Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL34784

Application Paradigms and Challenges of Large Language Models in Personalized Recommendations

Qiuyu Jin

School of Computer Science, Guangdong University of Finance, Guangzhou, China
13738484869@163.com

Abstract. Personalized recommendation systems are a key technology to solve information overload, but traditional methods rely heavily on user identity features and interaction, facing cold-start and cross-domain bottlenecks. Large language models(LLMs) provide new paths for recommendation systems with their semantic understanding and generation capabilities. This paper systematically reviews LLM technical routes in personalized recommendation via literature analysis, summarizing three paradigms: semantic representation enhancement(SRE) for data sparsity, generative recommendation(GR) for unified task modeling, and discriminative matching and interaction(DMI) for interactive decision-making. SRE alleviates the cold start problem by generating semantic embeddings; GR reconstructs the recommendation task into text generation to achieve task unity and interpretability; DMI utilizes reasoning capabilities to support complex preference understanding. This paper analyzes typical scenarios such as e-commerce, news, and music videos, and summarizes the challenges in three aspects: computing efficiency, generation controllability, and the evaluation system. Research shows that LLMs are driving the evolution of recommendation systems from collaborative signal statistics to semantic understanding.

Keywords: LLM, Semantic representation enhancement, Generative recommendation, Discriminative matching and interaction

1. Introduction

Personalized recommendation systems are widely used in Internet scenarios such as e-commerce, news, video & music, and have iteratively evolved through collaborative filtering, matrix factorization, neural collaborative filtering, and graph neural networks [1], but still suffer from inherent flaws: heavy reliance on user identity and explicit interaction records, and limited ability to capture item semantics. This leads to poor performance in cold-start and cross-domain scenarios.

Recently, LLMs such as GPT and LLaMA have achieved technological breakthroughs, opening new avenues for recommender system development [2]. By constructing fine-grained semantic features from item text descriptions, LLMs use their powerful semantic understanding, zero-shot learning, and generative abilities to both alleviate the cold-start problem and generate natural language explanations, enhancing model interpretability.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL34784

LLM research for recommendations is advancing quickly, yet current reviews rely on model architecture as a classification basis, missing a task-driven synthesis. Thereby, the paper is problem-oriented and divides existing research into three major paradigms through literature analysis: SRE for data sparsity, GR for unified task modeling, and DMI for interactive decision-making. Furthermore, this study systematically reviews the three paradigms and offers a comprehensive assessment. By highlighting major research advances and open challenges—such as evaluation efficiency and output controllability—it aims to provide a systematic technical reference for future research.

2. Major LLM methods for personalized recommendation

Based on how they function in recommendation systems, LLMs can be categorized into three paradigms: SRE for data sparsity, which uses LLMs as feature encoders to generate high-quality semantic embeddings; GR for unified task modeling, which reframes recommendation tasks as text generation and directly outputs results; and DMI for interactive decision-making, which leverages LLMs' reasoning capabilities to perform user-item matching via natural language prompts.

2.1. Semantic representation enhancement for data sparsity

SRE uses LLMs as feature extractors to convert user-item information into semantic embeddings for downstream models, serving as a mainstream solution to data sparsity. Recently, this field has moved beyond static, single semantic modeling into three innovative directions: item semantic encoding, semantic graph structure, and knowledge-semantic fusion.

2.1.1. Item semantic encoding

Early methods are represented by UniSRec, as shown in Figure 1, which relies on BERT to generate fixed static item semantic embeddings to achieve basic cross-domain cold start recommendations [3]. However, there are still rigid flaws in semantic modeling, and adaptation limitations when facing dynamic changes in user preferences and complex cross-domain scenarios.

