He presents tutorials on the LLM-based recommendations in SIGIR-AP 23 and WWW 2024. He has served as the PC member and reviewer for the top conferences and journals including TOIS, RecSys and CIKM.

Jizhi Zhang[3] is a Ph.D. student at University of Science and Technology of China (USTC), supervised by Prof. Fuli Feng and Prof. Xiangnan He. His research interest lies in the recommender system and LLMs. He has several publications in top conferences such as SIGIR, WWW, RecSys, EMNLP, and ACL. He presents tutorials on the LLM-based recommendations in SIGIR 24 and WWW 2024.

Yang Zhang[4] is a research fellow at the National University of Singapore. He received his Ph.D. from the University of Science and Technology of China. His research focuses on LLM/agent personalization & recommendation. His first-author publications appear in top conferences/journals such as SIGIR and ACL, and he has received the Best Paper Honorable Mention at SIGIR 2021. He also serves as the reviewer for conferences/journals such as WWW, SIGIR, KDD, and TKDE.

Wenjie Wang[5] is a professor at the University of Science and Technology of China. He received Ph.D. in Computer Science from National University of Singapore in 2023. His research interests cover causal recommendation, data mining, and multimedia. He has published over 30 recommendation papers in top conferences and journals such as SIGIR, KDD, WWW, TOIS, and TIP. Moreover, he has served as the PC member and reviewer for the top conferences and journals including TPAMI, TOIS, SIGIR, WWW, and KDD. He has rich tutorial experience and presented tutorials in WWW 2022, SIGIR 2023, SIGIR-AP 2023, and WWW 2024.

Fuli Feng[6] is a professor at the University of Science and Technology of China. His research interests include information retrieval, data mining, causal inference, and multi-media processing. He has over 60 publications appeared in several top conferences such as SIGIR, WWW, and SIGKDD, and journals including TKDE and TOIS. He has received the Best Paper Honourable Mention of SIGIR 2021 and Best Poster Award of WWW 2018. Moreover, he has served as the reviewer for several top conferences and journals, including SIGIR, WWW, SIGKDD, NeurIPS, ICML, ICLR, ACL, TOIS, TKDE, TNNLS, TPAMI. He has rich teaching experience and has organized tutorials at SIGIR'23, WWW'21&22&23, and RecSys'21.

[2]https://data-science.ustc.edu.cn/_upload/tpl/14/26/5158/template5158/author/keqin-bao.htmll.
[3]https://data-science.ustc.edu.cn/_upload/tpl/14/26/5158/template5158/author/jizhi-zhang.html.
[4]https://scholar.google.com/citations?user=M9NcazMAAAAJ.
[5]https://scholar.google.com/citations?user=Ma5DtmoAAAAJ&hl=en.
[6]https://scholar.google.com.sg/citations?user=QePM4u8AAAAJ&hl=en.

3 Topic and relevance

3.1 Motivation

Recommender systems are a cornerstone of modern digital services, enabling personalized user experiences regarding products, services, and content [1]. Over the years, language models (LMs) have played a sustained and significant role in advancing recommender algorithms with model architecture and learning paradigms, laying the groundwork in steering LLMs for recommendation.

Regarding model architecture, sequential recommendation has drawn inspiration from the Transformer architecture, giving rise to innovative approaches like BERT4Rec [34] and SASRec [22]. Regarding learning paradigms, recommender models have also embraced the paradigm of pre-training and fine-tuning [25, 31]. Notably, earlier models such as P5 [17] and M6-Rec [12] have explored the potential of a unified LM to seamlessly handle various recommendation tasks. However, it is worth noting that these earlier studies depend on medium-size LMs with limited generalization capabilities, often necessitating extensive training data [17, 19].

Recently, the advent of powerful LLMs has significantly advanced the field of recommender systems, giving rise to a new technical stack tailored for leveraging LLMs in recommendation (LLM4Rec). Based on the technical stack, current research in LLM4Rec has focused on bridging this gap across multiple stages, including:

- *Model architecture.* Various item tokenizers are proposed to bridge recommendation space to the language space with semantics, diversity, and collaborative information [27, 39]. Besides, novel LLM architectures are specifically designed for recommendation [30, 45].
- *Pre-training*. LLM4Rec adopts a pre-training and post-training paradigm [11], where LLMs are trained on vast world knowledge and item corpus via Supervised Fine-Tuning (SFT) to understand item content in recommendation scenarios, potentially facilitating better understanding of complex user preference in the subsequent learning stage.
- *Post-training*. On top of pre-trained LLMs, the user behavior data is utilized to fine-tune LLMs to understand user heterogeneous behavior and preference alignment via different training strategies such as SFT [24] and Direct Preference Optimization (DPO) [2].
- *Inference*. For recommendation, different strategies are proposed to achieve additional objectives beyond accuracy, including in-context learning to incentivize the generalization ability [33]. Besides, various decoding strategies are proposed to accelerate the LLM inference speed [29] and alleviate the recommendation bias issue [7, 16].
- *Deployment*. Issues such as low deployment efficiency, trustworthiness and agent-based LLM4Rec systems have attracted increasing attention [8, 9, 36, 37, 47]. Existing progress has been made in boosting the deployment efficiency from different perspectives such as data [26] and model [43]. Besides, a series of work identifies unfairness and popularity bias issues in LLM4Rec [20, 41, 49].

