
A Survey on Sequential Recommendation Systems: Leveraging LLMs and NLP for Personalized User Behavior Modeling

www.surveyx.cn

Abstract

This survey paper provides a comprehensive overview of sequential recommendation systems, emphasizing the integration of large language models (LLMs) and natural language processing (NLP) to enhance personalized user behavior modeling. The paper is structured to explore the evolution, methodologies, and challenges of sequential recommendation systems, with a focus on addressing cold-start issues, dynamic user behavior, and data sparsity. It highlights the role of LLMs and NLP in refining user preference modeling and improving recommendation accuracy through innovative frameworks and methodologies. Key advancements include the use of multimodal data, embedding fusion, and novel architectural frameworks such as dual-tower retrieval and graph neural networks. The survey also discusses the impact of personalization on user satisfaction and engagement, underscoring the importance of ethical considerations and user privacy. Future directions suggest continued exploration of adaptive models, enhanced computational efficiency, and robust user behavior modeling techniques. Overall, the integration of LLMs and NLP represents a significant advancement in the field, offering new opportunities for innovation and improvement in user-centric recommendation solutions.

1 Introduction

1.1 Structure of the Survey

This survey offers a comprehensive overview of sequential recommendation systems, emphasizing the integration of large language models (LLMs) and natural language processing (NLP) to enhance personalized user behavior modeling. Section 2 presents background information and core concepts, including definitions and the evolution of recommender systems. Section 3 analyzes the architecture and methodologies of sequential recommendation systems, addressing challenges such as cold-start problems and dynamic user behavior, while highlighting innovative approaches and future directions. The role of LLMs and NLP is examined in Section 4, discussing their advantages, challenges, and innovative frameworks that utilize these technologies, as evidenced by benchmarks like [1]. Section 5 focuses on personalized recommendations and user behavior modeling, exploring techniques for modeling user behavior and the effects of personalization on user satisfaction and engagement. Current challenges and future directions, including data sparsity, scalability, and ethical considerations, are identified in Section 6. The survey concludes in Section 7, summarizing key insights and underscoring the importance of integrating LLMs and NLP in advancing sequential recommendation systems. Throughout, we incorporate findings from recent studies, such as the dual-tower retrieval architecture proposed in [2], to provide a nuanced understanding of the field. The following sections are organized as shown in Figure 1.

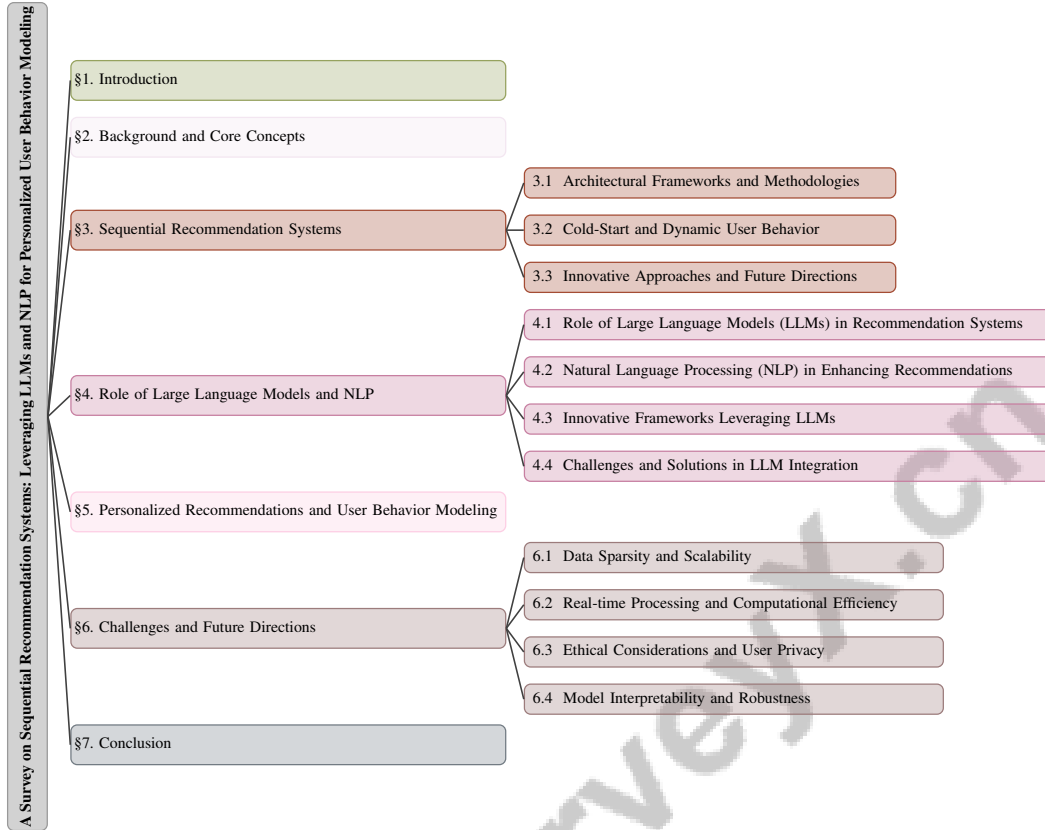


Figure 1: chapter structure

2 Background and Core Concepts

2.1 Definitions and Key Terms

Sequential recommendation focuses on predicting user preferences by analyzing past interactions, emphasizing the need to understand dynamic user behavior [3, 4]. Key elements include collaborative signals, semantic information, and addressing the long-tail challenge, which are crucial for comprehending these systems [5]. A fundamental aspect is the next-item recommendation, particularly in zero-shot settings, aiming to predict the subsequent item a user might engage with [6].

Recommender systems utilize advanced algorithms to suggest relevant items based on historical interactions, significantly impacting sectors like e-commerce and media streaming by enhancing user engagement. Recent progress, especially through pre-trained language models (PLMs) and frameworks like PRECISE, has improved these systems by merging collaborative signals with semantic information. This integration addresses long-tail items and cold-start scenarios [7, 5]. Large Language Models (LLMs) enhance recommendations by interpreting user-generated content, while Natural Language Processing (NLP) analyzes textual data linked to interactions, offering insights into user intents crucial for intent-aware recommendations.

Personalized recommendations consider unique preferences and situational contexts to predict next-best actions [8]. User behavior modeling is essential, analyzing interactions to boost recommendation accuracy and satisfaction [9]. Integrating LLMs aids in discovering latent item relations, enhancing recommendations through user interaction data and textual information [10]. Challenges such as long-tail distributions of user interactions and item popularity necessitate advanced techniques for mitigation [11].