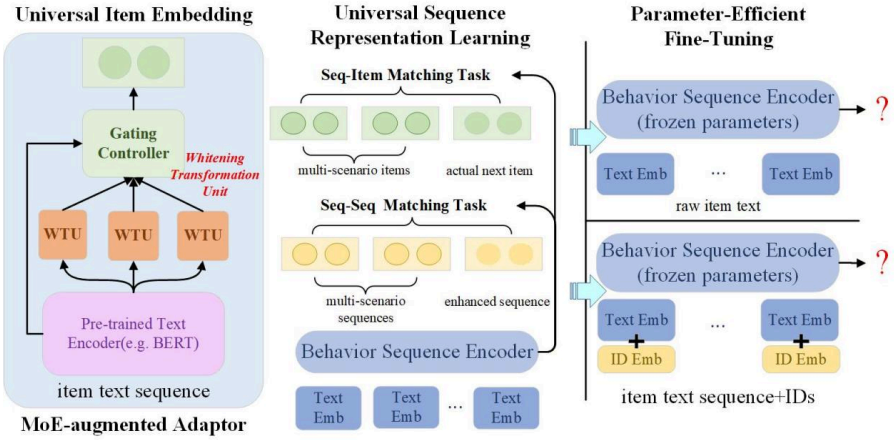


Figure 1. Framework of UniSRec

DOI: 10.54254/2755-2721/2026.GL34784

2.1.2. Semantic graph structure

SAGERec first utilizes LLMs to construct a semantic item relation graph, as shown in Figure 2, with a dynamically adaptive edge-weight learning mechanism, and then incorporates a lightweight Transformer to boost sequential recommendation accuracy, breaking the limitations of traditional fixed semantic representations [4].

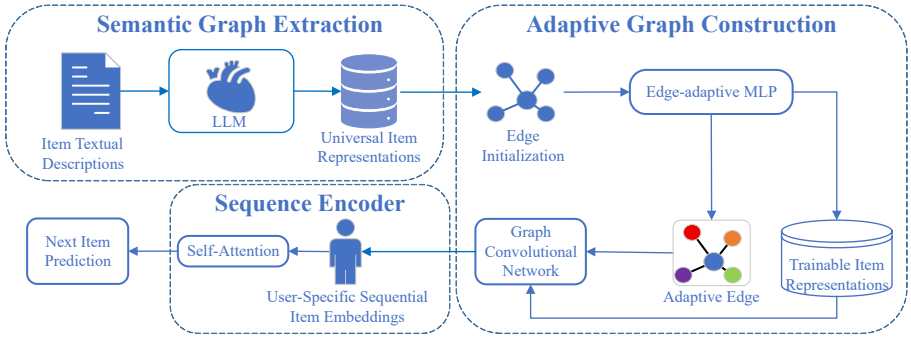


Figure 2. Framework of SAGERec

2.1.3. Knowledge-semantic fusion

K-RagRec, a latest breakthrough in this field, injects structured knowledge from a knowledge graph into LLMs via retrieval augmentation to enhance semantic representation, as illustrated in Figure 3 [5]. This method employs a popularity selective retrieval mechanism and a graph encoding architecture to extract high-quality structured information, thus improving inference efficiency while effectively mitigating the problem of neglecting knowledge structural relationships.