Given these developments, this tutorial aims to provide a timely and comprehensive overview of LLM4Rec through the lens of this emerging technical stack. In addition to the progress of LLM4Rec, this tutorial will demonstrate and analyze its critical challenges such as complex user modeling and retraining efficiency. Last but not least, this tutorial will summarize key insights and outline several promising future research directions, including foundational open-domain models, personalized content generation [38], and test-time scaling. We hope this tutorial will enhance understanding and spark further exploration in this rapidly evolving field.

Necessity and timely of this tutorial. LLM4Rec has quickly become a vibrant and rapidly evolving research area. In recent years, a surge of studies has built up a dedicated technical stack to address unique challenges of LLM4Rec [3, 5, 20, 49, 51, 53]. Besides, some relevant workshops are hosted at CIKM'23[7] and WWW'24[8]. Furthermore, researchers also actively organize special issues on TOIS[9] and TORS[10]. Given the potential and rapid development of LLM4Rec, it is a suitable time to conduct this tutorial, benefiting researchers and industrial developers to learn the progress and future directions of LLM4Rec.

3.2 Objective

The aim of this tutorial is to offer a comprehensive grasp of LLM development for recommendation, encompassing the motivation, significance, current advancements, hurdles, and future directions. Aligned with the LLM4Rec technical stack—spanning architecture, learning paradigms, and deployments—it equips participants with practical insights to apply LLMs in recommendation and related domains such as personalized advertising.

3.3 Relevance

This tutorial is acutely relevant to the core themes of SIGIR, with a specific focus on user modeling and recommendation, poised to inspire advancements in other associated web applications. Previous version of this tutorial has been discussed in Section 1. We will also examine the broader, unique challenges that arise in the sphere of LLM4Rec. Additionally, three workshops covering analogous themes are scheduled for CIKM 2023[11], and another two for WWW 2024[12] and WWW 2025[13], indicating a surge of interest in this domain. The RecSys tutorial and CIKM/WWW workshops have all attracted considerable attention. Given SIGIR's strong alignment with RecSys, CIKM, and WWW, we expect this tutorial will attract substantial interest from the SIGIR community.

3.4 Outline

We present an outline of the topics to be covered with timing:

- Introduction. (15 Min, Fuli Feng)
  - Organization of the tutorial.
  - Background of recommender systems.
  - Architecture and learning paradigm of LMs.
  - LMs for recommendation [12, 17, 25, 34].
- Progresses of LLM4Rec (60 Min, Keqin Bao, Jizhi Zhang, Xinyu Lin)
  - Development of LLMs [56].
  - Overview of LLM4Rec Technical Stack.
  - Model Architecture.
    - Tokenizer for recommendation [27, 46].
    - RecLLM architecture [30, 45].
  - Pre-training of LLMs for knowledge understanding.
  - Post-training of LLM4Rec.
    - User behavior understanding via SFT, DPO, etc [3, 15].
    - User preference reasoning via SFT, RL, etc [14, 35].
  - Inference of LLM4Rec
    - In-context learning for recommendation [13, 21].
    - Decoding strategies for recommendation [7, 16].
  - Deployment of LLM4Rec
    - Trustworthiness of LLM4Rec [20, 36, 55].
    - Agent-related LLM4Rec [47, 48, 51].
    - Efficiency of LLM4Rec deployment [26, 29].
- Q&A. (5 Min)
- Break. (10 Min)
- Open Problems and Challenges of LLM4Rec. (60 Min, Yang Zhang, Wenjie Wang)
  - Heterogenous user behavior understanding
    - Tokenization for open-domain user behavior.
    - User understanding with CF knowledge [53, 54].
  - Lifelong user behavior understanding
    - RecLLM with lifelong memory.
    - Incremental learning of LLM4Rec.
  - Evaluation
    - Lack of new data for evaluation.
    - Lack of more comprehensive features for items and users.
    - Insufficient data diversity.
- Conclusion and future directions. (25 Min, Fuli Feng)
  - Conclusion.
  - Foundational open-domain recommender models.
  - Open-ended personalized content generation [44].
  - Test-time scaling law of LLM4Rec.
  - Recommender system for LLM-based agent platform [51].
- Q&A. (5 Min)

