GR-LLMs: Recent Advances in Generative Recommendation Based on Large Language Models

Zhen Yang **Haitao Lin** **Jiawei Xue** **Ziji Zhang**
AMAP, Alibaba Group
{zhongming.yz,lht416932,xuejiawei.xjw,zhangziji.zzj}@alibaba-inc.com

Abstract

In the past year, Generative Recommendations (GRs) have undergone substantial advancements, especially in leveraging the powerful sequence modeling and reasoning capabilities of Large Language Models (LLMs) to enhance overall recommendation performance. LLM-based GRs are forming a new paradigm that is distinctly different from discriminative recommendations, showing strong potential to replace traditional recommendation systems that are heavily dependent on complex, hand-crafted features. In this paper, we provide a comprehensive survey designed to facilitate further research on LLM-based GRs. Initially, we outline the general preliminaries and application cases of LLM-based GRs. Subsequently, we introduce the main considerations during the industrial applications of GRs. Finally, we explore promising directions for LLM-based GRs. We hope that this survey contributes to the ongoing advancement of the GR domain.

1 Introduction

Recommendation systems (Adomavicius and Tuzhilin, 2005; Li et al., 2024b), which aim to recommend the items (e.g., e-commerce products, micro-videos, news, and point-of-interests) by implicitly inferring user interest from the user's profile and historical interactions, are ubiquitous in the modern digital landscape, serving as critical interfaces for navigating the vast sea of information and choices available online. The effectiveness of recommendation systems has been a driving force behind the success of numerous online platforms, from e-commerce giants and social networks to content streaming services and news aggregators.

With the advancement of recommendation systems, modeling algorithms have roughly undergone three different technological paradigms, namely **machine learning-based recommendation (MLR)**, **deep learning-based recommendation (DLR)**, and **generative recommendation (GR)**. The **MLR** primarily relies on traditional machine learning algorithms, often built upon explicit feature engineering. Key techniques include collaborative filtering (Breese et al., 2013; He et al., 2017; Sarwar et al., 2001; Linden et al., 2003), which predicts user preferences based on similarity with other users or items, and content-based filtering, which recommends items similar to those a user has liked based on item attributes. Matrix factorization techniques (Koren et al., 2009; Rendle et al., 2012), such as singular value decomposition or alternating least squares, are also central to this era, aiming to learn latent factors that represent user and item preferences to predict missing entries in a user-item interaction matrix. While **MLR** focuses on capturing statistical patterns in historical data, it frequently faces challenges in addressing data sparsity and the cold-start problem (where new users or items lack sufficient interaction data). Feature engineering remains essential for providing meaningful input features. The **DLR** leverages the power of deep neural networks to automatically learn complex, non-linear representations directly from raw or sparse features (Tang and Wang, 2018; Xue et al., 2017; Zhang et al., 2017; Chen et al., 2019; Xue et al., 2025). In industrial recommendation systems, **DLR** has been used for nearly a decade, typically with inputs that include many well-designed handcrafted features to improve model performance. The primary challenge with **DLR** models lies in the trade-off between effectiveness and efficiency. Their relatively small parameter size often makes scaling more difficult, hindering convenient increases in model capacity to enhance recommendation quality.

Traditional recommendation paradigms, i.e., the **MLR** and **DLR**, focus on predicting a similarity or rank score based on hand-crafted feature engineering and intricate cascaded modeling structure, making them brittle, difficult to interpret, and requiring significant manual effort for maintenance and adaptation to new data or domains. In recent years, a paradigm shift has been catalyzed by the emergence and rapid development of large language models (LLMs). These models, exemplified by architectures like GPT, BERT, encoder-decoder Transformer, and others, trained on vast amounts of text data, demonstrate remarkable capabilities in understanding and generating human language (Floridi and Chiriatti, 2020; Liu et al., 2019; Guo et al., 2025a). They excel at capturing complex statistical dependencies in sequential data, performing sophisticated reasoning tasks, and exhibiting a deep understanding of context and semantics. This unprecedented power in sequence modeling and general-purpose reasoning has naturally led to the emergence and rapid advancement of **GRs**. In particular, the **GR** systems have seen great progress in the past year. SASRec (Kang and McAuley, 2018) first proposes to predict the next user-interacted item through an autoregressive approach based on a transformer model. Zhai et al. (2024) propose HSTU, a new transformer framework for better modeling sequences and inference efficiency, which is followed by Huang et al. (2025) on ranking tasks. To deal with a large number of items in industrial-scale recommendation, TIGER (Rajput et al., 2023) incorporates the idea of RQ-VAE (Zeghidour et al., 2021) to learn to transform items into multiple semantic IDs, largely reducing the vocabulary size. Building upon similar semantic ID encoding, OneRec (Deng et al., 2025) employs a Mixture of Experts (MoE) architecture (Dai et al., 2024) and a Direct Preference Optimization (DPO) strategy (Rafailov et al., 2023) to further improve the recommendation ability. For industrial-scale generative recommendation, MTGR (Han et al., 2025) proposes incorporating the cross features used in **DLR** and finds that excluding cross features severely damages the model's performance. Qiu et al. (2025) and Zheng et al. (2025) further propose an end-to-end generative architecture that unifies online advertising ranking as one model. Moreover, Jiang et al. (2025) points out that the LLM can be used as a universal recommendation learner, and they propose URM, which can perform well on versatile recommendation tasks.