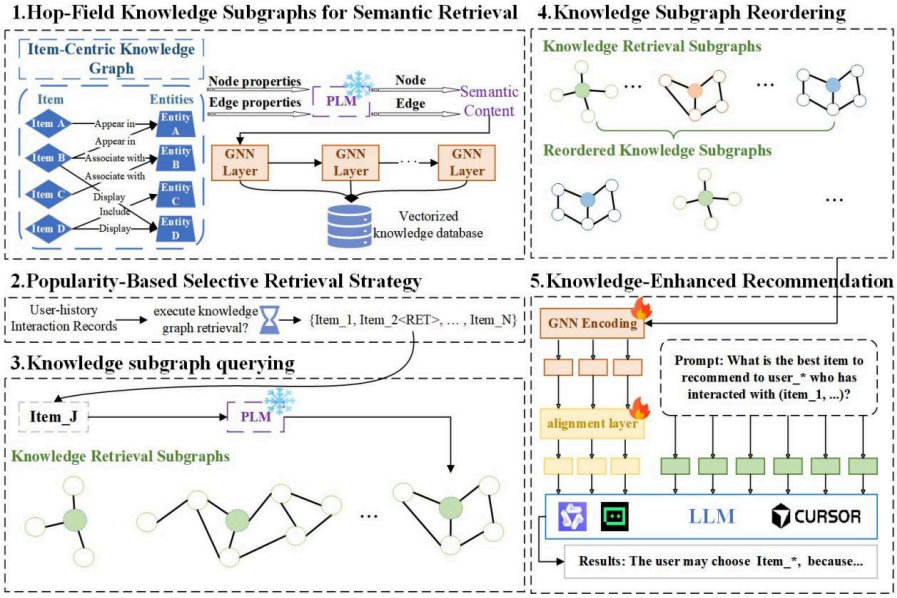


Figure 3. Overall workflow of K-RagRec

2.2. Generative recommendation for unified task modeling

GR relies on LLMs to perform end-to-end retrieval, ranking, and explanation generation, overcoming the limitations of traditional models that only output prediction scores and suffer from

DOI: 10.54254/2755-2721/2026.GL34784

fragmented task scenarios, and can be divided into two categories: instruction fine-tuning and prompt-based inference.

2.2.1. Instruction fine-tuning

Traditional models struggle with feature sharing and general representation learning. However, P5 unifies various recommendation sub-tasks into a text generation framework and achieves cross-task representation sharing, effectively enhancing generalization performance [6]. Though fully utilizing multi-task data, it requires updating all model parameters, resulting in a relatively high training cost.

2.2.2. Prompt-based inference

To reduce the high cost and complex implementation of instruction fine-tuning, the prompt reasoning paradigm represented by ChatGPT activates LLMs' zero-shot reasoning capabilities via prompt design without parameter updates. It adapts to cold-start and explainable recommendation scenarios [7], yet suffers from hallucination, out-of-domain recommendations, and context length limitations.

Retrieval-Augmented Generation introduces an external knowledge base to enhance input while preserving zero fine-tuning. OptiRAG-Rec combines RAG with a multi-head early exit mechanism and graph convolutional networks to boost retrieval efficiency and reduce computation without sacrificing accuracy [8], yet current RAG methods still struggle to balance retrieval accuracy and inference efficiency.

2.3. Discriminative matching and interaction for interactive decision-making

Based on interaction depth, DMI methods fall into three categories.

Prompt-based matching uses structured prompts to guide LLMs in making judgments. For instance, GollaRec integrates user-item interaction graphs into a multi-modal LLM via Graph-of-Thought, capturing structural information through graph instruction tuning [9].

Chain-of-thought reasoning decomposes recommendations into multi-step inference, producing a transparent logical chain from preference analysis to final recommendation. GREAME, a recent breakthrough, unifies collaborative-semantic alignment with curriculum-based reasoning to stabilize inference under sparse data [10].

Agent-based interaction treats the recommender as an autonomous agent capable of multi-turn dialogue and dynamic tool execution. DeepAgent achieves autonomous thinking and tool discovery. Using autonomous memory folding and ToolPO, it attains SOTA performance in multi-tool calling and open retrieval, significantly enhancing stability and accuracy in long interactions [11].

3. Typical application scenarios and major challenges

3.1. Typical application scenarios

3.1.1. E-commerce recommendation

In e-commerce, where product text is abundant and cold-start is severe, LLMs extract fine-grained semantics from titles, specs, and reviews to conquer cold-start barriers, generate natural language explanations to boost trust, and support cross-domain recommendations, thus handling frequent updates and large traffic [2].

DOI: 10.54254/2755-2721/2026.GL34784

3.1.2. News recommendation

News recommendation faces challenges such as strong timeliness, complex topics, and dynamic user interests. LLMs analyze news themes and sentiment tendencies, overcoming the limitations of keyword matching. With zero-shot prompting capabilities, they quickly respond to emerging hot topics, effectively mitigating filter bubbles and improving recommendation diversity and user satisfaction [2].

