Research on Large Model Empowerment for Recommender Systems

Yan Yang, Sai Wang
Computer School, Central China Normal University, Wuhan, China

Abstract: To address challenges such as data sparsity and cold start issues in traditional recommender systems, this paper explores the application and challenges of large model empowerment in recommender systems. It outlines three primary application directions: optimizing existing algorithms through "pre-training-fine-tuning," transforming recommendation models via conversational interaction and AIGC, and enabling proactive decision-making through agents. It analyzes bottlenecks including computational cost, data privacy, model interpretability, and long-term adaptability. The paper concludes that future progress requires advancing integration through technical optimization, scenario adaptation, and risk management to support personalized recommendation development in the digital economy.

Keywords: Recommender System; Large Language Models; Intelligent Agents

1. Introduction

Recommender systems analyze users' historical behavior data and preference features to deliver personalized content, finding widespread application across e-commerce, social media, content streaming, and other domains [1]. Core technical approaches primarily fall into three categories: collaborative filtering, content-based recommendation, and hybrid recommendation [2]. Collaborative filtering relies on similarities between users and items for recommendations; content-based recommendation analyzes item characteristics to predict users' latent interests; hybrid recommendation methods combine the strengths of multiple techniques to deliver more comprehensive and precise results [3]. However, traditional recommender systems often face significant limitations when addressing data sparsity and cold-start problems [4]. In recent years, the rapid evolution of large language models (LLMs) has introduced novel approaches and technical pathways to tackle these core challenges [5][6].

Large models refer to deep learning models with parameter scales reaching extremely high levels. Their core technical architecture, exemplified by the Transformer, represents one of the central technological forms in the current field of artificial intelligence [7]. In terms of parameter scale, these models typically exceed billions of parameters, with some advanced models reaching hundreds of billions to trillions of parameters. This massive parameter capacity enables them to process vast amounts of data and capture complex patterns, demonstrating exceptional performance in core tasks such as information extraction and pattern recognition [8][9]. They have already achieved technological implementation and breakthroughs in multiple domains including text processing, image recognition, and multimodal interaction. In practical applications, numerous representative large model products have emerged globally. For instance, OpenAI's ChatGPT series, with its latest iteration GPT-4 exceeding 1 trillion parameters, leverages deep semantic understanding and high-quality language generation capabilities to efficiently handle complex dialog interactions and multi-step task processing [10]. Meta (formerly Facebook)'s LLMam3.1 model, built on 750 billion parameters, excels in fusing text, image, and audio data. It analyzes users' cross-media behaviors and preferences to deliver precise, personalized recommendation services [11]. Additionally, domestic tech companies continue to advance in this field. Baidu's knowledge-enhanced large model and iFlytek's Spark AI large model, among others, have developed differentiated technical advantages tailored to specific scenarios, supporting intelligent upgrades across industries. From a technology implementation perspective, hardware innovation provides critical support for large model development. The widespread adoption and performance improvements of specialized computing hardware like GPUs and

TPUs have significantly reduced computational costs during model training and deployment, accelerating their transition from laboratory technologies to industrial applications. Looking ahead, large models will see deeper integration into recommender systems, with their capabilities expanding into diverse scenarios such as education, healthcare, and e-commerce. In education, they can deliver personalized knowledge recommendations and tutoring plans based on user learning data; in healthcare, they can analyze multimodal medical data to assist in diagnosis and generate treatment suggestions; in e-commerce, they can precisely match user consumption needs with product information, enhancing transaction efficiency and user experience. Overall, large models will become a core technological driver propelling digital transformation across industries and elevating user service quality [12].

2. Applications of Large Models in Recommender Systems

As large model technology continues to advance, recommender systems will encounter both new opportunities and challenges. Key directions include the direct application of large models within existing recommendation algorithms, the transformation of traditional recommendation paradigms, and the emergence of intelligent recommendation agents.

2.1 Direct Application of Large Models in Existing Recommendation Algorithms

The core logic of directly applying large models within existing recommendation algorithm frameworks lies in leveraging technological integration to address inherent limitations of traditional algorithms, thereby enhancing the generalization capabilities and scenario adaptability of recommender systems. While traditional recommendation algorithms (such as collaborative filtering, matrix factorization, and factor machines) can deliver stable performance in scenarios with sufficient data, they commonly face critical challenges including data sparsity, cold-start problems, and insufficient feature representation capabilities. Leveraging their massive parameter count for strong fitting capabilities and the domain-agnostic knowledge accumulated during pre-training, large models can be directly integrated into existing recommendation workflows through the "pre-training - fine-tuning" technical paradigm, thereby overcoming these core challenges.