2.2 Evolution of Recommender Systems

Recommender systems have evolved from basic content-based and collaborative filtering techniques, which relied on explicit user feedback and static datasets, to more advanced methods. Initial approaches faced challenges like data sparsity and the cold-start problem, with foundational methods such as Matrix Factorization and Markov Chains struggling with sparse datasets, leading to the development of similarity-based methods [12].

The advent of deep learning marked a pivotal shift, enabling the modeling of complex patterns in user interactions and improving recommendation accuracy [13]. Despite these advancements, challenges such as training inefficiencies in large parameter spaces and persistent cold-start issues remain [14]. The long-tailed distribution of user interactions and item popularity further complicates the development of effective sequential recommendation systems [15].

Recent innovations address these challenges through multi-behavior sequential recommendation systems, enhancing model performance by reducing inefficiencies and noise in user behavior data [14]. Graph neural networks have been explored to capture the dynamic nature of user preferences over time, improving the accuracy of current interest estimations based on historical data [16]. Knowledge graphs and auxiliary item relationships enrich contextual understanding and recommendation accuracy, especially in cold-start scenarios [17].

The application of natural language processing techniques introduces a new paradigm in recommender systems, unifying various tasks and enhancing the understanding of user intents and preferences [18]. However, existing LLM-based methods still face challenges related to inference time and computational efficiency [13].

The field is moving towards dynamic models that adapt to evolving user preferences and behaviors, emphasizing the need to address both temporal and contextual dimensions of user interactions [19]. The development of sequence-aware recommendation tasks highlights the importance of context adaptation, trend detection, repeated recommendations, and order constraints [20]. Current models continue to grapple with long-range dependencies and computational complexity, necessitating novel approaches. Additionally, biases in recommender systems, particularly from closed feedback loops, remain a concern as they can hinder user discovery and decision-making [21].

In recent years, the evolution of sequential recommendation systems has garnered significant attention within the academic community. These systems, which aim to predict user preferences based on their historical interactions, have become increasingly sophisticated. To better understand the various components that contribute to their effectiveness, it is essential to examine the underlying architectural frameworks and the challenges they face.

As illustrated in Figure 2, this figure presents a hierarchical categorization of architectural frameworks, challenges, and innovative approaches in sequential recommendation systems. It highlights three primary frameworks: experience-based, transaction-based, and interaction-based. Additionally, the figure addresses critical challenges such as the cold-start problem and dynamic user behavior, while also exploring novel methodologies and future research directions. By integrating these elements, we can gain a comprehensive understanding of the current landscape and the potential advancements in this field.

3 Sequential Recommendation Systems

3.1 Architectural Frameworks and Methodologies

Sequential recommendation systems have evolved significantly, employing diverse architectural frameworks and methodologies to enhance predictive accuracy and personalization. These systems are primarily categorized into experience-based, transaction-based, and interaction-based approaches, each utilizing deep learning techniques to address the complexities of user behavior [22]. Experience-based frameworks, like Side Sequential Network Adaptation (SSNA), optimize large language models (LLMs) by fine-tuning specific layers, leveraging pre-trained knowledge for effective recommendations [23]. Adasplit further refines user preferences into dynamic sub-sequences, capturing the evolving nature of user behavior [24].

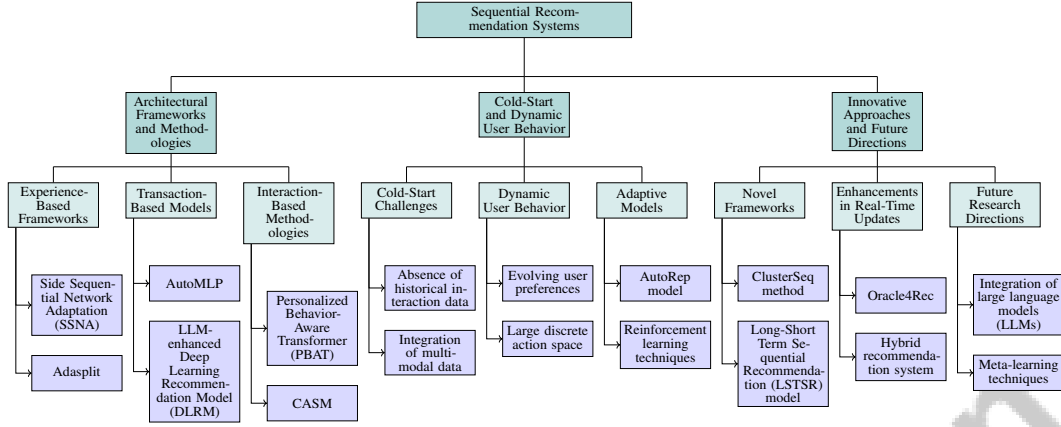


Figure 2: This figure illustrates the hierarchical categorization of architectural frameworks, challenges, and innovative approaches in sequential recommendation systems. It highlights experience-based, transaction-based, and interaction-based frameworks, addresses cold-start and dynamic user behavior challenges, and explores novel methodologies and future research directions.

Transaction-based models, such as AutoMLP, utilize multi-layer perceptron (MLP) architectures to model both long- and short-term user interests with linear computational complexity [25]. The LLM-enhanced Deep Learning Recommendation Model (DLRM) integrates multi-modal information through deep learning and LLMs to improve accuracy [26]. Interaction-based methodologies, exemplified by the Personalized Behavior-Aware Transformer (PBAT), capture nuanced user interactions to enhance recommendation performance [27]. CASM employs context-aware multi-head self-attention layers to capture dependencies among heterogeneous historical interactions [28].

Generative approaches like GenRec redefine sequential recommendation as a generative task, using a sequence-to-sequence Transformer architecture to align generated sequences with user preferences [29]. The M-GPT model extracts multi-behavior dependencies, generating multifaceted sequential patterns by capturing temporal preferences [30]. LRMRec ensures rapid inference and incremental processing, addressing inefficiencies in self-attention-based models [31]. PeaPOD enhances the recommendation process by distilling user preferences into personalized soft prompts [32].

Innovative methodologies, such as CALRec, enhance next-item prediction through a two-stage fine-tuning of pre-trained LLMs with contrastive and language modeling losses [33]. The Efficient Continuous Control framework (ECoC) utilizes unified actions from normalized user and item spaces to enhance reinforcement learning-based recommendations [34]. Additionally, the LLM-ESR framework merges semantic embeddings from LLMs with collaborative signals to improve recommendations for long-tail users and items [11].

