A Survey on Large Language Models for Recommendation Systems

Junlei Yang *

Department of journalism and communication, Nanjing Normal University, Nanjing, 210023, China

* Corresponding Author Email: yangjunlei0722@163.com

Abstract. Large language models (LLMs) offer new approaches to overcoming the limitations of recommendation systems in representation, user modeling, and decision-making. Accordingly, recommendation systems are evolving from paradigms primarily based on behavioral statistics and discriminative learning toward new paradigms featuring semantic understanding, reasoning-based decision-making, and autonomous interaction. This paper systematically reviews recent advances in LLM-empowered recommendation systems. The principal advantages of LLMs in semantic modeling, knowledge transfer, and generative reasoning are analyzed. Then, the representative technical paradigms including representation-enhanced recommendation, user-interest and context enhancement, prompt learning and task reformulation, and generative recommendation are subsequently reviewed and summarized. Finally, the paper analyzes major challenges related to hallucination, inference efficiency, privacy and security, fairness and trustworthiness, and recommendation agents, and outlines future research directions.

Keywords: Large language models, recommendation systems, generative recommendation, prompt learning, recommendation agents.

1. Introduction

Recommendation systems address information overload by matching users with relevant items in large information spaces. They usually rely on collaborative filtering or deep learning and operate through data collection, candidate retrieval, and ranking. Today, e-commerce platforms, streaming services, and social networks all use recommendation systems extensively. However, conventional systems face three main limitations. First, ID-based representations contain little semantic information, which worsens cold-start problems. Second, these systems often fail to understand complex behaviour sequences, distinguish between long-term preferences and short-term intentions, or recognise fine-grained and long-tail interests. Third, these decisions are often opaque and thus reduce user trust. The reasons for these deficiencies include a dependence on historical interaction data, vector similarity and supervised discriminative learning.

LLMs provide new functions for recommendations. Pre-trained on large amounts of text data, they have learned the rules of language and general knowledge. LLMs have captured deep semantic relationships and utilized context more effectively than traditional models. They can also make rational choices in a zero-shot or few-shot setting by using the knowledge and reasoning abilities learned in pre-training [1]. The above strengths can make a recommendation system more intelligent and less behaviourally driven.

Research on LLM-based recommendations is growing quickly, but existing work spans several parallel directions and lacks unified organization. Therefore, this paper will introduce a general system for the field. We divide the main ways into two categories: how many LLMs are integrated and how much autonomy the system has. We also cover model adaptation, deployment, evaluation and domain application, and finally introduce open problems and future directions.

2. Foundations of Recommendation Systems and Large Language Models

As shown in Figure 1, the four main types of LLMs address the shortcomings of traditional recommendation systems in semantic representation, user modelling and decision-making. Together,

these paradigms increase the accuracy, personalisation, interpretability, interaction and generation capabilities. LLMs enhance traditional recommendation systems in three ways. First, ID-based representations lack semantics and have problems with cold-start items. LLMs can encode titles, descriptions, images, knowledge graphs and other rich information to address this problem. Second, the traditional model cannot distinguish between stable preferences and short-term wishes, and it also fails to record rare or subtle interests. LLMs can use long contexts and reasoning to infer latent needs from sparse behaviour and rephrase recommendation tasks as language tasks. Third, the traditional model is not transparent. LLMs can generate recommendation lists and personalized explanations, while retrieval augmentation, counterfactual reasoning, and evidence grounding can make these explanations more reliable.

