Large Language Models for Recommendation: Past, Present, and Future

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

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

Xinyu Lin
xylin1028@gmail.com
National University of Singapore

Yang Zhang
zyang1580@gmail.com
University of Science and Technology of China

Wenjie Wang
wenjiewang96@gmail.com
National University of Singapore

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

ABSTRACT
Large language models (LLMs) have significantly influenced recommender systems, spurring interest across academia and industry in leveraging LLMs for recommendation tasks. This includes using LLMs for generative item retrieval and ranking, and developing versatile LLMs for various recommendation tasks, potentially leading to a paradigm shift in the field of recommender systems. This tutorial aims to demystify the Large Language Model for Recommendation (LLM4Rec) by reviewing its evolution and delving into cutting-edge research. We will explore how LLMs enhance recommender systems in terms of architecture, learning paradigms, and functionalities such as conversational abilities, generalization, planning, and content generation. The tutorial will shed light on the challenges and open problems in this burgeoning field, including trustworthiness, efficiency, online training, and evaluation of LLM4Rec. We will conclude by summarizing key learnings from existing studies and outlining potential avenues for future research, with the goal of equipping the audience with a comprehensive understanding of LLM4Rec and inspiring further exploration in this transformative domain.

CCS CONCEPTS
• Information systems → Recommender systems.

KEYWORDS
Large Language Models, Recommender Systems, Generative Recommendation, Generative Models

ACM Reference Format:

1 FORMAT AND INTENDED AUDIENCE
• 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. Large language models (LLMs) have significantly influenced recommender systems, spurring interest across academia and industry in leveraging LLMs for recommendation tasks. This encompasses initiatives such as deploying LLMs for generative item retrieval and ranking, as well as exploring the potential to develop universal LLMs capable of addressing diverse recommendation tasks, which may herald a paradigm shift in the field of recommender systems. This tutorial aims to demystify LLM4Rec by reviewing its evolution and delving into cutting-edge research. We will explore how LLMs enhance recommender systems in terms of architecture, learning paradigms, and functionalities such as conversational abilities, generalization, planning, and content generation. The tutorial will shed light on the challenges and open problems in this burgeoning field, including trustworthiness, efficiency, online training, and evaluation of LLM4Rec. We will conclude by summarizing key learnings from existing studies and outlining potential avenues for future research, with the goal of equipping the audience with a comprehensive understanding of LLM4Rec and inspiring further exploration in this transformative domain.

2 PRESENTERS
Keqin Bao1 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.

*Main contact author.

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 components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '24, July 14–18, 2024, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0431-4/24/07...$15.00
https://doi.org/10.1145/3626772.3661383

Jizhi Zhang 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-AP 23 and WWW 2024.

Xinyu Lin is a Ph.D. candidate at the University of Singapore, under the supervision of Prof. Tat-seng Chua. Her research interests lie in recommender systems, and her work has been published in top conferences and journals such as SIGIR, WWW, CIKM, and TOIS. Moreover, she has served as the reviewer and PC member for the top conferences and journals, such as SIGIR and TOIS.

Yang Zhang is a final-year Ph.D. student at the University of Science and Technology of China. His research interest lies in the recommender system. He has over ten publications that appeared in top conferences/journals. His work on recommendation has received the Best Paper Honorable Mention in SIGIR 2021. He has served as the PC member and reviewer for the top conferences and journals including TKDE, TOIS, RecSys, WSDM, WWW, SIGIR, and KDD. He has presented tutorials at WWW 2022, SIGIR 2023, SIGIR-AP 2023, and WWW 2024.

Wenjie Wang is a research fellow at the School of Computing, National University of Singapore. 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, ACM MM, 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 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, TOIS, TKDE, TNNLS, TPAMI. He has rich teaching experience and has organized tutorials at SIGIR'23, WWW'21&22&23, and RecSys'21.