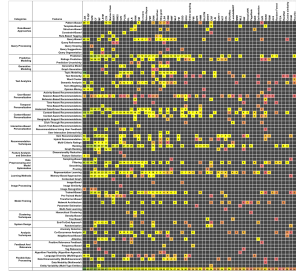
These methodologies collectively advance sequential recommendation systems by addressing challenges such as data sparsity, dynamic user behavior, and personalization needs. Techniques like graph-based learning, meta-learning, and transformer architectures significantly enhance prediction accuracy and user satisfaction. The SLMREC framework introduces a small language model specifically for sequential recommendations, utilizing knowledge distillation to align representations with larger models [3]. The PRECISE framework integrates collaborative signals and semantic information through a Mixture of Experts (MoE) structure for embedding fusion [5]. Novel approaches like time modeling and multi-task learning further enhance session-based model performance [35], while the RDRSR model employs a dynamic interest discriminator (DID) and allocator (DIA) to group user interactions into sub-sequences for representation [4].

The STDP framework incorporates statistical information into model pre-training to mitigate random noise effects in user action sequences, advancing sequential recommendation capabilities [36].

As illustrated in Figure 3, leveraging architectural frameworks and methodologies in sequential recommendation systems is crucial for enhancing recommendation precision and relevance. Figure 3 showcases various frameworks, including the "Dependency Score Booster for Event Type and Item Classification," which utilizes a neural network architecture for classifying event types and items based on input sequences. This highlights the integration of deep learning techniques to

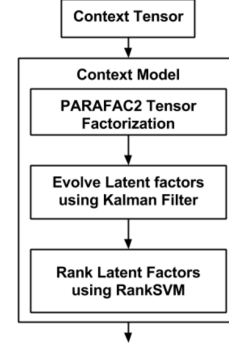


(a) Dependency Score Booster for Event Type and Item Classification[37]



(b) The image represents a table with a grid of cells, each cell containing a number representing the performance of different recommendation algorithms on various categories and features. The table is organized into rows and columns, with each cell containing a specific number.[38]

Fig. 3. Context Model to get relevance scores



(c) Context Model to get relevance scores[39]

Figure 3: Examples of Architectural Frameworks and Methodologies

enhance classification accuracy. Additionally, the performance table emphasizes the importance of categorizing recommendation algorithms, facilitating structured evaluations of their effectiveness. The "Context Model to get relevance scores" illustrates a flowchart employing PARAFAC2 tensor factorization and a Kalman filter to rank latent factors, demonstrating sophisticated methodologies for context-aware recommendations. Collectively, these examples underscore the diverse strategies employed in sequential recommendation systems to optimize user experience and predictive accuracy [37, 38, 39].

3.2 Cold-Start and Dynamic User Behavior

The cold-start problem poses a significant challenge in sequential recommendation systems due to the absence of historical interaction data for new users and items, hindering accurate recommendation generation [6]. Traditional models often struggle to capture interaction-level dependencies among diverse behavior types, limiting their effectiveness in dynamic environments [30]. Moreover, unidirectional architectures, such as Mamba, restrict the capacity to leverage future user-item interactions, leading to suboptimal predictions [40].

Addressing the cold-start issue necessitates innovative techniques capable of recommending items without prior user interaction data [6]. Integrating multi-modal data, such as text and images, enriches the recommendation process by providing additional context and enhancing personalization, although efficient utilization of such data remains challenging [26].

The dynamic nature of user preferences complicates the recommendation process, as interests can shift rapidly over time. Traditional methods exhibit inflexibility in adapting to these changes, necessitating frameworks like LIRD that can dynamically adjust to evolving user preferences [41]. Additionally, the large discrete action space associated with numerous items poses a significant challenge in learning effective recommendation policies, underscoring the need for models capable of managing this complexity [34].

To enhance the modeling of dynamic user behavior, it is crucial to capture collaborative information and semantic representations simultaneously, particularly in cold-start scenarios where such information is sparse [5]. Understanding the asymptotic behavior of user discovery within feedback loops is essential for improving recommendation accuracy, as these loops can significantly impact user engagement and exploration [21].

Developing adaptive models that continuously learn and adjust to evolving user preferences is essential for addressing cold-start situations and dynamic user behavior challenges. These models must integrate diverse data sources, such as user interactions and contextual information, while accommodating varying significance of different behaviors over time. Recent advancements like the AutoRep model demonstrate the importance of creating dynamic representations that reflect changing user behavior groups, while reinforcement learning techniques can simulate the evolution of user preferences, ultimately enhancing the accuracy and relevance of recommendations in real-world applications [42, 43, 24]. These advancements are critical for improving the personalization and effectiveness of recommendation systems.

3.3 Innovative Approaches and Future Directions

Recent advancements in sequential recommendation systems have introduced innovative frameworks and methodologies that address existing limitations while exploring new research directions. The ClusterSeq method exemplifies a novel approach by employing a clustering-based meta-learning framework that accommodates the unique preferences of minor users, thus enhancing personalization [44]. This approach highlights the importance of tailoring recommendations to diverse user groups to improve overall system efficacy.

The Long-Short Term Sequential Recommendation (LSTSR) model innovates by explicitly encoding short-term preferences and utilizing a memory mechanism with message passing to capture dynamic collaborative signals, effectively combining long- and short-term preferences [45]. This method provides a comprehensive view of user behavior, crucial for accurately predicting future interactions.

Oracle4Rec introduces an oracle-guiding module that minimizes the discrepancy between past and future information, aligning historical data with prospective user actions [46]. This innovation enhances the robustness of recommendation systems in dynamic environments.

The dual-tower retrieval architecture, combined with a self-supervised multi-modal pre-training approach [2], improves generalization across diverse datasets and enhances retrieval capabilities, leveraging multi-modal data to enrich the recommendation process and addressing traditional single-modal limitations.

In real-time updates, the hybrid recommendation system proposed by [43] supports online personalization without retraining model parameters, allowing for instantaneous updates to user and item history representation vectors, essential for maintaining recommendation relevance in rapidly changing environments.