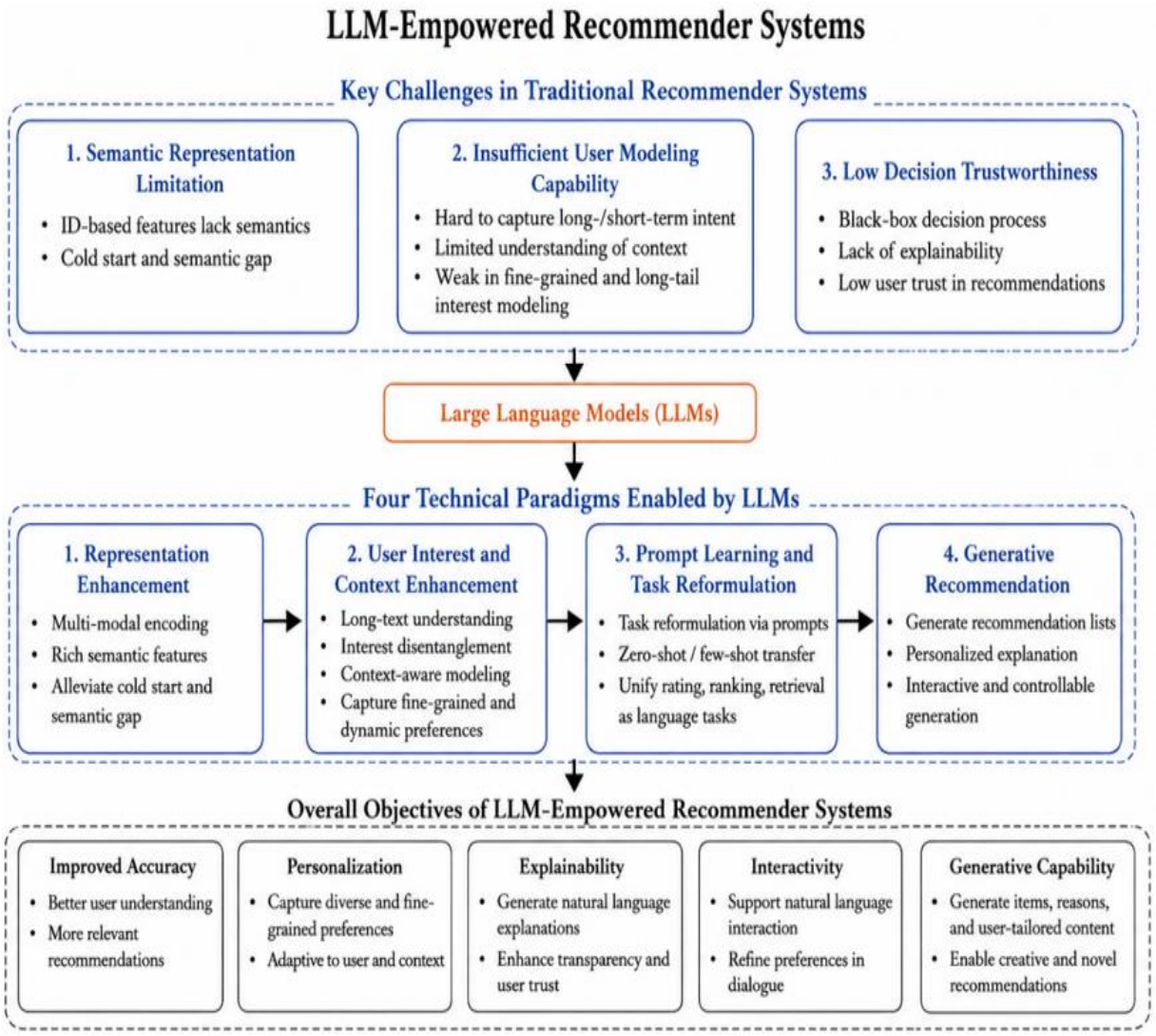


Fig 1. Overview of key challenges, technical paradigms, and objectives of LLM-enpowered recommendation system.

Industrial deployment remains challenging. LLMs can be slow at inference, may hallucinate, and may expose private data or amplify bias. Effective systems therefore need human oversight and strong ethical safeguards in addition to LLM reasoning and generation. The following sections review four core paradigms: representation-enhanced recommendation, user-interest and context enhancement, prompt learning and task reformulation, and generative recommendation.

3. Major Paradigms and Key Technologies for LLM-Empowered Recommendation

Table 1. Comparison of LLM-based recommendation technologies.

Method Category	Main Approach	Representative Work	Principal Advantages and Effects
Representation-enhanced recommendation	Construct unified semantic representations with LLMs and fuse multimodal information such as text and images	Modal Recommendation (BERT + Swin + SASRec) [2]; P4R [3]; NoteLLM-2 [4]	Enriches item semantics, improves the matching of user interests to item features, alleviates cold-start and long-tail problems, and improves accuracy and generalization.
User-interest and context enhancement	Dynamically model user preferences through context awareness, long- and short-term interest modeling, and reasoning	LSCAR [5]; LLM-based user-interest modeling [6]; CoT-based user-intent modeling [7]	More accurately captures the evolution of user interests, improves sequential, conversational, and complex interactive recommendation, and strengthens the semantic richness and interpretability of user profiles.
Prompt learning and task reformulation	Reformulate recommendation as natural-language tasks and transfer pretrained knowledge through prompts and parameter-efficient fine-tuning	Prompt4NR [8]; RAG-based Recommendation [7];Adapter-based Transfer Learning [23]; PPR [24]	Activates the knowledge acquired during LLM pre-training, improves task transferability, enhances factual consistency through external knowledge retrieval, and achieves better performance in zero-shot, few-shot, and cold-start recommendation scenarios with lower training costs.
Generative recommendation	Model recommendation as generation and combine retrieval-augmented generation, agents, and trustworthy generation for autonomous recommendation	GenRec [8]; Agent4Rec, AgentCF, and RecMind [10]–[12]; CEF [13]	Enables a paradigm shift from preference prediction to recommendation generation, strengthens complex requirement understanding, multi-turn interaction, and autonomous planning capabilities, and improves personalization, interpretability, and reliability in open-domain recommendation scenarios.