Regarding data sparsity, large models leverage general semantic knowledge acquired during pre-training (such as deep feature representations of text and images) to augment sparse samples in user-item interaction data. Taking e-commerce recommendation as an example, when a new product lacks user interaction records due to its recent launch, large models can perform semantic similarity matching between pre-trained feature representations (derived from product titles, detail page text, and images) and users' historical preferences. This effectively mitigates the cold-start problem for items. At the algorithmic optimization level, large models can directly replace core components in traditional recommendation algorithms: During the fine-ranking stage, large models based on the Transformer architecture replace traditional Multi-Layer Perceptrons (MLPs). By leveraging self-attention mechanisms to capture long-range dependencies in user behavior sequences, they more accurately uncover dynamic user preferences. In the recall phase, the large model's semantic retrieval capabilities enable the construction of a content-semantic recall channel, complementing traditional collaborative filtering recall to significantly enhance the diversity and relevance of recall results.

Furthermore, the multimodal processing capabilities of large language models can expand the data source dimensions of existing recommendation algorithms. Traditional recommendation algorithms primarily rely on text and numerical structured data, whereas large models can integrate and process multimodal data such as text, images, audio, and video. Through cross-modal feature alignment techniques, they transform users' multidimensional behaviors (e.g., watching videos, listening to audio, posting text-image content) into representation vectors within a unified feature space, enabling more comprehensive user interest modeling. Taking short-video recommendation as an example, large models can simultaneously analyze visual content features, audio style characteristics, and textual semantic features from user comments. This generates recommendations that better align with users' comprehensive preferences, enhancing recommendation accuracy while optimizing user satisfaction.

2.2 Transformation of Traditional Recommendation Models

The integration of large models is driving fundamental transformations in traditional recommendation paradigms across three dimensions: interaction logic, content formats, and service models. This shifts away from the static "push-to-receive" framework toward a new paradigm characterized by proactive and personalized recommendations.

At the interaction logic level, traditional recommendation models primarily rely on unidirectional interactions like "user browsing - system push," where user needs are expressed through indirect actions such as clicks and bookmarks, resulting in low efficiency in conveying user demands. Leveraging mature natural language processing (NLP) technology, large language models (LLMs) have pioneered a novel "conversational recommendation" interaction model: users can directly describe their needs in natural language (e.g., "Recommend a lightweight down jacket suitable for winter outdoor running"). The LLM extracts key demand features through semantic analysis (scenario: winter outdoor running; attributes: lightweight; category: down jacket), dynamically adjust recommendation strategies based on historical preferences, and support multi-round follow-up queries (e.g., "Do you require waterproof functionality?") to achieve precise demand fulfillment. This interaction model substantially reduces the cost of user demand expression while significantly enhancing the relevance and efficiency of recommendations.

At the content format level, traditional recommender systems center on "existing content distribution"—selecting resources matching user preferences from pre-existing libraries. This approach suffers from content homogenization and innovation deficits. Large model-driven generative AI (AIGC) technology propels recommender systems from "content distributors" to "content creators," ushering in a new "generative recommendation" paradigm. In news recommendation scenarios, large models can generate personalized news summaries or feature reports based on users' past reading preferences (e.g., interest in technology or preference for in-depth analysis articles). In music recommendation scenarios, they can create exclusive music snippets or playlists tailored to users' preferred genres, instruments, and rhythmic characteristics. In educational recommendation scenarios, they can generate customized practice questions and explanatory content based on students' learning progress and knowledge gaps. This generative recommendation not only enriches content diversity but also enhances user engagement and value through "exclusive content creation."

At the service paradigm level, traditional recommender systems focus on fulfilling single-scenario needs (e.g., product recommendations in e-commerce, series recommendations on video platforms). Large models' knowledge integration capabilities propel recommender systems toward "cross-scenario, full-domain services." By constructing a unified, cross-domain preference model for users, large models enable demand linkage across different scenarios. For example, after a user searches for "5-day tour to Sanya, Hainan" on a travel platform, the recommender system can: - Link with e-commerce platforms to push travel essentials like sunscreen and sandals - Link with video platforms to recommend Sanya travel guide videos - Link with food platforms to suggest local Sanya cuisine This creates a closed-loop "scenario - need - service" closed loop, delivering end-to-end intelligent recommendations.