Future research in sequential recommendation systems is likely to focus on further integrating large language models (LLMs) and enhancing model adaptability, efficiency, and diversity in recommendations. The exploration of meta-learning techniques, as demonstrated by the metaCSR framework, optimizes initialization for new users by identifying common patterns, representing a promising direction for addressing the cold-start problem [47]. Additionally, developing frameworks that integrate change point detection into the recommendation process will be vital for ensuring timely and accurate adaptations to evolving user preferences.

The innovative approaches and frameworks discussed in recent literature significantly enhance the field of sequential recommendation systems by addressing challenges such as personalized interest sustainability and contextual information integration. These advancements explore new research directions, including open-domain and explainable sequential recommendations, leveraging cutting-edge techniques like seq2seq models and multi-modal data integration. Collectively, they pave the way for developing more sophisticated and adaptive recommendation solutions that better capture both short-term user interests and long-term preferences, ultimately leading to improved user experiences [48, 49, 50, 20, 51].

4 Role of Large Language Models and NLP

The integration of Large Language Models (LLMs) and Natural Language Processing (NLP) is pivotal in the evolution of recommendation systems, enhancing user preference modeling and recommendation accuracy. This section explores the significant contributions of LLMs, focusing on

innovative frameworks and methodologies that optimize user interactions and improve the overall recommendation experience.

4.1 Role of Large Language Models (LLMs) in Recommendation Systems

LLMs advance recommendation systems by analyzing intricate user interactions, capturing complex behavior patterns to enhance personalization and accuracy. The PRECISE framework refines item representation by integrating ID and semantic embeddings, improving user interest modeling [5]. Similarly, the EIMF framework extracts explicit semantic interests from user data, enriching the recommendation process [8]. The SIGMA model leverages a bidirectional structure to enhance user preference modeling over time [40], while SLMREC demonstrates LLM efficiency with reduced parameters [3]. Dual-view modeling in LLM-enhanced methods addresses long-tail user and item challenges by caching semantic embeddings [11].

LLMs are also crucial in zero-shot next-item recommendations, predicting preferences without extensive training data [6]. Integrating recommender system knowledge into LLMs for natural language generation tailored to recommendation tasks offers promising avenues for narrative enhancement [52]. Frameworks like LIRD continuously update strategies based on interactions, enhancing personalization and relevance [41]. The ECoC framework employs unified action representation to facilitate continuous RL-based policies, adapting to dynamic environments [34]. LLMs enhance recommendation systems by deepening user interaction understanding and pushing innovation boundaries, as demonstrated by the RDRSR model’s adaptive interest modeling [4].

4.2 Natural Language Processing (NLP) in Enhancing Recommendations

NLP enhances recommendation systems by transforming user-generated content into actionable insights, refining personalization and accuracy. The P5 framework exemplifies this by converting interactions, descriptions, metadata, and reviews into unified natural language sequences, improving recommendation precision [18]. Generating natural language explanations for recommendations is crucial for contextualizing user ratings and item features, enhancing engagement and trust [53].

By leveraging NLP, systems analyze diverse information sources, producing more accurate recommendations. Advanced NLP techniques align user preferences with item characteristics, generating personalized narratives and leveraging user-generated content to improve understanding. This approach fosters an engaging user experience, increasing satisfaction with recommendation outcomes. Frameworks like P5 ensure tailored suggestions resonate with user preferences [18, 43, 52].

4.3 Innovative Frameworks Leveraging LLMs

Innovative frameworks leveraging LLMs enhance sequential recommendation systems by integrating advanced language processing with traditional methodologies. The HLLM architecture decouples item and user modeling, utilizing pre-trained LLM capabilities [54]. The E4SRec framework uses ID sequences for direct recommendation generation [55]. The MMRec framework integrates multimodal data, enhancing ranking model learning [26]. CALRec employs a two-tower fine-tuning approach, improving performance through mixed training objectives [33].

LLM-ESR utilizes semantic embeddings to enhance long-tail user and item recommendations, outperforming traditional methods [11]. The Laser framework optimizes LLMs with virtual tokens, enhancing parameter efficiency [56]. The LANCER framework improves personalization and accuracy by incorporating domain-specific knowledge [57]. CASM integrates multiple user behaviors, enhancing outcomes [28].

The REGEN dataset enriches recommendations by integrating user narratives [52]. LLMGR combines LLMs with GNNs for relational data modeling [10]. Future research may merge GNNs with LLMs to enhance performance [30]. These frameworks illustrate LLMs and NLP’s transformative impact on personalization, accuracy, and explainability, paving the way for sophisticated, user-centric solutions.

4.4 Challenges and Solutions in LLM Integration

Integrating LLMs into recommendation systems presents challenges in computational efficiency, semantic embedding quality, and preference modeling. Self-attention mechanisms require recomputing

item-to-item weights, incurring high costs [31]. LLM redundancy in intermediate layers hinders deployment [3]. Generating effective semantic embeddings depends on model capacity and data quality [5]. Modeling preferences in vast recommendation spaces is difficult, especially in zero-shot scenarios [6]. Iterative systems limit user discovery, affecting recommendation diversity [21].

Solutions include the ECoC framework for efficient training and continuous action space handling, improving engagement through policy learning [34]. MLP-based models reduce complexity and parameter count, capturing sequential correlations efficiently [58]. A comprehensive strategy enhances LLM integration, focusing on computational efficiency, semantic enrichment, and adaptability to preferences. Leveraging LLMs for natural language and multi-modal data analysis can improve recommendation accuracy and relevance, addressing challenges in interest diversity and coherence [59, 26, 60]. Overcoming these challenges leads to effective, scalable solutions, enhancing user satisfaction and engagement.

5 Personalized Recommendations and User Behavior Modeling

Understanding and modeling user behavior is pivotal in enhancing the effectiveness of personalized recommendation systems. This section explores various techniques designed to capture the complexities of user interactions and preferences, thereby refining the personalization process and delivering more relevant and timely recommendations.

5.1 Techniques for Modeling User Behavior

Modeling user behavior is essential for advancing the personalization and precision of recommendation systems. The PRECISE framework exemplifies this by employing embedding fusion and structured training processes to enhance user behavior modeling [5]. Linear Recurrent Dynamics (LRD) further demonstrates efficacy in uncovering complex item relationships, thereby enhancing user preference understanding [9].

In zero-shot next-item recommendation scenarios, the NIR prompting method effectively captures user preferences without extensive historical data [6]. The theoretical framework by [21] emphasizes the iterative nature of user behavior in collaborative filtering systems, highlighting the need to consider iterative interactions for improved recommendation accuracy.