Works mentioned above show that LLM-based GRs represent a fundamental departure from traditional discriminative methods. This generative aspect offers several compelling advantages. Firstly, it allows for greater explainability, enabling systems to communicate why a particular item is recommended, thereby building user trust and facilitating feedback loops. Secondly, it inherently supports creativity and novelty, as LLMs can suggest items beyond the most predictable ones based on past behavior, potentially helping users discover new interests. Thirdly, the unified language model approach potentially simplifies system design by reducing the need for complex, hand-crafted feature engineering and separate modules for different tasks. Furthermore, the scaling laws in LLMs have great potential to raise the performance ceiling of generative recommendation systems.

Recognizing the transformative potential and the rapid pace of development in this area, there is a growing need for a comprehensive survey that synthesizes the current knowledge on GRs. While initial explorations into this space have been conducted, the field is evolving rapidly, and a systematic survey is crucial to help researchers and practitioners navigate the landscape, understand the core concepts and techniques, learn from existing applications, and identify promising avenues for future work. In this paper, we provide a comprehensive survey aimed at facilitating further research and development in LLM-based GRs. We structure our survey to first outline the general preliminaries and foundational concepts of LLM-based GRs. Subsequently, we delve into the diverse application cases and real-world deployments of these systems. Finally, we critically analyze the main considerations and challenges encountered when applying LLM-based GRs in demanding industrial scenarios. We conclude by exploring promising future research directions. We hope that this survey contributes significantly to the ongoing advancement and maturation of the GR domain.

2 Preliminaries

2.1 Large language models

Large language models, which are trained on vast amounts of text data, have demonstrated significant capabilities in natural language processing (Bubeck et al., 2023; Yang et al., 2025a; Grattafiori et al., 2024). Given an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, LLMs are trained to optimize the probability $P(x_t|x_{<t};\theta)$ with the next-token prediction format, where θ represents the parameters of the model and $x_{<t}$ indicates the tokens before x_t. Initially, LLMs were primarily text-based, but they have evolved to handle multi-modal data, integrating text with images, audio, and video (Team

et al., 2023; Liu et al., 2023b,a; Yang et al., 2023). With the ability to support multi-modal inputs and outputs, large models can perform a variety of sequence generation tasks.

2.2 Traditional cascaded recommendations

Traditional recommendation systems widely adopt multi-stage cascaded architectures to balance computational efficiency and prediction accuracy (Burges, 2010; Chang et al., 2023; Wang et al., 2011). As illustrated in Figure 1, a typical cascaded recommendation system includes three sequential stages: recall, pre-ranking, and ranking. Although efficient in practice, existing methods typically treat each stage independently, where the effectiveness of each isolated stage serves as the upper bound for the subsequent stage, thereby limiting the performance of the overall recommendation system. Many previous works (Fei et al., 2021; Gallagher et al., 2019; Huang et al., 2023b; Wang et al., 2024b) have been proposed to enhance overall recommendation performance by enabling interaction among different stages, but they still maintain the traditional cascade paradigm. Recently, the GRs have emerged as a promising paradigm to serve as a unified architecture for end-to-end generation (Qiu et al., 2025; Deng et al., 2025).

3 Application Settings of GR

In the past year, various GR systems have achieved significant business benefits in practical industrial settings. There are two different branches of approaches applying GRs in online recommendation: the first branch is cooperating with the corresponding modules of traditional cascaded systems; the other is to apply generative models directly for end-to-end recommendations. This section will systematically summarize and analyze recent works based on their specific application settings.

3.1 Recall

Recall is a foundational step that narrows the candidate item pool to a subset potentially relevant to a user. This stage is critical for supporting the subsequent ranking stage and prioritizes efficient implementation over massive datasets (McAuley, 2022). LLMs can be leveraged in the recall phase in three approaches: prompt-based, token-based, and embedding-based recall methods.

Prompt-based methods generate recall results by providing user information to pre-trained LLMs through customized queries. Pre-trained LLMs, such as Qwen3 (Yang et al., 2025a), are equipped with extensive world knowledge, which enables the recall of items even when user information is implicit or sparse. Specifically, LLMTreeRec (Zhang et al., 2024b) constructs prompts to guide LLMs in summarizing user interests, inferring related item categories, and recalling specific items through item trees. Note that the parameters in LLMs remain frozen and are not fine-tuned in this method. Besides, SyNeg (Li et al., 2024a) utilizes LLMs to synthesize negative samples that are difficult to classify. These generated hard negative samples are then fused with the retrieved negatives for model fine-tuning.

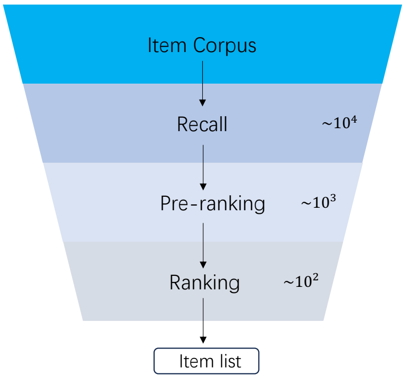

Figure 1: A typical cascade ranking system, which includes three stages from the top to the bottom: Recall, Pre-ranking, and Ranking.

