AI Open 7 (2026) 169–193

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Review

A survey on generative recommendation: Data, model, and tasks

Min Hou [a], Le Wu [a,*], Yuxin Liao [a], Yonghui Yang [b], Zhen Zhang [a], Yu Wang [a], Changlong Zheng [a], Han Wu [a], Richang Hong [a]

[a] Hefei University of Technology, Hefei, 230009, Anhui, China
[b] National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Recommender system
Generative model
Large language model

ABSTRACT

Recommender systems serve as a foundational infrastructure in modern information ecosystems, helping users navigate the expanding digital content space and discover items aligned with their preferences. At their core, recommender systems address a fundamental research problem: matching users with items. Over the past decades, the field has experienced successive technological paradigm shifts, from collaborative filtering and matrix factorization in the machine learning era to sophisticated neural architectures in the deep learning era. Recently, the emergence of generative models, especially large language models (LLMs) and diffusion models have sparked a new paradigm: generative recommendation, which reconceptualizes the recommendation problem as a generation task rather than a discriminative scoring procedure. This survey provides a comprehensive examination of this paradigm through a unified tripartite framework spanning data, model, and task dimensions. Rather than simply categorizing works, we systematically decompose approaches into operational stages—data augmentation and unification, model alignment and training, task formulation and execution. At the data level, generative models enable knowledge-infused augmentation and agent-based simulation while unifying heterogeneous signals. At the model level, we taxonomize LLM-based methods, large recommendation models, and diffusion approaches, analyzing their alignment mechanisms and innovations. At the task level, we illuminate new capabilities including conversational interaction, explainable reasoning, and personalized content generation. We identify five key advantages: world knowledge integration, natural language understanding, reasoning capabilities, scaling laws, and creative generation. We critically examine challenges in benchmark design, model robustness, and deployment efficiency, while charting a roadmap toward intelligent recommendation assistants that fundamentally reshape human-information interaction.

1. Introduction

Recommender systems (RSs) aim to recommend the items (e.g., E-commerce products, micro-videos, musics, and news) by inferring user interest from historical interactions, user profiles, and context information. RSs serve as an effective solution to alleviate information overload, to facilitate users seeking desired information, and to increase the traffic and revenue of service providers. Currently, RSs are extensively deployed across diverse domains, including e-commerce, social media, education, online video, and music services, establishing them as one of the most pervasive user-centered AI applications in modern information systems.

Over the past decade, RSs have made significant advancements, largely driven by improvements in deep learning architectures and the growing availability of large-scale user behavior data. Traditional RSs can be viewed as a discriminative matching paradigm, focusing on learning an effective matching function between users and items. In the 1990s, early research developed heuristics for content-based and collaborative filtering methods (Resnick et al., 1994). In the 2000s, Matrix Factorization (Koren et al., 2009) gained prominence, especially after the Netflix Prize, marking a key milestone in the development of RSs. By the mid-2010s, advancements in neural networks, such as CNNs, RNNs, GNNs, and Transformers, led to a transition toward deep learning-based recommendation methods. These approaches take advantage of deep learning's powerful representation capabilities to handle complex data, enabling non-linear mapping of user–item interactions and enhancing user/item representation learning by encoding intricate semantics (e.g., text, images, social networks, and knowledge graphs) (Wu et al., 2022). Despite these advancements, traditional RSs

* Corresponding author.
E-mail addresses: hmhoumin@gmail.com (M. Hou), lewu.ustc@gmail.com (L. Wu), yuxinliao314@gmail.com (Y. Liao), yyh.hfut@gmail.com (Y. Yang), zhangz.199911@gmail.com (Z. Zhang), wangyu20001162@gmail.com (Y. Wang), changlongzheng419@gmail.com (C. Zheng), ustcwuhan@gmail.com (H. Wu), hongrc.hfut@gmail.com (R. Hong).

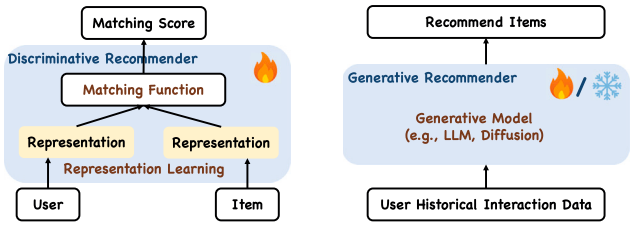
https://doi.org/10.1016/j.aiopen.2026.05.002
Received 1 November 2025; Received in revised form 5 March 2026; Accepted 6 May 2026
Available online 13 May 2026

still face several challenges. They remain reliant on limited semantic knowledge, often manually processed, and struggle with sub-optimal performance for small-scale models. Additionally, they depend on fixed candidate sets, require task-specific architectures and training objectives, encounter difficulties in cold-start scenarios, and struggle to provide transparent, context-rich explanations.