These methodologies collectively advance recommendation systems by integrating diverse strategies that model the intricate dynamics of user interactions. By combining context-based methods with traditional collaborative filtering, these approaches enhance personalization and accuracy, addressing challenges like data sparsity and improving bias-variance trade-offs. They support real-time updates to user and item profiles, boosting system responsiveness to new interactions, and raise questions about potential user preference manipulation, highlighting the need for ethical considerations in recommendation algorithms [61, 43].

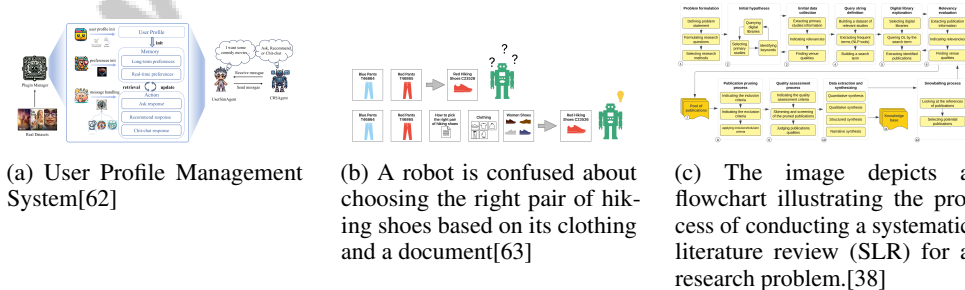


Figure 4: Examples of Techniques for Modeling User Behavior

As shown in Figure 4, diverse methodologies are employed in modeling user behavior to enhance personalized recommendation systems. The User Profile Management System manages profiles by initializing long-term and real-time preferences, foundational for tailoring experiences. The second image illustrates decision-making complexities when variables intersect, and the third depicts a

systematic literature review process, showcasing structured approaches to understanding research problems [62, 63, 38].

5.2 Impact of Personalization on User Satisfaction and Engagement

Personalization significantly enhances user satisfaction and engagement by aligning recommendations with individual preferences. This tailored approach fosters deeper user connections and increases interaction likelihood, as seen in Conversational Recommender Systems (CRS) which improve engagement and speed of recommendations compared to traditional methods [64].

By leveraging user-specific data, personalized systems offer relevant suggestions, enhancing satisfaction and encouraging platform use. This involves comprehensive analysis and forecasting of preferences using advanced modeling techniques that capture user behavior subtleties, as evidenced by research in user intent modeling and data integration [65, 61, 43, 38]. Users experience intuitive interactions, increasing satisfaction and engagement.

Personalized systems' adaptability ensures recommendations remain relevant, sustaining engagement and mitigating fatigue. Advanced language models and contextual knowledge incorporation enhance relevance, fostering engaging experiences by aligning content with sustained interests [57, 49]. Personalization's impact underscores the importance of advanced systems capable of understanding unique user preferences.

5.3 Future Directions in User Behavior Modeling

Future research in user behavior modeling for recommendation systems will focus on enhancing data processing, integrating diverse approaches, and optimizing objectives while ensuring privacy and security. Multi-behavior sequential recommendation systems will capture complex user interactions, leading to accurate and personalized recommendations [66].

Advanced models like NeuMF for friend recommendations and bidirectional models like BERT will enhance recommendation quality by providing deeper insights into user preferences [43]. These models leverage interaction contextual richness, improving preference prediction accuracy.

Extending frameworks like LLMGR for multi-domain recommendations will allow flexible interactions and improved performance across contexts [10]. This will enable systems to adapt to broader needs, enhancing applicability and effectiveness.

Future user behavior modeling will prioritize robust, adaptable, and secure frameworks navigating complex interactions across domains. This evolution will incorporate dynamic representation learning and pre-trained language models for enhanced preference understanding through textual data. Addressing challenges like data sparsity and improving bias-variance trade-offs will lead to personalized, accurate recommendations responding to real-time interactions [67, 43, 24]. These advancements will pave the way for sophisticated, user-centric solutions, enhancing satisfaction and engagement.

6 Challenges and Future Directions

Enhancing sequential recommendation systems requires addressing complex challenges related to data sparsity and scalability, which are critical for the effectiveness and efficiency of these systems. As user interactions evolve, understanding these challenges is essential for developing robust systems. The subsections below examine specific difficulties posed by data sparsity and scalability, highlighting innovative strategies to mitigate these obstacles and improve recommendation accuracy.

6.1 Data Sparsity and Scalability

Data sparsity and scalability significantly hinder sequential recommendation systems' ability to deliver precise, personalized suggestions. Data sparsity, especially among long-tail users and items, can impair model performance due to insufficient collaborative signals [11]. This issue is exacerbated in LLM-based recommendation systems (LLMRS), which require extensive datasets for optimal performance [59].

Scalability challenges arise from the computational demands of processing large datasets and complex algorithms. The LIRD method efficiently manages vast item spaces, addressing scalability and data sparsity [41]. Similarly, the LRURec framework offers a scalable solution with space complexity akin to RNN-based recommenders, effectively tackling data sparsity [31].

The PRECISE framework exemplifies overcoming data sparsity and cold-start challenges by capturing user interests across diverse scenarios, enhancing model adaptability [5]. Additionally, integrating multimodal data, as in the MMRec framework, enriches recommendations with contextual information, though care must be taken to avoid overfitting [26].

Despite these advancements, position bias in LLMs remains a significant challenge, necessitating further research to mitigate its impact [68]. The quality of user and item embeddings also affects performance in sparse datasets, underscoring the need for robust embedding techniques to enhance accuracy [34].

Innovative methodologies are required to balance computational efficiency with capturing complex user interactions. Integrating document context-based methods with collaborative filtering addresses data sparsity and supports online personalization, enabling real-time updates to user and item representations. Using large language models to extract semantic information enriches understanding of user interests, particularly in complex scenarios involving long-tail items and cold-start situations. These innovations contribute to more effective recommendation systems catering to diverse user bases while maintaining high accuracy and relevance [5, 43].

6.2 Real-time Processing and Computational Efficiency

Real-time processing and computational efficiency are critical for sequential recommendation systems, affecting user experience and system performance. The dynamic nature of user interactions necessitates advanced processing capabilities for timely and contextually relevant recommendations, adapting to temporal changes, contextual information, and user-generated content [61, 57, 43]. However, computational demands can be substantial, especially in large-scale environments.