Token-based methods map user behavior sequences to token sequences, formulating the recall problem as the next token prediction task. These methods offer strong flexibility in token sequence construction, model architecture design, and information encoding and decoding, supporting various industrial deployment scenarios. Notably, HSTU (Zhai et al., 2024) reformulates the recall task within a generative framework. Similarly, KuaiFormer (Liu et al., 2024) defines the recall process as next token prediction, incorporating hierarchical user behavior summarization across early, middle, and recent action sequences. The model has been deployed to industrial-scale platforms serving 400 million daily active users with a significant lift in both offline and online evaluation metrics. To unify diverse recall objectives (e.g., items users will click on), Universal Retrieval Model (URM) (Jiang et al., 2025) encodes these objectives as components within token sequences to provide objective-aware recall outcomes.

Embedding-based methods leverage LLMs as encoders to generate item embeddings, which are subsequently integrated into classical DLR methods. For example, MoRec employs pre-trained vision and text encoders to obtain item embeddings (Yuan et al., 2023) for downstream prediction modules such as DSSM (Huang et al., 2013) and SASRec (Kang and McAuley, 2018).

3.2 Rank

The ranking stage plays a pivotal role in determining the final relevance and diversity of candidate items presented to users. Compared to DLR, LLMs can accurately model user preferences mainly through chronological user behavior sequences, which eliminates the need for extensive feature engineering. Furthermore, LLMs possess a theoretical foundation in scaling laws, which allows LLM-based ranking systems to overcome the performance bottlenecks inherent in DLR approaches by leveraging increased model scale (Zhai et al., 2024).

Current approaches to integrating LLMs into the ranking stage can be categorized into two paradigms: generative architectures and hybrid integration architectures, depending on whether they retain the traditional DLR framework. Generative architectures abandon DLR frameworks, employing LLMs to directly process user behavior sequences and generate candidate scores through explicitly supervised tasks such as CTR prediction. As mentioned above, GR (Zhai et al., 2024) constructs chronological sequences by integrating user-related behaviors and features, reformulating the ranking task as a sequential transduction task. Notably, GR marks the first observation of the scaling laws inherent in LLMs within large-scale recommendation systems. GenRank (Huang et al., 2025) proposes an action-oriented sequence organization that treats items as positional context and focuses on predicting user actions associated with each candidate item. DFGR (Guo et al., 2025b) introduces a dual-flow generative architecture that decouples user behavior sequences into parallel real and fake flows to address computational inefficiencies in GR. It merges real and fake action-type tokens to model heterogeneous user behaviors while maintaining end-to-end efficiency.

Hybrid integration architectures, in contrast, leverage LLMs to generate highly informative representations that are integrated as supplementary features to DLR, thereby enhancing the performance of existing systems. As the representative approach, LEARN (Jia et al., 2025) adapts a fixed LLM as the Content-Embedding Generation (CEG) module to preserve open-world knowledge while bridging the gap between general and collaborative domains via a twin-tower structure comprising a user tower and an item tower. LEARN deploys its LLM-derived representations as complementary features within existing ranking models in industrial scenarios, achieving substantial improvements in practical applications. HLLM (Chen et al., 2024a) employs a hierarchical LLM architecture that sequentially employs a trainable ITEM LLM and USER LLM to extract item-specific and user-specific representations, respectively. In online scenarios, HLLM integrates the high-level item and user representations through a late fusion approach. SRP4CTR (Han et al., 2024) enhances CTR prediction by integrating self-supervised sequential recommendation pre-training with CTR models through a fine-tuned architecture that employs FG-BERT for multi-attribute side information encoding and a uni cross-attention block to transfer knowledge from pre-trained sequences to item-specific predictions efficiently. To integrate the strengths of both paradigms, MTGR (Han et al., 2025) employs a generative architecture based on HSTU to model user-level data while retaining raw features, including cross features designed for DLR.

3.3 End-to-end recommendation

The goal of end-to-end recommendation is to directly output the recommendation results based on users' historical behaviors. A key distinction from the recall task lies in whether the output inherently possesses ranking capability and whether it can fully replace the conventional cascaded recommendation pipeline. The main advantage of end-to-end recommendation is that it could avoid error propagation and objective misalignment caused by pipeline-based methods.

Most recall-focused methods (e.g., TIGER (Rajput et al., 2023), COBRA(Yang et al., 2025b) and URM (Jiang et al., 2025)) are trained with the objective of next item prediction, emphasizing top-k hit rate during evaluation. In contrast, end-to-end approaches often enhance the model's ranking ability through additional post-training stages built upon generative models. OneRec (Deng et al., 2025) first deploys the end-to-end recommendation model to the industrial scenario, targeting the generation of session-wise recommendation lists as its core task. It employs an iterative reward-model-based DPO alignment strategy to progressively refine the model. It completely replaces the cascaded online recommendation pipeline with a single model, achieving significant improvements on online feedback metrics. OneSug (Guo et al., 2025c) employs list-wise preference alignment, leveraging online feedback signals for reinforcement training to generate more preferable query suggestions. EGA-V2 (Zheng et al., 2025) is specifically optimized for advertising recommendation scenarios, where it trains a reward model on top of next item prediction to evaluate the expected cumulative reward of an ad sequence, thereby improving the overall expected return of recommended ads. Although EGA-V1 (Qiu et al., 2025) is primarily designed as a ranking model, it considers the full set of candidate advertisements as input, which places it within the scope of end-to-end recommendation.