Table 1 compares the four paradigms. LLM-based recommendation has evolved from shallow to deep integration. Representation enhancement enriches user and item representations and eases semantic sparsity and cold-start problems. Dynamic user modeling then tracks changes in user interests. Prompt learning maps conventional tasks to language understanding and generation, allowing systems to reuse pretrained knowledge. Generative recommendation goes further by treating recommendation as an end-to-end generation task and combining retrieval, agents, and trustworthy generation. This progression moves recommendation systems from prediction toward planning, interaction, and decision-making.

3.1. Representation-Enhanced Recommendation

Representation-enhanced recommendation uses semantic encoders to extract features from item titles, descriptions, images, and other content. It maps these features into a shared vector space and combines them with user behavior and collaborative signals for matching. Traditional methods rely heavily on user and item IDs learned from past interactions and therefore have poor performance when dealing with new items, long-tail items and cross-domain tasks. Existing multimodal methods may not use text and images fully or properly balance their contributions. LLMs solve the above

problems by encoding the item content more deeply or generating structured item profiles. They reduce reliance on IDs and sparse interaction data while enriching item representations.

To address semantic sparsity in ID-based recommendation, ref. [2] jointly fine-tunes BERT and a Swin Transformer, then feeds their multimodal representations into SASRec for sequential recommendation. The method matches or exceeds ID-based models on ordinary and popular items. To make better use of textual attributes in graph collaborative filtering without directly deploying an LLM at high cost, ref. [3] proposes P4R. P4R prompts an LLM to create item profiles, encodes them with a pretrained language model, and fuses the results with collaborative signals through graph convolution and contrastive learning. It improves Recall, NDCG, and MRR at relatively low cost. To address weak visual modeling and modality imbalance. Ref. [4] proposes NoteLLM-2. It models text and images separately and uses visually gated late fusion to control their contributions. The method improves multimodal recommendation performance and raises click-through and engagement in online evaluation.

3.2. User-Interest and Context Enhancement

User-interest and context enhancement combine long-term behaviour, recent interactions and context such as time and place, and the current task to model changing preferences. Traditionally, only static profiles of the rating matrix or behaviour sequence are used. They cannot differentiate between stable preferences and short-term explorations and thus fail to infer hidden motives from dialogue and logs. LLMs can handle clicks, queries, reviews and multi-turn conversations simultaneously. Through long-context understanding and reasoning, they extract fine-grained interests and explain changes in preference.

Ref. [5] proposes LSCAR to model long- and short-term interests simultaneously. It has a user-context interaction module, an autoencoding multi-interest module and an ensemble fusion-gate module. LSTM, multihead self-attention and gated fusion are used to track changes in interest more accurately for sequential recommendation. Given the new demands that traditional labels cannot meet, reference [6] has been used to analyze multi-turn dialogues and dynamically update user interests. Better understanding of the demand and conversational recommendation. To uncover motives that cannot be revealed through historical choices alone, ref. [7] uses chain-of-thought prompting to guide an LLM in preference formation and converts interactions into structured user profiles. Thus, we can show the hidden interests more effectively and improve both the accuracy and interpretability of the complex preference model.

3.3. Prompt Learning and Task Reformulation

Prompt learning and task reformulation use natural-language or continuous prompts to express click prediction, rating prediction, ranking, and explanation as language tasks. The system adapts by updating prompts or a small set of auxiliary modules. Conventional recommendations often need separate models and full training for different tasks. Full fine-tuning of pretrained models also requires many parameters and can transfer poorly in zero-shot and few-shot settings. LLMs can instead use task instructions and pretrained knowledge. Personalized prompts, parameter-efficient tuning, and retrieved external information make this adaptation more economical.