2.3 Driving Intelligent Recommendation Agents

Large language models provide core technological support for building intelligent recommendation agents, elevating recommender systems from "passive response tools" to "proactive decision-making agents." This endows them with planning, memory, interaction, and adaptive capabilities, enabling end-to-end processing of complex recommendation tasks.

From a technical architecture perspective, the intelligent recommendation agent based on large models utilizes the large model as the "core decision-making unit" to integrate the planning module, memory module, and multimodal interaction module, forming a complete intelligent decision-making chain. The large model undertakes the functions of demand understanding, strategy generation, and task scheduling: it understands users' explicit demands and implicit preferences through natural language processing technology, and generates recommendation strategies by combining pre-trained knowledge with real-time data; the planning module breaks down recommendation tasks based on user goals (such as breaking down "weekend family leisure plan" into subtasks like "attraction recommendation - transportation planning - dining reservation - accommodation selection") and formulates a step-by-step execution plan; the memory module is divided into short-term memory and long-term memory, with the short-term memory storing the interactive information of the user's current session (such as real-time feedback of "disliking crowded attractions"), and the long-term memory storing the user's historical preferences, behavioral habits, and scene adaptation characteristics (such as "preference for parent-child leisure activities" and "tendency towards short-distance self-driving for weekend travel"). Dynamic updates and precise matching of preferences are achieved through the large model's memory invocation mechanism.

Functionally, the intelligent recommendation agent possesses three core capabilities: First, dynamic demand adaptation-capturing real-time shifts in user needs and contextual changes (e.g., automatically switching recommendations from outdoor weekend picnics to indoor family venues and home movies when rain is forecasted). Second, multi-objective optimization capability, balancing user satisfaction, content diversity, and platform commercial goals during recommendations (e.g., in e-commerce recommendations, ensuring product-user preference alignment while expanding categories to boost new discovery rates, all while meeting platform sales and profit targets). Third, autonomous learning and evolution capabilities. By continuously learning from user feedback (including explicit evaluations like "not interested" or "highly accurate recommendations," as well as implicit feedback such as dwell time and purchase conversion rates), the system dynamically optimizes recommendation strategies, achieving a closed-loop evolution of "use - feedback - iteration."

From an application perspective, intelligent recommendation agents have demonstrated practical value across multiple domains: In education, agents can generate personalized learning paths based on student data (e.g., types of incorrect answers, study duration), recommend tailored course resources, and simulate the role of a "private tutor" through real-time interactive Q&A. In healthcare, agents integrate user health metrics (age, medical history, exam reports) with lifestyle habits (dietary preferences, exercise frequency) to recommend tailored wellness plans, educational content, and medical services, empowering users in health decisions; In office settings, AI agents analyze user tasks (e.g., "draft quarterly sales reports") to recommend relevant historical documents, data charts, and template tools. They also integrate with calendar systems to schedule task execution times, boosting productivity. As large-model technology continues to evolve, intelligent recommendation agents will advance toward "contextualization, personalization, and autonomy," becoming core hubs connecting user needs with service resources.

To clearly illustrate the three pathways through which large models empower recommender systems, their core comparisons are summarized in Table 1.

Table 1. Comparison of the Three Major Application Directions of Large Models in Recommender Systems

Application Direction	Core Enabling Mechanism	Key Value Proposition

1. Existing Algorithm Optimization	Knowledge enhancement;
	Architecture upgrade;
	Multimodal fusion
		Directly improves recommendation accuracy and generalization, effectively alleviating data sparsity and cold-start problems.

2. Transformation of Traditional Paradigms	Conversational recommendation;
	Generative recommendation;
	Cross-domain service integration
		Transforms the passive interaction paradigm into an active, exclusive, and coherent recommendation experience.

3. Agent-Driven Recommendation	Human-like architecture;
	Dynamic decision-making;
	Continuous evolution
		Elevates the recommender system to an “active decision-making assistant” with comprehension, planning, and evolutionary capabilities.