Integrating Large Language Models (LLMs) enhances personalization and accuracy but increases computational costs. The EIMF framework, despite extracting explicit semantic interests, faces challenges related to computational efficiency during LLM inference [8]. Similarly, the TSSR's end-to-end learning paradigm incurs high training costs, requiring significant resources [69].

Innovative approaches like the A-LLMRec model, which integrates with various Collaborative Filtering Recommender Systems (CF-RecSys), reduce training and inference times [70]. Hardware-aware scanning acceleration algorithms further enhance computational efficiency, supporting real-time processing needs [71].

Certain methodologies, such as the MARS framework, may struggle with sparse interaction data, affecting performance and efficiency [72]. Similarly, DGSF may face challenges in unpredictable user interactions or sparse data scenarios [73].

Achieving real-time processing and computational efficiency requires integrating advanced algorithms with hardware optimizations to deliver timely, contextually relevant recommendations while managing training time and resource utilization. By improving predictive accuracy and addressing challenges like the cold start problem and popularity bias, researchers can create a more responsive and effective digital marketing experience [74, 75, 76, 77]. Addressing these challenges enhances the responsiveness and scalability of recommendation systems, improving user satisfaction and engagement.

6.3 Ethical Considerations and User Privacy

Ethical considerations and user privacy are paramount in developing sequential recommendation systems, particularly as these systems utilize Large Language Models (LLMs) to process extensive user interaction data. LLMs introduce challenges, including potential biases in recommendations due to training data characteristics and prompting strategies [6]. Addressing these biases is crucial for fairness and equity in recommendation systems, as the quality of candidate sets and prompting strategies significantly influence outcomes.

Reliance on collaborative signals in sparse data environments raises ethical concerns about recommendation accuracy and fairness, as these signals may not adequately represent diverse user interactions [21]. Targeted exploration strategies and optimizations are needed to enhance user discovery and address blind spots within algorithms, improving the system’s ability to provide equitable suggestions.

User privacy remains critical, especially given the computational demands of processing multimodal inputs and potential unauthorized data access. Techniques like differential privacy and federated learning are explored to protect user data while maintaining recommendation effectiveness [4]. Future research will focus on improving allocation processes to enhance computational efficiency and support privacy-preserving practices.

By prioritizing ethical considerations and user privacy, developers can create systems that deliver personalized and accurate suggestions while respecting user privacy, maintaining trust and satisfaction. This approach aligns recommendation systems with ethical standards and user expectations by integrating advanced methodologies that balance engagement with transparency, prioritizing user autonomy and minimizing manipulative practices. Frameworks to benchmark manipulation levels and incorporate collaborative and semantic information ensure systems respect user preferences, leading to a responsible and user-centric experience [61, 5].

6.4 Model Interpretability and Robustness

Model interpretability and robustness are crucial for transparency, reliability, and user trust in sequential recommendation systems. Interpretability is vital for understanding how models like PRECISE integrate collaborative and semantic information to enhance recommendations [5]. Structured prompting strategies guide Large Language Models (LLMs) in making accurate recommendations, emphasizing clear model interpretability [6].

Autonomous discovery of latent relations, as demonstrated by [9], improves model adaptability and performance, highlighting the necessity of robust models that adjust to new data and interactions. The LLMGR framework integrates LLMs and Graph Neural Networks (GNNs), leveraging their complementary strengths to enhance interpretability and robustness [10].

Future research aims to integrate few-shot learning capabilities into models like SLMREC, enhancing adaptability to new datasets without full retraining [3]. This adaptability is crucial for maintaining robustness in rapidly changing data landscapes, ensuring systems remain effective over time.

Emphasizing interpretability and robustness is essential for building user trust and enhancing engagement, as systems must accurately capture dynamic user preferences through collaborative signals and semantic relatedness while integrating short-term interests with longer-term patterns. This ensures recommendations are relevant and reliable [69, 20]. Developing transparent and resilient models advances the field toward more sophisticated and user-centric recommendation solutions.

7 Conclusion

Incorporating Large Language Models (LLMs) and Natural Language Processing (NLP) into sequential recommendation systems marks a pivotal enhancement in personalization, interpretability, and overall system efficacy. The CA-RNN model exemplifies this by integrating contextual information into user behavior modeling, achieving superior performance over traditional methods. Similarly, the MLSA4Rec framework, which merges Mamba with self-attention mechanisms, underscores the potential of advanced computational techniques in sequential recommendations.

The LEADRE framework stands out as a groundbreaking industrial application, leveraging generative retrieval to significantly improve ad relevance and diversity in display advertising. This demonstrates the capacity of LLMs to elevate commercial applications through enhanced personalization and diversity. Furthermore, the DELRec model illustrates the effective fusion of conventional and LLM-based methodologies, resulting in more precise and context-aware recommendations.

Innovative frameworks such as GCL4SR, which utilize global context through transition graphs, have surpassed existing methods, while the LRURec model consistently outperforms state-of-the-art sequential recommenders, highlighting its real-world applicability. Additionally, the dual-tower retrieval architecture confirms the critical role of multimodal integration in boosting recommendation system performance across diverse datasets.

The advancements brought by LLMs and NLP are crucial in the evolution of sequential recommendation systems, fostering innovation and enhancing user experience. These technologies not only improve recommendation personalization and accuracy but also enhance model interpretability and robustness, paving the way for more sophisticated and user-centric solutions. Ongoing research and development in this dynamic area remain essential to unlock the full potential of LLMs and NLP in shaping the future of personalized recommendations.