For news recommendation, ref. [8] proposes Prompt4NR, which reformulates click prediction as a cloze-style task. It uses discrete, continuous, and hybrid prompts to model the relation between user interests and news content, outperforming conventional fine-tuning. To reduce nonexistent-item generation and factual errors, ref. [9] applies retrieval-augmented generation. It retrieves evidence from item repositories, knowledge bases, and user histories and gives that evidence to the LLM as context, improving factual consistency and reliability. To reduce the cost of full fine-tuning, ref. [10] inserts lightweight adapters into pretrained language models and visual encoders. Updating only about 2%–3% of parameters support recommendation transfer and can match, or sometimes exceed, full fine-tuning. For few-shot, zero-shot, and cold-start sequential recommendation, ref. [11] proposes Personalized Prompts for Sequential Recommendation (PPR). PPR derives soft prompts from user

profiles and combines them with prompt-guided contrastive learning. It outperforms full fine-tuning and classical sequential methods in several cold-start settings, while its prompt-only version remains competitive.

3.4. Generative Recommendation

Generative recommendation treats recommendation as an autoregressive generation task. Given user history and current needs, the model directly generates item text, semantic IDs, recommendation lists, or personalized explanations. Conventional systems use multistage retrieval and ranking pipelines whose components are usually optimized separately. As a result, they struggle to unify need understanding, result generation, and natural-language interaction. Generative models can also invent items, produce incorrect explanations, and incur high inference costs. LLMs can integrate understanding, reasoning, and generation, while memory, planning, tool use, and counterfactual reasoning can strengthen interaction and decision-making.

To simplify retrieval and ranking pipelines, ref. [12] proposes GenRec. It represents items with textual identifiers or semantic IDs and autoregressively generates target items from user behavior. GenRec learns the joint distribution of users and items, simplifies the pipeline, and achieves competitive results. To support autonomous planning and sustained interaction, ref. [13][14][15] introduce long-term memory, task planning, multi-agent collaboration, and tool use. Agent4Rec uses user agents to simulate behavior for policy evaluation; AgentCF lets user and item agents collaborate through natural-language interaction; and RecMind combines long-term memory with self-reflection to plan complex tasks. These systems improve recommendation quality, interaction, and decision efficiency in complex settings. To make rationales better reflect decision evidence and to assess fairness, ref. [16] proposes CEF. It uses counterfactual reasoning to identify decisive factors and evaluate fairness, while faithfulness checks, knowledge graphs, and evidence citations improve agreement between explanations and decisions.

4. Challenges and Future Directions

4.1. Accuracy, Hallucination, and Factual Consistency.

Hallucination is a serious problem of reliability for recommendation systems using LLMs. In the absence of strong connections with a real item catalogue, the model may generate reasonable but fake items and attribute/bibliographic information. Its explanations may not be the same as those provided by the retrieval and ranking models. Therefore, researchers have added knowledge augmentation and hard constraints to solve this problem. Retrieval-augmented generation can provide high-confidence knowledge bases and structured knowledge graphs as factual anchors. Trie-based constrained decoding can limit the output to valid candidate identifiers. Fact-checking and explicit evidence attribution after generation can help align the generated text with the recommendation behaviour.

4.2. Efficiency, Cost, and Industrial Deployment.

Industrial Recommendations need to handle high concurrency, thousands of requests per second, and millisecond-level response times. Large parameter counts and token-by-token decoding of LLMs are also issues. A multi-stage funnel needs to retrieve and rank many candidates at once, but autoregressive generation is serial and has high latency. Item catalogs and behaviour streams are also changing at a high speed, so if the parameters are not updated promptly, LLMs will be unable to respond to new trends and immediate negative feedback effectively. The System will accelerate development and run independently at different levels. An LLM can be used to build deep user representations in a nearline or offline layer, and a lightweight online model can use these dense vectors for fast retrieval and ranking. Dynamic Request Routing, Item-side Semantic Caches, Quantization and Structured Pruning reduce costs. Small on-device models can collect real-time signals, such as swipes and dwell time, and the cloud model will update the global knowledge asynchronously. This Design will be relatively fresh, private and convenient.

4.3. Privacy, Security, Fairness, and Controllability.

Comprehensive personalisation needs considerable amounts of user data and is therefore more susceptible to privacy violations, manipulation or discrimination. LLMs may have memorised private communication data, and hidden prompts added to product descriptions or reviews aim to change recommendations. LLMs may also inherit popularity bias from historical data, reducing exposure for long-tail and cold-start items. The above strong reasons may reinforce filter bubbles and feedback loops. Systems need multi-level technical defences and ecosystem-wide alignment. Federated Learning and Differential Privacy can support joint modeling without exposing raw data. Causal inference and counterfactual reasoning can mitigate popularity confounding, and fairness rewards and debiasing penalties can enhance long-tail exposure. Independent Safety Guardrails and real-time semantic review can filter out unsafe explanations and prevent adversarial manipulation.