3. Challenges and Limitations

Despite demonstrating immense potential and innovation in recommender systems, large language models still face significant challenges and limitations in practical applications.

3.1 Computational Overhead and Scalability Bottlenecks

The training and inference processes of large models exhibit strong dependence on hardware computational power. With parameter counts typically ranging from billions to trillions, they require sustained computational support from large-scale GPU or TPU clusters. This directly leads to high costs during system deployment, including hardware procurement, data center maintenance, and electricity consumption. For small and medium-sized enterprises or resource-constrained scenarios, the investment required for a complete computing infrastructure often exceeds their capacity, creating significant barriers to technological adoption. Even with sufficient computational resources, large models face severe challenges in delivering real-time responsiveness during high-concurrency recommendation scenarios, such as e-commerce sales events or peak traffic periods for short videos. Complex model inference processes can easily increase recommendation latency, thereby degrading user experience. When optimizing response performance through techniques like model compression and quantization, organizations often face the trade-off of reduced recommendation accuracy. Balancing computational costs, system performance, and recommendation precision has thus become the core challenge constraining the large-scale deployment of large model recommender systems.

3.2 Data Quality Deficiencies and Privacy Security Risks

The core efficacy of recommender systems relies on data-driven foundations, while large models impose even more stringent requirements on the "massive scale" and "high quality" of training data. On one hand, if training data contains noise or inherent biases (such as extreme preference data in user behavior or sample skew during data collection), large models amplify these flaws through deep learning processes. This leads to homogenized and discriminatory recommendations, manifesting as persistent push of single-type content, neglect of niche user needs, or unfair recommendations based on sensitive characteristics like gender or region. On the other hand, training large models requires integrating multidimensional information including historical user behavior, personal preference data, and contextual information. Such datasets often contain privacy-sensitive content such as consumption records, health metrics, and social relationships. Although privacy-preserving techniques like federated learning and differential privacy can mitigate data leakage risks to some extent, critical challenges remain unresolved in practical applications. These include balancing data anonymization with model performance, establishing compliant cross-platform data sharing mechanisms, and defining clear boundaries for user privacy authorization. Any data breach or misuse would directly trigger legal compliance risks and user trust crises, severely hindering the system's widespread adoption.

3.3 Lack of Model Interpretability and Weakened Trust Mechanisms

Traditional recommendation algorithms (e.g., collaborative filtering, logistic regression) exhibit high transparency in their recommendation logic. Their decision-making basis can be quantitatively explained through feature weight analysis and similarity calculation processes. In contrast, large models centered on the Transformer architecture are inherently "black-box" models. Their recommendation decisions rely on complex nonlinear mapping relationships among massive parameters, making it difficult to trace the specific generation mechanisms of recommendation results through a logical chain. This lack of interpretability triggers multiple issues: From the user perspective, the inability to clarify the basis for generated recommendations can erode trust in the system. Precise recommendations without apparent justification may even provoke resistance due to perceived excessive privacy intrusion. From a developer perspective, when recommendation outcomes deviate (e.g., irrelevant content or misleading consumption information), pinpointing root causes in model parameter settings or training processes becomes challenging, resulting in inefficient troubleshooting and system optimization. From a regulatory perspective, the lack of explainability prevents authorities from effectively verifying whether recommendation practices comply with relevant laws and regulations (such as the Advertising Law and

Anti-Unfair Competition Law). This significantly increases regulatory complexity and compliance risks, thereby limiting the application scope of large-model recommender systems in critical sectors like finance and healthcare, where controllability demands are exceptionally high.

3.4 Insufficient Long-Term Adaptability and Shortcomings in Dynamic Demand Matching

User demands are not static but dynamically evolve over time, across contexts, and with personal states (e.g., shifting interests, emerging needs). Concurrently, external environments introduce uncertainties like market trend shifts, policy adjustments, and breaking events. Large models, characterized by lengthy training cycles and high update costs, typically rely on offline training and periodic fine-tuning based on historical data, making it difficult to swiftly capture real-time shifts in user demands. On one hand, models can fall into a "path dependency" trap, persistently pushing content highly aligned with users' past preferences while failing to respond promptly to new interests, resulting in "lagging" recommendation outcomes. On the other hand, when confronted with entirely new application scenarios or niche user demands, large models struggle to achieve rapid learning and adaptation due to the lack of relevant samples in their training data, leading to the derivative issue of "dynamic cold start." Furthermore, the "continuous learning" technology of large models remains in its developmental stage. The technical framework for real-time absorption of new data and dynamic updating of recommendation logic while preventing the forgetting of historically effective knowledge is still immature. This results in insufficient long-term adaptability, making it difficult for models to consistently align with users' dynamically changing demand characteristics.