3.5 Qualification of presenters

We have been working on recommender systems for a long time with a series of publications [18, 52] that emerged in top-tier conferences and journals. Recently, we have also released several well-known papers about LLM4Rec [3, 5, 20, 28, 49], some of which are published at RecSys'23 [5, 49], SIGIR'24 [26, 33], and WWW'24 [20]. Besides, we organized a relevant workshop titled "Recommendation with Generative Models" at CIKM'2[14]/WWW'24[15] and a special issue on TOIS[16]. Moreover, our team has rich tutorial experience and has conducted more than 10 tutorials at various conferences including SIGIR, WWW, WSDM, CIKM, and RecSys [10, 23, 32, 40, 42]. We thus believe this tutorial would be attractive and insightful.

4 Tutorial Details

- **Tutorial materials**. The slides will be released on the tutorial website. Organizers can obtain copyright permission.

[7]https://uobevents.eventsair.com/cikm2023/workshops.
[8]https://www2024.thewebconf.org/program/workshops/
[9]https://dl.acm.org/journal/tois/calls-for-papers.
[10]https://dl.acm.org/journal/tors/calls-for-papers.
[11]https://uobevents.eventsair.com/cikm2023/workshops
[12]https://www2024.thewebconf.org/program/workshops/
[13]https://www2025.thewebconf.org/full-schedule
[14]https://rgm-cikm23.github.io/.
[15]https://generative-rec.github.io/workshop/
[16]https://dl.acm.org/journal/tois/calls-for-papers.

- **Organization details**. Additionally, we can prepare pre-recorded lectures if deemed necessary. Moreover, we are open to live-streaming the tutorial via popular video-streaming platforms.

References

[1] Qingyao Ai et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* 4 (2023), 80–90.

[2] Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. 2024. Aligning Large Language Model with Direct Multi-Preference Optimization for Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 76–86.

[3] Keqin Bao et al. 2023. A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems. *CoRR* abs/2308.08434 (2023). arXiv:2308.08434

[4] Keqin Bao et al. 2023. Large Language Models for Recommendation: Progresses and Future Directions. *SIGIR-AP* (2023).

[5] Keqin Bao et al. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *RecSys*. ACM, 1007–1014.

[6] Keqin Bao, Jizhi Zhang, Xinyu Lin, Yang Zhang, Wenjie Wang, and Fuli Feng. 2024. Large language models for recommendation: Past, present, and future. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2993–2996.

[7] Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. 2024. Decoding Matters: Addressing Amplification Bias and Homogeneity Issue in Recommendations for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10540–10552.

[8] Lukas Berglund et al. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv:2309.12288.*

[9] Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Agentic Feedback Loop Modeling Improves Recommendation and User Simulation. *SIGIR* (2025).

[10] Jiawei Chen et al. 2021. Bias Issues and Solutions in Recommender System. In *RecSys*. 825–827.

[11] Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, et al. 2023. Leveraging large language models for pre-trained recommender systems. *arXiv preprint arXiv:2308.10837.*

[12] Zeyu Cui et al. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084* (2022).

[13] Sunhao Dai et al. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *RecSys*. ACM, 1126–1132.

[14] Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Large Language Models for Recommendation with Deliberative User Preference Alignment. *arXiv preprint arXiv:2502.02061.*

[15] Yue Feng et al. 2023. A Large Language Model Enhanced Conversational Recommender System. *arXiv preprint arXiv:2308.06212* (2023).

[16] Chongming Gao, Mengyao Gao, Chenxiao Fan, Shuai Yuan, Wentao Shi, and Xiangnan He. 2025. Process-Supervised LLM Recommenders via Flow-guided Tuning. *arXiv preprint arXiv:2503.07377* (2025).

[17] Shijie Geng et al. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *RecSys*. 299–315.

[18] Xiangnan He et al. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.

[19] Yupeng Hou et al. 2022. Towards universal sequence representation learning for recommender systems. In *SIGKDD*. 585–593.

[20] Meng Jiang et al. 2024. Item-side Fairness of Large Language Model-based Recommendation System. In *WWW*.

[21] Wang-Cheng Kang et al. 2023. Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction. *CoRR* abs/2305.06474 (2023).

[22] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*. IEEE, 197–206.

[23] Wenqiang Lei et al. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *SIGIR*. 2425–2428.

[24] Yuxuan Lei, Jianxun Lian, Jing Yao, Xu Huang, Defu Lian, and Xing Xie. 2024. Recexplainer: Aligning large language models for explaining recommendation models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1530–1541.