3 TOPIC AND RELEVANCE

3.1 Motivation

Recommender systems have become an integral part of our digital lives, providing personalized recommendations regarding products, services, and content [1]. Looking back at the history of recommender systems, language models (LMs) are essential in shaping the landscape of recommender algorithms.

In terms of model architecture, sequential recommendation has drawn inspiration from the Transformer architecture, giving rise to innovative approaches like BERT4Rec [31] and SASRec [21]. Regarding learning paradigms, recommender models have also embraced the two-stage paradigm of pre-training and fine-tuning [23, 29], as well as prompt learning for multiple tasks [9, 14]. Notably, earlier models such as P5 [14] and M6-Rec [9] have explored the potential of a unified LM to seamlessly handle various recommendation tasks, such as click-through rate prediction, sequential recommendation, explanation generation, and conversational recommendation. However, it is worth noting that these earlier studies depend on medium-size LMs with limited generalization capabilities, often necessitating extensive recommendation data for training [14, 16].

Recently, the advent of powerful LLMs has significantly sparked technical advances in the field of recommender systems, opening up new horizons of leveraging LLMs for recommendation (shorted as LLM4Rec). These LLMs exhibit remarkable language understanding and generation capabilities, generalization prowess, and planning abilities [61]. In light of these, in-context learning-based [10] and instruction tuning-based recommendations [4, 53] with few-shot data, have garnered substantial attention within the research community. As such, it is really opportune to organize this tutorial to examine the progress of LLM4Rec.

In addition to the progress of LLM4Rec, this tutorial will demonstrate and analyze its critical challenges. The trustworthiness issues in recommendations generated by LLMs remain a primary concern, including problems like bias & fairness, privacy, and safety. Besides, efficiency is another essential aspect. LLMs are computationally demanding, thus deploying LLMs into real-time recommendation systems requires careful optimization and resource management. Additionally, the dynamic nature of user preferences necessitates timely model retraining to keep recommendations up to date. It is also challenging to achieve timely retraining in LLM4Rec. Lastly, recommendation data might encompass collaborative filtering signals, long interaction sequences from users, and high-dimensional features, which are also hard for LLMs to capture or utilize.

Last but not least, this tutorial will summarize the implications of previous work and present future research directions. In the future, it is essential to construct the evaluation benchmarks for LLM4Rec. Moreover, personalized AI-generated content might significantly facilitate the boom of generative recommendations [35]. As we delve into this tutorial, we aim to empower researchers and practitioners with knowledge and inspiration about LLM4Rec, inspiring further exploration and innovation in this emerging field.

Necessity and timely of this tutorial. LLM4Rec is a new and hot topic, but developing vigorously. In the past years, there emerge extensive relevant papers [2, 4, 19, 52, 55, 59]. Besides, some relevant workshops are hosted at CIKM’23 and WWW’24. Furthermore, researchers also actively organize special issues on TOIS and TORS. It is undoubted that more research endeavors will surface in this field, leading to more reliable and powerful recommender systems. Given the potential and rapid development of LLM4Rec, it is a suitable time to conduct this tutorial, benefiting

---

2. https://scholar.google.com/citations?user=O_2xUAAAAJ&hl=en
5. https://scholar.google.com/scholar?user=QePMtskAAAJ&hl=en
6. https://www2024.thewebconf.org/program/workshops/
7. https://dl.acm.org/journal/tois/calls-for-papers
8. https://dl.acm.org/journal/tors/calls-for-papers
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. Additionally, this tutorial strives to empower participants with the expertise and capabilities needed to harness LLMs for their own projects and make contributions to the progress of related domains, such as 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. In the lead-up to its proposal for SIGIR’24, several related tutorials have emerged. Among them, one was presented at RecSys’23 [18], which primarily illuminated the technical dimensions and predominantly featured research on medium-sized language models. Our tutorial, however, concentrates more intently on the application of LLMs to recommendation systems, exploring the foundational reasons and benefits for developing LLMs in this context. We will also examine the broader, unique challenges that arise in the sphere of LLM4Rec. Additionally, two workshops covering analogous themes are scheduled for CIKM 2023\textsuperscript{13}, and another for WWW 2024\textsuperscript{12}, indicating a surge of interest in this domain. The RecSys tutorial and CIKM/WWW workshops have all attracted considerable attention. Given SIGIR’s strong ties with RecSys, CIKM, and WWW, we expect that our tutorial will similarly captivate the SIGIR community, generating significant interest and engagement.