Beyond industrial applications, several academic studies have focused on how to learn better ranking capabilities within the end-to-end framework. S-DPO (Chen et al., 2024b) is the first work to apply the DPO algorithm to recommendation models. By considering multiple negative samples exposed at the same time, S-DPO extends the binary preference optimization of DPO into a multi-preference setting, thereby improving the ranking quality of recommendations. RosePO (Liao et al., 2024) further enhances the construction of negative samples by incorporating factors such as popularity and semantic similarity, generating harder negatives for more effective reinforcement learning. SPRec (Gao et al., 2025) introduces a self-replay mechanism, using the model's predicted results as negative samples, which increases the difficulty of distinguishing between positive and negative examples and leads to stronger generalization. Combining these reinforcement learning strategies with industrial applications represents a promising direction for the development of end-to-end recommendations.

4 Main Considerations and Challenges

As a new paradigm in recommendations, GR faces many challenges in industrial applications; this section will detail the key issues that need to be considered during the application process.

4.1 Training pipelines

A core issue in GR is how to design the training methodology and objectives to align with the recommendation task. Based on the number of stages involved in the training process, we categorize existing approaches into two main types: single-stage training and multi-stage training. Table 1 summarizes these methods by category, training objective, and distinctive strategies.

4.1.1 Single-stage training

In single-stage training, the model produces the final recommendation output through a single training phase, typically focusing on only one specific task—either recall or ranking. For recall tasks, a common training objective is next item prediction. Methods such as HSTU (Zhai et al., 2024) and URM (Jiang et al., 2025) train models based on this objective, aiming to predict the top-k items as the recall results. Given the large scale of item spaces, these approaches often employ negative sampling during training to reduce the computational cost of the softmax operation. In contrast, KuaiFormer (Liu et al., 2024) adopts in-batch softmax, generating embeddings directly for recall purposes. For ranking tasks, a typical training objective is CTR prediction. MTGR (Han et al., 2025) explicitly models the relationship between user sequences and candidate items, predicting CTR values to replace traditional ranking models, thereby improving the recommendation performance.

4.1.2 Multi-stage training

Multi-stage training generally involves two phases: pre-training and fine-tuning. By defining distinct training objectives at each stage, the model can learn different capabilities. This category can be further divided into two subcategories, depending on how the pretrained models are utilized during fine-tuning: representation-based fine-tuning and model-based fine-tuning.

Representation-based finetuning Methods in this category (e.g., HLLM (Chen et al., 2024a), LEARN (Jia et al., 2025), and LUM (Yan et al., 2025)) primarily target ranking tasks. During the pre-training phase, they use contrastive learning and InfoNCE loss (Oord et al., 2018) to generate user and item embeddings. These embeddings are then used as features during the fine-tuning phase, where a traditional DLR model is trained to enhance the ranking performance. QARM (Luo et al., 2024) further quantizes the embedding into semantic IDs for downstream training, making these information learnable. Notably, LUM employs a three-stage pipeline: the first stage focuses solely

Table 1: A summary of learning targets and post-training strategies of recent GR works.

	Methods	Training Objective	Training Strategy & Loss
Single Stage	HSTU (Zhai et al., 2024)	next item / action prediction	cross entropy
	URM (Jiang et al., 2025)	next item prediction	cross entropy
	KuaiFormer (Liu et al., 2024)	user-item embedding similarity	in-batch contrastive learning
	MTGR (Han et al., 2025)	ctr prediction	binary classification
Multi Stage	OneRec (Deng et al., 2025)	next item prediction	DPO loss, iterative alignment
	OneSug (Guo et al., 2025c)	next item prediction	DPO loss, list-wise preference alignment
	EGA-V2 (Zheng et al., 2025)	next item prediction	auction-based preference alignment
	EGA-V1 (Qiu et al., 2025)	ctr prediction	auction-based preference alignment
	QARM (Luo et al., 2024)	item alignment	contrastive learning, id as ranking feature
	HLLM (Chen et al., 2024a)	ctr prediction	contrastive learning, embedding as ranking feature
	LEARN (Jia et al., 2025)	ctr prediction	contrastive learning, embedding as ranking feature
	LUM (Yan et al., 2025)	next item prediction	contrastive learning, embedding as ranking feature

on next item prediction; the second stage learns to generate user and item embeddings through contrastive learning; the third stage trains DLMs for recall and ranking by using user and item embeddings as input features.

Model-based finetuning This category largely falls under the paradigm of end-to-end recommendation, where the pre-training phase learns the ability to predict next items and the finetuning phase then enhances the model's ranking capability in specific application scenarios using reinforcement learning. Both OneRec (Deng et al., 2025) and OneSug (Guo et al., 2025c) follow this framework, applied to video recommendation and query suggestion, respectively. EGA-V2 (Zheng et al., 2025) and EGA-V1 (Qiu et al., 2025) are specifically designed for advertising scenarios, achieving notable improvements in end-to-end ad recommendation and ad ranking.

4.2 Inference efficiency

The increased inference latency accompanied by the complex architectures of LLMs presents another challenge that hinders the deployment of GR models in real-world industrial scenarios. Currently, substantial efforts have been dedicated to optimizing decoding speed, primarily focusing on sequence compression, model architecture optimization, and specialized modeling and decoding tricks tailored for the recall and ranking stages.