5. Conclusion

Shortcomings of traditional recommendation systems in semantic representation, user modelling and recommendation decisions are discussed in this paper. Four LLM-based paradigms were reviewed: representation enhancement, user-interest and context enhancement, prompt learning and task reformulation, and generative recommendation. It also introduced their applications and the problems of inference efficiency, hallucination, factual inconsistency, privacy, security and scalability. LLMs will probably be added to support the traditional recommendation system in practice. Retrieval and Ranking Models can still provide real-time performance and computational efficiency; LLMs add semantic understanding, reasoning, multi-turn interaction and autonomous decision-making. Together, they can predict the user's needs and changes in that need to offer proactive plans and complete tasks. Retrieval augmentation, parameter-efficient tuning, multimodal modelling, long-term memory and agent technologies have all advanced; unified recommendation foundation models and recommendation agents may constitute the next generation of recommendation systems.

Reference

[1] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., & Chen, E. (2024). A survey on large language models for recommendation. World Wide Web, 27(5), 1–31.

[2] Yuan, Z., Yuan, F., Song, Y., Li, Y., Yang, F., & Pan, Y. (2023). Where to go next for recommendation systems? ID- vs. modality-based recommendation models revisited. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023).

[3] Chen, J., & Suzumura, T. (2024). A prompting-based representation learning method for recommendation with large language models. arXiv. https://arxiv.org/abs/2409.16674

[4] Zhang, C., Zhang, H., Wu, S., Wu, D., Xu, T., Zhao, X., Gao, Y., Hu, Y., & Chen, E. (2025). NoteLLM-2: Multimodal large representation models for recommendation. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), 2815–2826.

[5] Thaipisutikul, T., Shih, T., Enkhbat, A., & Aditya, W. (2022). Exploiting long- and short-term preferences for deep context-aware recommendations. IEEE Transactions on Computational Social Systems, 9(4), 1237–1248.

[6] Friedman, L., Ahuja, S., Allen, D., Tan, Z., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., Chu, B., Chen, Z., & Tiwari, M. (2023). Leveraging large language models in conversational recommendation systems. arXiv preprint, abs/2305.07961.

[7] Yang, F., Yue, Y., Li, G., Payne, T. R., & Man, K. L. (2024). Chain-of-thought prompting empowered generative user modeling for personalized recommendation. Neural Computing and Applications, 36, 21723–21742.

[8] Zhang, Z., & Wang, B. (2023). Prompt learning for news recommendation. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23).

[9] Palma, D. (2023). Retrieval-augmented recommendation system: Enhancing recommendation systems with large language models. In Proceedings of the 17th ACM Conference on Recommendation Systems (RecSys '23), 1369–1373.

[10] Fu, J., Yuan, F., Song, Y., Yuan, Z., Cheng, M., Cheng, S., Zhang, J., Wang, J., & Pan, Y. (2024). Exploring adapter-based transfer learning for recommendation systems: Empirical studies and practical insights. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), 208–217.

[11] Wu, Y., Xie, R., Zhu, Y., Zhuang, F., Zhang, X., Lin, L., & He, Q. (2024). Personalized prompt for sequential recommendation. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3376–3389.

[12] Ji, J., Li, Z., Xu, S., Hua, W., Ge, Y., Tan, J., & Zhang, Y. (2024). GenRec: Large language model for generative recommendation. In Advances in Information Retrieval: 46th European Conference on Information Retrieval (ECIR 2024), 494–502.

[13] Zhang, A., Chen, Y., Sheng, L., Wang, X., & Chua, T. (2024). On generative agents in recommendation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24).

[14] Zhang, J., Hou, Y., Xie, R., Sun, W., McAuley, J., Zhao, X., Lin, L., & Wen, J. (2024). AgentCF: Collaborative learning with autonomous language agents for recommendation systems. In Proceedings of the ACM Web Conference 2024 (WWW '24).

[15] Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Huang, X., Lu, Y., & Yang, Y. (2024). RecMind: Large language model powered agent for recommendation. In Findings of the Association for Computational Linguistics: NAACL 2024, 4351–4364.

[16] Ge, Y., Tan, J., Zhu, Y., Xia, Y., Luo, J., Liu, S., Fu, Z., Geng, S., Li, Z., & Zhang, Y. (2022). Explainable fairness in recommendation. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), 681–691.