3.1.3. Video & music recommendation

In this domain, content tags are sparse, and manual annotation incurs high costs. Traditional recommendations relying on interactive data struggle to mine deep semantic features. LLMs learn semantic representation from texts such as introductions and subtitles, reducing over-reliance on artificial tags, accurately capturing users' implicit preferences, and achieving both accuracy improvement and long-term experience optimization [2].

3.2. Major challenges

Despite the significant potential of LLMs in recommendation systems, several critical challenges remain unresolved. In terms of computational efficiency, LLMs suffer from high inference latency and large memory footprint, making it difficult to meet the low-latency requirements of real-time recommendation systems [2]. Regarding generation controllability, models may suffer from hallucinations, recommending non-existent items or fabricating unreasonable recommendation rationales [8]. For the evaluation system, traditional ranking metrics such as NDCG and Hit Rate are not fully compatible with generative outputs, highlighting an urgent need to establish unified and automated evaluation standards [2].

4. Conclusion

This paper reviews three LLM paradigms for personalized recommendation: SRE alleviates data sparsity, GR unifies tasks and enables interpretability, and DMI supports preference reasoning. Despite challenges in efficiency, controllability, and evaluation, LLMs are shifting RS from collaborative-statistical to cognitive-semantic models.

However, this literature-based study lacks quantitative validation, leaving cross-paradigm performance comparison and in-depth discussions of efficiency, privacy, and fairness for future work. Future research may focus on lightweight LLM architectures for real-time deployment, hallucination mitigation, unified benchmarks, and multimodal trustworthy recommender agents.

References

[1] Rahmatikargar B, Zadeh P M, Kobti Z. Two Decades of Recommender Systems: From Foundational Models to State-of-the-Art Advancements (2004–2024) [J/OL]. ACM/IMS Journal of Data Science, 2024.

[2] Lin J H, Dai X Y, Xi Y J, et al. How Can Recommender Systems Benefit from Large Language Models: A Survey [J]. ACM Transactions on Information Systems, 2025, 43(2): 1-47.

[3] Hou Y P, Mu S L, Zhao W X, et al. Towards Universal Sequence Representation Learning for Recommender Systems [C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 585-593.

[4] Cui, Wanna, Lam H-K. SAGERec: Semantic-Aware Global Graph-Enhanced Representation Learning for Sequential Recommendation. Electronics. 2025; 14(24): 4844.

DOI: 10.54254/2755-2721/2026.GL34784

[5] Wang S J, Fan W Q, Feng Y, et al. Knowledge graph retrieval-augmented generation for LLM-based recommendation [J/OL]. arXiv preprint, 2025. https: //arxiv.org/abs/2501.02226.
[6] Geng S, Liu S, Fu Z, et al. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5) [J]. 2022.
[7] Di Palma D, Biancofiore G M, Anelli V W, et al. Evaluating ChatGPT as a Recommender System: A Rigorous Approach [J]. 2023.
[8] Zhou H, Gu H, Liu X, et al. The Efficiency vs. Accuracy Trade-off: Optimizing RAG-Enhanced LLM Recommender Systems Using Multi-Head Early Exit [J]. 2025.
[9] Yi Z X, Iadh Ounis. A Multi-modal Large Language Model with Graph-of-Thought for Effective Recommendation [C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. Albuquerque: Association for Computational Linguistics, 2025: 1591-1606.
[10] Hong M J, Zhou Z T, Guo Z R, et al. Generative Reasoning Recommendation via LLMs [J/OL]. arXiv preprint arXiv: 2510.20815, 2025.
[11] Li X, Jiao W, Jin J, et al. DeepAgent: A General Reasoning Agent with Scalable Toolsets [J]. In Proceedings of the ACM Web Conference 2026 (WWW '26). Association for Computing Machinery, New York, NY, USA, 2219–2230.