Compressing the sequence length serves as an effective strategy to fundamentally reduce computational costs. GenRank (Huang et al., 2025) introduces an action-oriented sequence organization framework, which treats items as positional context, halving the input sequence length. DFGR (Guo et al., 2025b) reduces input sequence length by merging user-item interactions into single tokens through a real-flow and fake-flow. KuaiFormer (Liu et al., 2024) employs an adaptive item compression mechanism to reduce input sequence length by grouping earlier user interactions into coarsely aggregated representations while retaining fine-grained modeling of recent items, thereby decreasing sequence length without sacrificing recall performance.

Some work improves decoding efficiency by making slight adjustments to the model architecture. HSTU (Zhai et al., 2024) replaces the standard softmax in attention computation with a pointwise aggregated attention mechanism, reducing the quadratic complexity of self-attention to linear complexity. EGA-V1 (Qiu et al., 2025) proposes RecFormer, which introduces a cluster-attention mechanism in the Global Cluster-Former module to replace standard self-attention, reducing computational complexity by dynamically grouping keys/values into semantically coherent clusters via a learnable cluster matrix.

Some tailored tricks are proposed to accelerate the inference speed during the recall and rank stages. To reduce the vocabulary size during recall, some approaches (Rajput et al., 2023; Yang et al., 2024a; Yin et al., 2024) attempt to replace item IDs with the hierarchical semantic IDs. These semantic IDs are generated via RQ-VAE, which encodes item content into semantically meaningful token sequences, enabling efficient knowledge sharing across similar items by leveraging hierarchical semantic structure for autoregressive item prediction. RPG (Hou et al., 2025) generates long, unordered semantic IDs in parallel using optimized product quantization, trains with MTP loss to integrate sub-item semantics. To further enhance the inference efficiency of semantic ID-based methods, EGA-V2 (Zheng et al., 2025) employs Multi-Token Prediction (MTP) inference to enhance scalability and alignment with business objectives. Additionally, URM (Jiang et al., 2025) introduces matrix decomposition and probabilistic sampling instead of TopK selection to reduce computational complexity. The key to enhancing inference efficiency in the ranking stage lies in how to efficiently score all candidate items. Zhai et al. (2024) introduce M-FALCON that processes multiple candidate items in a single forward pass by modifying causal attention masks to ensure mutual invisibility among candidates. Similar to M-FALCON, MTGR (Han et al., 2025) introduces a customized masking strategy to prevent information leakage while enabling efficient candidate scoring.

4.3 Cold start and world knowledge

The cold start problem refers to the challenge of generating accurate recommendation outcomes when there is insufficient data, particularly for newly registered users and newly uploaded items (Lam et al., 2008; Wei et al., 2020). LLMs offer two primary strategies for mitigating the cold start problem in recommendation systems. (1) Information Augmentation. The goal is to enhance the input data used in recommendations by incorporating new embeddings and knowledge generated by LLMs. For example, SAID (Hu et al., 2024), proposed by Ant Group, generates item embeddings based on textual information and integrates them into downstream recommendation tasks. Analogously, CSRec (Yang et al., 2024b) fuses metadata-based and common sense-based knowledge derived from LLMs as side information to enhance recommendations. (2) Model Reasoning. The central idea is that LLMs can directly produce recommendation results by leveraging patterns learned from large-scale training data. A representative approach is LLM-Rec (Lyu et al., 2024), which employs carefully designed prompt strategies to derive effective recommendation solutions.

For the above information augmentation and model reasoning approaches, the underlying rationale for the positive impact of LLMs can be attributed to *world knowledge*. Here, world knowledge refers to the extensive contextual and conceptual knowledge inherent in LLMs, which stems from their training on large-scale datasets spanning diverse domains (Zhang et al., 2023). For instance, Llama 3 (Grattafiori et al., 2024) and Qwen3 (Yang et al., 2025a) were pre-trained using 15 and 36 trillion multilingual tokens, respectively, with domain diversity. The world knowledge embedded in pre-trained LLMs allows recommendation systems to effectively learn user-item interaction patterns during the cold start stage. Specifically, LC-Rec (Zheng et al., 2024) integrates language semantics from Llama with collaborative signals to attain world knowledge and task-specific characteristics in recommendation systems. Notably, recent research has revealed that item representations linearly mapped from language representations in LLMs enhance recommendation performance, demonstrating the value of world knowledge in recommendation systems (Sheng et al., 2024). Furthermore, the world knowledge within LLMs can be refined by adding external domain-specific knowledge using the Retrieval-Augmented Generation technique (Arslan et al., 2024).

The world knowledge in LLMs for recommendation can be obtained from various data sources on users and items: images (Radford et al., 2021), videos (Covington et al., 2016), and speech (Cui et al., 2024). The integration of diverse data sources can be referred to as *multi-modal learning*. To better leverage the multi-modal representations, a potential solution is contrastive learning, such as image-text contrastive loss (Li et al., 2021), which is used to align image and text representations before representation fusion.

Table 2: Multi-modal LLMs for recommendation.