www.SurveyX.cn

References

- [1] Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, and Philip S. Yu. Llmrec: Benchmarking large language models on recommendation task, 2023.
- [2] Kunzhe Song, Qingfeng Sun, Can Xu, Kai Zheng, and Yaming Yang. Self-supervised multi-modal sequential recommendation, 2023.
- [3] Wujiang Xu, Qitian Wu, Zujie Liang, Jiaojiao Han, Xuying Ning, Yunxiao Shi, Wenfang Lin, and Yongfeng Zhang. Slmrec: Distilling large language models into small for sequential recommendation, 2025.
- [4] Weiqi Shao, Xu Chen, Jiashu Zhao, Long Xia, and Dawei Yin. Gumble softmax for user behavior modeling, 2022.
- [5] Chonggang Song, Chunxu Shen, Hao Gu, Yaoming Wu, Lingling Yi, Jie Wen, and Chuan Chen. Precise: Pre-training sequential recommenders with collaborative and semantic information, 2024.
- [6] Lei Wang and Ee-Peng Lim. Zero-shot next-item recommendation using large pretrained language models, 2023.
- [7] Nuofan Xu and Chenhui Hu. Enhancing e-commerce recommendation using pre-trained language model and fine-tuning, 2023.
- [8] Shutong Qiao, Chen Gao, Yong Li, and Hongzhi Yin. Llm-assisted explicit and implicit multi-interest learning framework for sequential recommendation, 2024.
- [9] Shenghao Yang, Weizhi Ma, Peijie Sun, Qingyao Ai, Yiqun Liu, Mingchen Cai, and Min Zhang. Sequential recommendation with latent relations based on large language model, 2024.
- [10] Naicheng Guo, Hongwei Cheng, Qianqiao Liang, Linxun Chen, and Bing Han. Integrating large language models with graphical session-based recommendation, 2024.
- [11] Qidong Liu, Xian Wu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Feng Tian, and Yefeng Zheng. Large language models enhanced sequential recommendation for long-tail user and item, 2024.
- [12] Ruining He and Julian McAuley. Fusing similarity models with markov chains for sparse sequential recommendation, 2016.
- [13] Hanbing Wang, Xiaorui Liu, Wenqi Fan, Xiangyu Zhao, Venkataramana Kini, Devendra Yadav, Fei Wang, Zhen Wen, Jiliang Tang, and Hui Liu. Rethinking large language model architectures for sequential recommendations, 2024.
- [14] Hao Wang, Yongqiang Han, Kefan Wang, Kai Cheng, Zhen Wang, Wei Guo, Yong Liu, Defu Lian, and Enhong Chen. Denoising pre-training and customized prompt learning for efficient multi-behavior sequential recommendation, 2024.
- [15] Kibum Kim, Dongmin Hyun, Sukwon Yun, and Chanyoung Park. Melt: Mutual enhancement of long-tailed user and item for sequential recommendation, 2023.
- [16] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. Sequential recommendation with graph neural networks, 2023.
- [17] Tingjia Shen, Hao Wang, Jiaqing Zhang, Sirui Zhao, Liangyue Li, Zulong Chen, Defu Lian, and Enhong Chen. Exploring user retrieval integration towards large language models for cross-domain sequential recommendation, 2024.
- [18] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt predict paradigm (p5), 2023.

-
- [19] Kostadin Cvejovski, Ramses J. Sanchez, Bogdan Georgiev, Jannis Schuecker, Christian Bauckhage, and Cesar Ojeda. Recurrent point processes for dynamic review models, 2020.
- [20] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *ACM computing surveys (CSUR)*, 51(4):1–36, 2018.
- [21] Sami Khenissi, Mariem Boujelbene, and Olfa Nasraoui. Theoretical modeling of the iterative properties of user discovery in a collaborative filtering recommender system, 2020.
- [22] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations, 2020.
- [23] Bo Peng, Ben Burns, Ziqi Chen, Srinivasan Parthasarathy, and Xia Ning. Towards efficient and effective adaptation of large language models for sequential recommendation, 2023.
- [24] Weiqi Shao, Xu Chen, Jiashu Zhao, Long Xia, and Dawei Yin. User behavior understanding in real world settings, 2022.
- [25] Muiyang Li, Zijian Zhang, Xiangyu Zhao, Wanyu Wang, Minghao Zhao, Runze Wu, and Ruocheng Guo. Automlp: Automated mlp for sequential recommendations, 2023.
- [26] Jiahao Tian, Jinman Zhao, Zhenkai Wang, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system, 2024.
- [27] Jiajie Su, Chaochao Chen, Zibin Lin, Xi Li, Weiming Liu, and Xiaolin Zheng. Personalized behavior-aware transformer for multi-behavior sequential recommendation, 2024.
- [28] Shereen Elsayed, Ahmed Rashed, and Lars Schmidt-Thieme. Context-aware sequential model for multi-behaviour recommendation, 2023.
- [29] Panfeng Cao and Pietro Lio. Genrec: Generative sequential recommendation with large language models, 2024.
- [30] Chuan He, Yongchao Liu, Qiang Li, Weiqiang Wang, Xin Fu, Xinyi Fu, Chuntao Hong, and Xinwei Yao. Multi-grained preference enhanced transformer for multi-behavior sequential recommendation, 2024.
- [31] Zhenrui Yue, Yueqi Wang, Zhankui He, Huimin Zeng, Julian McAuley, and Dong Wang. Linear recurrent units for sequential recommendation, 2023.
- [32] Jerome Ramos, Bin Wu, and Aldo Lipani. Peapod: Personalized prompt distillation for generative recommendation, 2025.
- [33] Yaoyiran Li, Xiang Zhai, Moustafa Alzantot, Keyi Yu, Ivan Vulić, Anna Korhonen, and Mohamed Hammad. Calrec: Contrastive alignment of generative llms for sequential recommendation, 2024.
- [34] Jun Wang, Likang Wu, Qi Liu, and Yu Yang. An efficient continuous control perspective for reinforcement-learning-based sequential recommendation, 2024.
- [35] Marleson R. O. Santana and Anderson Soares. Hybrid model with time modeling for sequential recommender systems, 2021.
- [36] Sirui Wang, Peiguang Li, Yunsen Xian, and Hongzhi Zhang. Beyond the sequence: Statistics-driven pre-training for stabilizing sequential recommendation model, 2024.
- [37] Mehdi Soleiman Nejad, Meysam Varasteh, Hadi Moradi, and Mohammad Amin Sadeghi. Designing a sequential recommendation system for heterogeneous interactions using transformers, 2022.
- [38] Siamak Farshidi, Kiyan Rezaee, Sara Mazaheri, Amir Hossein Rahimi, Ali Dadashzadeh, Morteza Ziabakhsh, Sadegh Eskandari, and Slinger Jansen. Understanding user intent modeling for conversational recommender systems: A systematic literature review, 2023.