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.
  - The development of LMs.
  - LMs for recommendation.
    - Utilizing LMs’ model structure for recommendation [23, 31].
    - LM as item encoder [16, 47].
    - Recommendation as natural language processing [9, 14].
  - Progresses of LLM4Rec (60 Min, Keqin Bao, Jizhi Zhang, Xinyu Lin)
    - Development of LLMs [61].
    - Benefits of LLMs for recommendations [24].
    - In-context learning-based recommendation.
      - Directly utilizing in-context learning for recommendation [10, 20, 28].
      - Combining in-context learning with conventional recommender models [17, 34, 41, 42].
    - Tuning-based LLM4Rec.
      - Training with raw textual description interactions [2, 4, 53].
- Training with multi-source information integration [13, 26, 46, 59].
- Agent-related LLM4Rec.
- LLM-based Agent as environment simulators [33, 49].
- LLM-based Agent as recommenders [12, 38, 40, 51, 55].
- Trustworthy LLM4Rec.
  - Privacy in LLM4Rec [7, 60].
  - OOD & robustness in LLM4Rec [5, 32].
  - Fairness & Bias in LLM4Rec [19, 39, 43, 52].
  - Explainable LLM4Rec [6, 57].
  - Safety in LLM4Rec [56].
- Q&A. (5 Min)
- Break. (10 Min)
- Open Problems and Challenges of LLM4Rec. (60 Min, Yang Zhang, Wenjie Wang)
  - Modeling
    - Lifelong user modeling in LLM4Rec [25].
    - Item/User representation.
  - Cost
    - Training/Inference/Deployment of LLM4Rec [11, 45].
  - Evaluation
    - Lack of new data for evaluation.
    - Lack of more comprehensive features for items and users.
    - Insufficient data diversity.
    - Lack of the interactive recommendation benchmark [37].

Conclusion and future directions. (25 Min, Fuli Feng)
  - Conclusion.
  - Generative recommendation with LLMs [35].
  - Recommender system for LLM-based agent platform [55].
  - Foundational recommender models [48, 50]

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 [15, 58] that emerged in top-tier conferences and journals. Recently, we have also released several well-known papers about LLM4Rec [2, 4, 19, 52], some of which are published at RecSys’23 [4, 52] and WWW’24 [19]. Besides, we organized a relevant workshop titled “Recommendation with Generative Models” at CIKM’23\textsuperscript{13}/WWW’24\textsuperscript{14} and a special issue on TOIS\textsuperscript{15}. We are familiar with this field and will provide a systematic review and prospects of LLM4Rec. Moreover, our team has rich tutorial experience and has conducted more than 10 tutorials at various conferences including SIGIR, WWW, WSDM, CIKM, and RecSys [8, 22, 30, 36, 44]. 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.

\textsuperscript{13}https://rgm-cikm23.github.io/
\textsuperscript{14}https://generative-rec.github.io/workshop/
\textsuperscript{15}https://dl.acm.org/journal/tois/calls-for-papers.
Previous editions. Before this tutorial for SIGIR’24, we offer this tutorial at SIGIR-AP’23 [3] and WWW’24 [54]. We intend to gather valuable audience feedback to refine and enhance the tutorial content specifically for the SIGIR community. Additionally, this domain is advancing swiftly, with innovative research surfacing on a monthly basis. At SIGIR’24, we will present the latest developments, ongoing challenges, and unresolved issues within LLM4Rec, informed by the most recent studies from 2024.

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

REFERENCES