[25] Jiacheng Li et al. 2023. Text Is All You Need: Learning Language Representations for Sequential Recommendation. In *SIGKDD*. ACM, 1258–1267.

[26] Xinyu Lin et al. 2024. Data-efficient Fine-tuning for LLM-based Recommendation. *arXiv preprint arXiv:2401.17197* (2024).

[27] Xinyu Lin, Haihan Shi, Wenjie Wang, Fuli Feng, Qifan Wang, See-Kiong Ng, and Tat-Seng Chua. 2025. Order-agnostic Identifier for Large Language Model-based Generative Recommendation. *arXiv preprint arXiv:2502.10833.*

[28] 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*. 1816–1826.

[29] Xinyu Lin, Chaoqun Yang, Wenjie Wang, Yongqi Li, Cunxiao Du, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025. Efficient Inference for Large Language Model-based Generative Recommendation. In *The Thirteenth International Conference on Learning Representations.*

[30] Kai Mei and Yongfeng Zhang. 2023. LightLM: a lightweight deep and narrow language model for generative recommendation. *arXiv preprint arXiv:2310.17488.*

[31] Zhaopeng Qiu et al. 2021. U-BERT: Pre-training User Representations for Improved Recommendation. In *AAAI*. AAAI Press, 4320–4327.

[32] Zhaochun Ren et al. 2018. Information Discovery in E-commerce: Half-day SIGIR 2018 Tutorial. In *SIGIR*. 1379–1382.

[33] Wentao Shi, Xiangnan He, Yang Zhang, Chongming Gao, Xinyue Li, Jizhi Zhang, Qifan Wang, and Fuli Feng. 2024. Large language models are learnable planners for long-term recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1893–1903.

[34] Fei Sun et al. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*. 1441–1450.

[35] Alicia Y Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H Chi, and Xinyang Yi. 2024. Leveraging LLM Reasoning Enhances Personalized Recommender Systems. *arXiv preprint arXiv:2408.00802.*

[36] Jindong Wang et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095* (2023).

[37] Lei Wang et al. 2023. RecAgent: A Novel Simulation Paradigm for Recommender Systems. *arXiv preprint arXiv:2306.02552* (2023).

[38] Wenjie Wang et al. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516* (2023).

[39] 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*. 2400–2409.

[40] Xiang Wang et al. 2020. Learning and reasoning on graph for recommendation. In *WSDM*. 890–893.

[41] Chen Xu et al. 2023. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *arXiv preprint arXiv:2311.07054* (2023).

[42] Jun Xu et al. 2018. Deep learning for matching in search and recommendation. In *SIGIR*. 1365–1368.

[43] Wujiang Xu, Qitian Wu, Zujie Liang, Jiaojiao Han, Xuying Ning, Yunxiao Shi, Wenfang Lin, and Yongfeng Zhang. 2025. SLMRec: Distilling Large Language Models into Small for Sequential Recommendation. In *The Thirteenth International Conference on Learning Representations.*

[44] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized Generation In Large Model Era: A Survey. *arXiv preprint arXiv:2503.02614.*

[45] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152.*

[46] Jianyang Zhai, Zi-Feng Mai, Chang-Dong Wang, Feidiao Yang, Xiawu Zheng, Hui Li, and Yonghong Tian. 2025. Multimodal Quantitative Language for Generative Recommendation. In *The Thirteenth International Conference on Learning Representations.*

[47] An Zhang et al. 2023. On Generative Agents in Recommendation. *arXiv preprint arXiv:2310.10108* (2023).

[48] Junjie Zhang et al. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. *arXiv preprint arXiv:2310.09233* (2023).

[49] Jizhi Zhang et al. 2023. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *RecSys*. ACM, 993–999.

[50] Jizhi Zhang et al. 2024. Large Language Models for Recommendation: Progresses and Future Directions. *WWW* (2024).

[51] Jizhi Zhang, Keqin Bao, Wenjie Wang, Yang Zhang, Wentao Shi, Wanhong Xu, Fuli Feng, and Tat-Seng Chua. 2024. Prospect Personalized Recommendation on Large Language Model-based Agent Platform. *arXiv preprint arXiv:2402.18240* (2024).

[52] Yang Zhang et al. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *SIGIR*. 11–20.

[53] Yang Zhang et al. 2023. Collm: Integrating collaborative embeddings into large language models for recommendation. *arXiv preprint arXiv:2310.19488* (2023).

[54] Yang Zhang, Keqin Bao, Ming Yan, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024. Text-like Encoding of Collaborative Information in Large Language Models for Recommendation. In *ACL*. 9181–9191.

[55] Jujia Zhao et al. 2024. LLM-based Federated Recommendation. https://api.semanticscholar.org/CorpusID:267682268

[56] Wayne Xin Zhao et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).