4. Conclusion

This paper systematically investigates the integrated application of large models and recommender systems. Results indicate that large models, leveraging their strong fitting capabilities, multimodal processing abilities, and accumulated general knowledge, provide breakthrough solutions for traditional recommender systems. By employing "pre-training - fine-tuning" to mitigate data sparsity and cold-start issues, leveraging conversational interaction and AIGC to drive the transformation of recommendation models toward "dynamic interaction + generative services," and relying on agent architecture to enable the system's leap toward becoming an "active decision-making assistant." This approach reshapes service logic and enhances user experience in fields such as e-commerce, education, and healthcare, while simultaneously expanding the application boundaries of recommender systems and establishing a new ecosystem of more precise and intelligent personalized recommendations. However, it should be noted that large-scale application of large models still faces challenges such as high computational costs, significant data privacy risks, weak model interpretability, and insufficient long-term adaptability. These issues not only constrain implementation scope but also clarify key areas for future research.

Looking forward, the integration of large models and recommender systems should advance along three interconnected fronts: technological optimization, scenario-specific customization, and risk governance. Technologically, the focus should be on exploring lightweight architectures, efficient continual learning, and explainability enhancement techniques to balance the trade-offs between performance and cost, as well as between adaptation and knowledge retention. For scenario adaptation, tailored solutions must be developed to meet the distinct demands of various industries, ensuring a tight fit between technological capabilities and practical business needs. In terms of risk governance, it is imperative to establish robust privacy protection and regulatory compliance frameworks to mitigate issues such as data leakage and algorithmic bias. In summary, large models have infused recommender systems with revolutionary momentum. Although significant challenges remain, continuous technological iteration, improved infrastructure, and the development of comprehensive industry standards will inevitably drive their large-scale adoption across more critical domains. They are poised to become the central hub connecting user needs with service resources, thereby providing crucial support for the high-quality development of the digital economy.

References

[1]WU Y, LU J. Multi Modal Information Generation and Recommendation Driven by Large Models [J]. Journal of Henan Normal University (Natural Science Edition), 2025, 53 (05): 145-151+181.

[2]ZHOU X. Research on Efficient Recommendation Algorithm Based on Large Language Model [D]. University of Electronic Science and Technology of China, 2025.

[3]LIU P, ZHANG M, WANG P, et al. Algorithm Optimization and Performance Evaluation of Intelligent Knowledge Recommendation System for Convenience Hotline Work Orders Based on Large Models [J]. Digital Technology and Applications, 2025, 43 (03): 16-18.

[4]YANG L. Personalized News Recommendation Method Enhanced by Large Language Model [D]. Lanzhou University, 2025.

[5]HUANG W, LI Z. Comparative Study of Recommendation Systems under Traditional Mode and Large Language Model [J]. Software Guide, 2025, 24 (02): 204-210.

[6]KA Z, ZHAO P, ZHANG B, et al. A Review of Recommendation Systems for Large Language Models [J]. Computer Science, 2024, 51 (S2): 11-21.

[7]ZHU M. Research on Personalized Resource Recommendation Method Based on Large Language Model [J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2024, 38 (05): 59-64.

[8]ZHOU X, DENG X, HUANG W. Large Models and Recommendation Systems Open a New Chapter in Personalized Recommendation [J]. Shanghai Informatization, 2024, (09):35-38.

[9]WU G, QIN H, HU Q, et al. Research on Large Language Models and Personalized Recommendations [J]. Journal of Intelligent Systems, 2024, 19 (06): 1351-1365.

[10]YE C. Overview of Large Language Model Recommendation Techniques [J]. Electronic Components and Information Technology, 2023, 7 (12): 127-131.

[11]WU Z, TANG Y, LIU H. Survey of Personalized Learning Recommendations [J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(1): 21-40.

[12]BROWN T, MANN B, RYDER N, et al. Language Models are Few-shot Learners [J]. Advances in Neural Information Processing Systems, 2020, 159: 1877-1901.