-
- [39] Biswarup Bhattacharya, Iftikhar Burhanuddin, Abhilasha Sancheti, and Kushal Satya. Intent-aware contextual recommendation system, 2017.
- [40] Ziwei Liu, Qidong Liu, Yejing Wang, Wanyu Wang, Pengyue Jia, Maolin Wang, Zitao Liu, Yi Chang, and Xiangyu Zhao. Sigma: Selective gated mamba for sequential recommendation, 2024.
- [41] Xiangyu Zhao, Liang Zhang, Long Xia, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for list-wise recommendations, 2019.
- [42] Weiqi Shao, Xu Chen, Long Xia, Jiashu Zhao, and Dawei Yin. Sequential recommendation with user evolving preference decomposition, 2022.
- [43] Meysam Varasteh, Mehdi Soleiman Nejad, Hadi Moradi, Mohammad Amin Sadeghi, and Ahmad Kalhor. An improved hybrid recommender system: Integrating document context-based and behavior-based methods, 2021.
- [44] Mohammadmahdi Maheri, Reza Abdollahzadeh, Bardia Mohammadi, Mina Rafiei, Jafar Habibi, and Hamid R. Rabiee. Clusterseq: Enhancing sequential recommender systems with clustering based meta-learning, 2023.
- [45] Huixuan Chi, Hao Xu, Hao Fu, Mengya Liu, Mengdi Zhang, Yuji Yang, Qinfen Hao, and Wei Wu. Long short-term preference modeling for continuous-time sequential recommendation, 2022.
- [46] Jiafeng Xia, Dongsheng Li, Hansu Gu, Tun Lu, Peng Zhang, Li Shang, and Ning Gu. Oracle-guided dynamic user preference modeling for sequential recommendation, 2024.
- [47] Xiaowen Huang, Jitao Sang, Jian Yu, and Changsheng Xu. Learning to learn a cold-start sequential recommender, 2021.
- [48] Ke Sun and Tieyun Qian. Seq2seq translation model for sequential recommendation, 2020.
- [49] Dongmin Hyun, Chanyoung Park, Junsu Cho, and Hwanjo Yu. Beyond learning from next item: Sequential recommendation via personalized interest sustainability, 2022.
- [50] Liwei Pan, Weike Pan, Meiyan Wei, Hongzhi Yin, and Zhong Ming. A survey on sequential recommendation, 2024.
- [51] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Is news recommendation a sequential recommendation task?, 2021.
- [52] Krishna Sayana, Raghavendra Vasudeva, Yuri Vasilevski, Kun Su, Liam Hebert, James Pine, Hubert Pham, Ambarish Jash, and Sukhdeep Sodhi. Beyond retrieval: Generating narratives in conversational recommender systems, 2024.
- [53] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. Automatic generation of natural language explanations, 2017.
- [54] Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling, 2024.
- [55] Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation, 2023.
- [56] Xinyu Zhang, Linmei Hu, Luhao Zhang, Dandan Song, Heyan Huang, and Liqiang Nie. Laser: Parameter-efficient llm bi-tuning for sequential recommendation with collaborative information, 2024.
- [57] Junzhe Jiang, Shang Qu, Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, Kai Zhang, Rui Li, Jiatong Li, and Min Gao. Reformulating sequential recommendation: Learning dynamic user interest with content-enriched language modeling, 2024.

-
- [58] Muyang Li, Xiangyu Zhao, Chuan Lyu, Minghao Zhao, Runze Wu, and Ruocheng Guo. Mlp4rec: A pure mlp architecture for sequential recommendations, 2022.
- [59] Angela John, Theophilus Aidoo, Hamayoon Behmanush, Irem B. Gunduz, Hewan Shrestha, Maxx Richard Rahman, and Wolfgang Maaß. Llmrs: Unlocking potentials of llm-based recommender systems for software purchase, 2024.
- [60] Mingze Wang, Shuxian Bi, Wenjie Wang, Chongming Gao, Yangyang Li, and Fuli Feng. Incorporate llms with influential recommender system, 2024.
- [61] Zhengbang Zhu, Rongjun Qin, Junjie Huang, Xinyi Dai, Yang Yu, Yong Yu, and Weinan Zhang. Understanding or manipulation: Rethinking online performance gains of modern recommender systems, 2023.
- [62] Lixi Zhu, Xiaowen Huang, and Jitao Sang. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems, 2024.
- [63] Elisabeth Fischer, Albin Zehe, Andreas Hotho, and Daniel Schlör. Modeling and analyzing the influence of non-item pages on sequential next-item prediction, 2025.
- [64] Yueming Sun and Yi Zhang. Conversational recommender system, 2018.
- [65] Fabian Paischer, Liu Yang, Linfeng Liu, Shuai Shao, Kaveh Hassani, Jiacheng Li, Ricky Chen, Zhang Gabriel Li, Xialo Gao, Wei Shao, Xue Feng, Nima Noorshams, Sem Park, Bo Long, and Hamid Eghbalzadeh. Preference discerning with llm-enhanced generative retrieval, 2024.
- [66] Xiaoqing Chen, Zhitao Li, Weike Pan, and Zhong Ming. A survey on multi-behavior sequential recommendation, 2023.
- [67] Zekai Qu, Ruobing Xie, Chaojun Xiao, Yuan Yao, Zhiyuan Liu, Fengzong Lian, Zhanhui Kang, and Jie Zhou. Thoroughly modeling multi-domain pre-trained recommendation as language, 2023.
- [68] Tianhui Ma, Yuan Cheng, Hengshu Zhu, and Hui Xiong. Large language models are not stable recommender systems, 2023.
- [69] Mingyue Cheng, Hao Zhang, Qi Liu, Fajie Yuan, Zhi Li, Zhenya Huang, Enhong Chen, Jun Zhou, and Longfei Li. Empowering sequential recommendation from collaborative signals and semantic relatedness, 2024.
- [70] Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system, 2024.
- [71] Chengkai Liu, Jianghao Lin, Hanzhou Liu, Jianling Wang, and James Caverlee. Behavior-dependent linear recurrent units for efficient sequential recommendation, 2024.
- [72] Hyunsoo Kim, Junyoung Kim, Minjin Choi, Sunkyung Lee, and Jongwuk Lee. Mars: Matching attribute-aware representations for text-based sequential recommendation, 2024.
- [73] Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. Dynamic graph neural networks for sequential recommendation, 2021.
- [74] Zhenghao Liu, Sen Mei, Chenyan Xiong, Xiaohua Li, Shi Yu, Zhiyuan Liu, Yu Gu, and Ge Yu. Text matching improves sequential recommendation by reducing popularity biases, 2023.
- [75] Aleksandr Petrov and Craig Macdonald. Effective and efficient training for sequential recommendation using recency sampling, 2022.
- [76] Xin Chen, Alex Reibman, and Sanjay Arora. Sequential recommendation model for next purchase prediction, 2023.
- [77] Annie Marsden and Sergio Bacallado. Sequential matrix completion, 2017.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

www.SurveyX.cn