	Data	Target
InteraRec (Karra and Tulabandhula, 2024)	Image, text	Product
I-LLMRec (Kim et al., 2025)	Image, text	Product
NoteLLM-2 (Zhang et al., 2024a)	Image, text	Note
TALKPLAY (Doh et al., 2025)	Audio, text	Music

The increasing feasibility of multi-modal learning in recommendation systems has been driven by the prosperity of multi-modal LLMs such as CLIP (Radford et al., 2021), vision transformer (Kim et al., 2021), and Qwen2-vl (Wang et al., 2024a). Typical multi-modal LLMs for recommendation are summarized in Table 2. Specifically, NoteLLM-2 (Zhang et al., 2024a), developed by Xiaohongshu, utilizes visual information within LLMs to recommend notes to users, resulting in a 6.4% increase in note exposures. Alternatively, TALKPLAY (Doh et al., 2025) is a multi-modal music recommendation system that encodes audio features, including lyrics and semantic tags, into LLMs to provide music recommendations. TALKPLAY demonstrates superior performance on the Million Playlist Dataset (Chen et al., 2018), which contains cold-start items in its test sets. Furthermore, InteraRec (Karra and Tulabandhula, 2024) is an online e-commerce product recommendation method that extracts valuable information from high-frequency web page screenshots. Together, given the standard representation learning frameworks in LLMs, these external multi-modal signals can be easily incorporated into existing LLM-based GRs to handle the cold start problem.

5 Future Directions

In this section, we explore promising directions for LLM-based GRs across the following aspects.

5.1 Model scaling

Since its observation and proposal, the scaling law has become the theoretical foundation for parameter scaling in large language models. When it comes to scaling, the traditional DLR has two significant drawbacks: 1) with the scaling of the length of the user behavior sequence, the DLR cannot efficiently process entire user behaviors, which limits the model's performance; 2) scaling incurs approximately linear costs in training and inference with the number of candidates, making the expenses unbearably high (Chen et al., 2021; Pi et al., 2020; Han et al., 2025). For GRs, recent studies have observed some scaling effects (Han et al., 2025; Wang et al., 2025; Huang et al., 2025). However, in these works, the model sizes are still limited to a relatively small level, such as 0.x B or 1.x B, and the performance improvements of models at much larger sizes have not been well validated. With a larger model size and longer user behavior sequence, it is an auspicious and challenging direction to train a more powerful generative recommendation model. Considering the requirements of line applications, it is also crucial to explore ways of efficient inference.

5.2 Data cleaning

As we all know, the quality of training data has a significant impact on the final performance of large language models. Few works in GRs have investigated how to perform data cleaning in the recommendation domain. Unlike traditional linguistic text corpora used in textual LLMs, training data in recommendation systems comprises not only item IDs but also multi-source side information with multi-modal characteristics. How to handle this heterogeneous side info is still an open question. The training corpora in GRs consist of user behavior sequences, posing unique challenges for quality assessment, as there exists no equivalent of grammaticality evaluation in natural language processing to discern the validity of behavioral sequences. Developing frameworks to evaluate behavioral sequence validity, implement quality-aware data curation through discriminative filtering, and establish dynamic training protocols conditioned on corpus quality represents a principled methodology for substantially improving recommendation performance (Huang et al., 2024).

5.3 One model for all

The core aspiration of LLMs is to achieve a universal architecture capable of accomplishing all diverse language tasks through prompt switching with a single model; recently, remarkable advancements in multimodal large models have further ignited researchers' enthusiasm for developing unified frameworks that support multiple modalities (Girdhar et al., 2023; Huang et al., 2023a; Yu et al., 2023; Yang et al., 2023; Zheng et al., 2024). Zhai et al. (2024) and Deng et al. (2025) unify the recall and rank in one GR model. Recently, Jiang et al. (2025) stepped further and proposed that GRs (URM in their work) can function as universal recommendation learners, capable of handling multiple tasks within a unified input-output framework, eliminating the need for specialized model designs. URM can handle multi-scenario recommendation (search included), multi-objective recommendation, long-tail item recommendation, etc. We posit that unifying the input and output, recommendation and search, through generative large models, which deliver customized recommendations by dynamically interpreting user instructions, will emerge as a promising research frontier in next-generation information retrieval.

6 Conclusions

In this paper, we have presented a comprehensive survey of LLM-based GRs with a focus on recent advancements. Initially, we outline the general preliminaries and application cases of LLM-based GRs. Subsequently, we introduce the main considerations when LLM-based GRs are applied in real industrial recommendation systems. Our survey also sheds light on their capabilities across diverse scenarios and promising future directions in this rapidly evolving field. We hope this survey can provide insights for researchers and contribute to the ongoing advancements in the GR domain.

Limitations

In this paper, we embark on a comprehensive exploration of the current LLM-based GRs landscape, presenting a synthesis from diverse perspectives enriched by our insights. Acknowledging the dynamic nature of GRs, it is plausible that certain aspects may have eluded our scrutiny, and recent advances might not be entirely encapsulated. Given the constraints of the page limits, we are unable to delve into all technical details and have provided concise overviews of the core contributions of mainstream GRs.

References

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on RAG with LLMs. *Procedia Computer Science*, 246:3781–3790.

John S Breese, David Heckerman, and Carl Kadie. 2013. Empirical analysis of predictive algorithms for collaborative filtering. *ArXiv preprint arXiv:1301.7363*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv preprint arXiv:2303.12712*.

Christopher JC Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81.

Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and 1 others. 2023. TWIN: Two-stage interest network for lifelong user behavior modeling in CTR prediction at Kuaishou. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3785–3794.

Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 527–528.

Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. 2024a. HLLM: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *ArXiv preprint arXiv:2409.12740*.

Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. 2021. End-to-end user behavior retrieval in click-through rateprediction model. *ArXiv preprint arXiv:2108.04468*.

Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic explainable recommendation based on neural attentive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 53–60.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024b. On softmax direct preference optimization for recommendation. *ArXiv preprint arXiv:2406.09215*.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 191–198.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *ArXiv preprint arXiv:2410.03751*.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *ArXiv preprint arXiv:2401.06066*.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. OneRec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *ArXiv preprint arXiv:2502.18965*.

Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025. TALKPLAY: Multimodal music recommendation with large language models. *ArXiv preprint arXiv:2502.13713*.

Hongliang Fei, Jingyuan Zhang, Xingxuan Zhou, Junhao Zhao, Xinyang Qi, and Ping Li. 2021. GemNN: Gating-enhanced multi-task neural networks with feature interaction learning for CTR prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2166–2171.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J Shane Culpepper. 2019. Joint optimization of cascade ranking models. In *Proceedings of the twelfth ACM International Conference on Web Search and Data Mining*, pages 15–23.

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. 2025.

SPRec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 5075–5084.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *ArXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv preprint arXiv:2501.12948*.

Hao Guo, Erpeng Xue, Lei Huang, Shichao Wang, Xiaolei Wang, Lei Wang, Jinpeng Wang, and Sheng Chen. 2025b. Action is all you need: Dual-flow generative ranking network for recommendation. *ArXiv preprint arXiv:2505.16752*.

Xian Guo, Ben Chen, Siyuan Wang, Ying Yang, Chenyi Lei, Yuqing Ding, and Han Li. 2025c. OneSug: The unified end-to-end generative framework for e-commerce query suggestion. *ArXiv preprint arXiv:2506.06913*.

Ruidong Han, Qianzhong Li, He Jiang, Rui Li, Yurou Zhao, Xiang Li, and Wei Lin. 2024. Enhancing CTR prediction through sequential recommendation pretraining: Introducing the SRP4CTR framework. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3777–3781.

Ruidong Han, Bin Yin, Shangyu Chen, He Jiang, Fei Jiang, Xiang Li, Chi Ma, Mincong Huang, Xiaoguang Li, Chunzhen Jing, and 1 others. 2025. MTGR: Industrial-scale generative recommendation framework in meituan. *ArXiv preprint arXiv:2505.18654*.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182.

Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. 2025. Generating long semantic IDs in parallel for recommendation. *ArXiv preprint arXiv:2506.05781*.

Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via LLM-based semantic embedding learning. In *Companion Proceedings of the ACM Web Conference 2024*, pages 103–111.

Lei Huang, Weitao Li, Chenrui Zhang, Jinpeng Wang, Xianchun Yi, and Sheng Chen. 2024. EXIT: An explicit interest transfer framework for cross-domain recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4563–4570.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 2333–2338.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, and 1 others. 2023a. Language is not all you need: Aligning perception with language models. *ArXiv preprint arXiv:2302.14045*.

Xu Huang, Defu Lian, Jin Chen, Liu Zheng, Xing Xie, and Enhong Chen. 2023b. Cooperative retriever and ranker in deep recommenders. In *Proceedings of the ACM Web Conference 2023*, pages 1150–1161.

Yanhua Huang, Yuqi Chen, Xiong Cao, Rui Yang, Mingliang Qi, Yinghao Zhu, Qingchang Han, Yaowei Liu, Zhaoyu Liu, Xuefeng Yao, and 1 others. 2025. Towards large-scale generative ranking. *ArXiv preprint arXiv:2505.04180*.

Jian Jia, Yipei Wang, Yan Li, Honggang Chen, Xuehan Bai, Zhaocheng Liu, Jian Liang, Quan Chen, Han Li, Peng Jiang, and 1 others. 2025. LEARN: Knowledge adaptation from large language model to recommendation for practical industrial application. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11861–11869.

Junguang Jiang, Yanwen Huang, Bin Liu, Xiaoyu Kong, Ziru Xu, Han Zhu, Jian Xu, and Bo Zheng. 2025. Large language models are universal recommendation learners. *ArXiv preprint arXiv:2502.03041*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE.

Saketh Reddy Karra and Theja Tulabandhula. 2024. InteraRec: Interactive recommendations using multimodal large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 32–43. Springer.

Kibum Kim, Sein Kim, Hongseok Kang, Jiwan Kim, Heewoong Noh, Yeonjun In, Kanghoon Yoon, Jinoh Oh, and Chanyoung Park. 2025. Image is all you need: Towards efficient and effective large language model-based recommender systems. *ArXiv preprint arXiv:2503.06238*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, pages 208–211.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024a. SyNeg: LLM-driven synthetic hard-negatives for dense retrieval. *ArXiv preprint arXiv:2412.17250*.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024b. A survey of generative search and recommendation in the era of large language models. *ArXiv preprint arXiv:2404.16924*.

Jiayi Liao, Xiangnan He, Ruobing Xie, Jiancan Wu, Yancheng Yuan, Xingwu Sun, Zhanhui Kang, and Xiang Wang. 2024. RosePO: Aligning LLM-based recommenders with human values. *ArXiv preprint arXiv:2410.12519*.

Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Chi Liu, Jiangxia Cao, Rui Huang, Kai Zheng, Qiang Luo, Kun Gai, and Guorui Zhou. 2024. KuaiFormer: Transformer-based retrieval at Kuaishou. *ArXiv preprint arXiv:2411.10057*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *ArXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv preprint arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint arXiv:1907.11692*.

Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, and 1 others. 2024. QARM: Quantitative alignment multi-modal recommendation at Kuaishou. *ArXiv preprint arXiv:2411.11739*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

Julian McAuley. 2022. *Personalized Machine Learning*. Cambridge University Press.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint arXiv:1807.03748*.

Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2685–2692.

Junyan Qiu, Ze Wang, Fan Zhang, Zuowu Zheng, Jile Zhu, Jiangke Fan, Teng Zhang, Haitao Wang, and Xingxing Wang. 2025. One model to rank them all: Unifying online advertising with end-to-end learning. *ArXiv preprint arXiv:2505.19755*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *ArXiv preprint arXiv:1205.2618*.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the*

10th International Conference on World Wide Web, pages 285–295.

Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. 2024. Language representations can be what recommenders need: Findings and potentials. *ArXiv preprint arXiv:2407.05441*.

Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 565–573.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint arXiv:2312.11805*.

Chunqi Wang, Bingchao Wu, Zheng Chen, Lei Shen, Bing Wang, and Xiaoyi Zeng. 2025. Scaling transformers for discriminative recommendation via generative pretraining. *ArXiv preprint arXiv:2506.03699*.

Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv preprint arXiv:2409.12191*.

Yunli Wang, Zhiqiang Wang, Jian Yang, Shiyang Wen, Dongying Kong, Han Li, and Kun Gai. 2024b. Adaptive neural ranking framework: Toward maximized business goal for cascade ranking systems. In *Proceedings of the ACM Web Conference 2024*, pages 3798–3809.

Tianxin Wei, Ziwei Wu, Ruirui Li, Ziniu Hu, Fuli Feng, Xiangnan He, Yizhou Sun, and Wei Wang. 2020. Fast adaptation for cold-start collaborative filtering with meta-learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 661–670. IEEE.

Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia.

Jiawei Xue, Zhen Yang, Haitao Lin, Ziji Zhang, Luzhu Wang, Yikun Gu, Yao Xu, and Xin Li. 2025. HGCL: Hierarchical graph contrastive learning for user-item recommendation. *ArXiv preprint arXiv:2505.19020*.

Bencheng Yan, Shilei Liu, Zhiyuan Zeng, Zihao Wang, Yizhen Zhang, Yujin Yuan, Langming Liu, Jiaqi Liu, Di Wang, Wenbo Su, and 1 others. 2025. Unlocking scaling law in industrial recommendation systems with a three-step paradigm based large user model. *ArXiv preprint arXiv:2502.08309*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *ArXiv preprint arXiv:2505.09388*.

Liu Yang, Fabian Paischer, Kaveh Hassani, Jiacheng Li, Shuai Shao, Zhang Gabriel Li, Yun He, Xue Feng, Nima Noorshams, Sem Park, and 1 others. 2024a. Unifying generative and dense retrieval for sequential recommendation. *ArXiv preprint arXiv:2411.18814*.

Shenghao Yang, Weizhi Ma, Peijie Sun, Min Zhang, Qingyao Ai, Yiqun Liu, and Mingchen Cai. 2024b. Common sense enhanced knowledge-based recommendation with large language model. In *International Conference on Database Systems for Advanced Applications*, pages 381–390. Springer.

Yuhao Yang, Zhi Ji, Zhaopeng Li, Yi Li, Zhonglin Mo, Yue Ding, Kai Chen, Zijian Zhang, Jie Li, Shuanglong Li, and Lin Liu. 2025b. Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations. *Preprint*, arXiv:2503.02453.

Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. 2023. TEAL: Tokenize and embed all for multi-modal large language models. *ArXiv preprint arXiv:2311.04589*.

Jun Yin, Zhengxin Zeng, Mingzheng Li, Hao Yan, Chaozhuo Li, Weihao Han, Jianjin Zhang, Ruochen Liu, Allen Sun, Denvy Deng, and 1 others. 2024. Unleash LLMs potential for recommendation by coordinating twin-tower dynamic semantic token generator. *ArXiv preprint arXiv:2409.09253*.

Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, and 1 others. 2023. SPAE: Semantic pyramid autoencoder for multimodal generation with frozen llms. *ArXiv preprint arXiv:2306.17842*.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? ID-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2639–2649.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, and 1 others. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *ArXiv preprint arXiv:2402.17152*.

Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2024a. NoteLLM-2: Multimodal large representation models for recommendation. *ArXiv preprint arXiv:2405.16789*.

Wenlin Zhang, Chuhan Wu, Xiangyang Li, Yuhao Wang, Kuicai Dong, Yichao Wang, Xinyi Dai, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024b. LLMTreeRec: Unleashing the power of large language models for cold-start recommendations. *ArXiv preprint arXiv:2404.00702*.

Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1449–1458.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *ArXiv preprint arXiv:2310.07343*.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1435–1448. IEEE.

Zuowu Zheng, Ze Wang, Fan Yang, Jiangke Fan, Teng Zhang, and Xingxing Wang. 2025. Beyond cascaded architectures: An end-to-end generative framework for industrial advertising. *ArXiv e-prints*, pages arXiv–2505.