Recently, generative models have made significant advances in artificial intelligence. In particular, Large Language Models (LLMs) and diffusion models have brought about revolutionary changes in AI-generated content (AIGC). Leveraging vast pre-training corpora and enhanced computational power, LLMs demonstrate emergent abilities, including in-context learning, complex reasoning, and so on. Meanwhile, diffusion models have led to remarkable breakthroughs in photo-realistic image generation. In recent years, the emergence of generative models has brought about a paradigm shift in the research and development of RSs. Distinguished from previous discriminative matching paradigms, generative recommendation aims to directly generate the target document or item to satisfy users' information needs as shown in Fig. 1. Embracing generative recommendation brings new benefits and opportunities for recommendation. In particular: (1) LLMs inherently possess formidable capabilities, such as vast world knowledge, semantic understanding, interactive skills, and instruction following. These inherent abilities can be leveraged to augment sparse recommendation data, enrich item representations, simulate user behaviors, and unify heterogeneous information across domains and modalities, thereby addressing long-standing data sparsity and cold-start challenges. (2) The tremendous success of LLMs stems from their generative learning paradigm and scaling laws. Applying generative learning to recommendation fundamentally revolutionizes the modeling approaches — from discriminative scoring to generative synthesis — enabling powerful architectures that can capture complex user–item interactions and demonstrate emergent capabilities through increased scale. (3) LLMs-based generative AI applications, such as ChatGPT, are gradually becoming a new gateway for users to access Web content. Developing generative recommendation enables new task paradigms including conversational interaction, explainable reasoning, and personalized content generation, which can be seamlessly integrated into these generative AI applications. To advance this rapidly evolving field, it is imperative to undertake a comprehensive survey of existing studies, building a technique map to systematically guide the research, practice, and service in generative model-empowered recommendation.

In this survey, we provide a comprehensive review of how generative models reshape recommendation. This survey organizes existing studies into three complementary perspectives: **data-level**, **model-level**, and **task-level**. This taxonomy reflects both the comprehensive influence of generative modeling across the entire recommendation pipeline and provides a systemic framework for understanding recent advances.

From a methodological perspective, the generative capability can be incorporated into recommendation processes at different stages. Data-level approaches focus on leveraging generative models to enhance or augment the data space—such as synthesizing user behaviors, generating item attributes, or mitigating data sparsity. Model-level approaches integrate generative mechanisms directly into the recommendation architecture. Task-level approaches further extend generative modeling to high-level objectives, including personalized content creation, explainable recommendation, and conversational recommendation tasks, etc. This taxonomy not only delineates the scope of generative recommendation along the end-to-end system pipeline — ranging from data preparation to model design and task realization — but also echoes the historical evolution of the field. Early research primarily employed generative modeling for data-level enhancement, addressing challenges such as data sparsity and bias through synthetic interactions or item generation. Subsequent works advanced toward model-level innovation, where generative inference and probabilistic reasoning became integral parts of recommendation architectures. More recently, task-level generation has emerged, extending generative capabilities to high-level applications such as personalized content creation, multi-modal recommendation, and explainable reasoning. Together, these three perspectives form a coherent and progressive framework that reflects both the methodological depth and practical breadth of generative recommendation research—from its early data augmentation roots to its modern role as an intelligent, generation-driven paradigm.



(a) Discriminative Recommender (b) Generative Recommender

Fig. 1. Discriminative Recommendation and Generative Recommendation.

1.1. How do we collect the papers?

We collected over 200 papers related to generative models (LLMs and diffusion models) for recommender systems. Initially, we searched top-tier conferences and journals such as ICML, ICLR, NeurIPS, ACL, SIGIR, KDD, WWW, RecSys, TKDE, TOIS to identify related works in the past 3 years. We also include some highly cited work in ArXiv. To collect relevant literature, we first defined two groups of search keywords. The first group is related to recommendation systems, including terms such as "recommendation", "recommender system", and "user modeling". The second group focuses on generative and foundation models, including "generative", "agent", "large language model", "reasoning", "foundation model", and "retrieval-augmented generation (RAG)". We then constructed combinations of these keywords to perform systematic searches.

1.2. Contributions of this survey

Research on generative recommendation is accelerating. A growing body of surveys has recently emerged to review this emerging research direction. In 2024 and before, Wu et al. (2024c) systematically reviewed the LLM-based recommendation systems. From the perspective of modeling paradigms, they categorized the studies into LLM Embeddings enhanced RS, LLM Tokens enhanced RS, and LLM as RS. (Lin et al., 2025a) introduced two orthogonal perspectives: where and how to adapt LLMs in recommender systems. They introduced the existing methods according to whether to tune LLM and whether to involve conventional recommendation models. (Zhao et al., 2024a) reviewed of LLM-empowered recommender systems from various aspects, including pre-training, fine-tuning, and prompting paradigms. (Deldjoo et al., 2024) connects key advancements in RS using Generative Models (Gen-RecSys), covering: interaction-driven generative models; the use of LLM and textual data for natural language recommendation; and the integration of multimodal models for generating and processing images/videos in RS. (Liu et al., 2024f) explored the advancements in multimodal pretraining, adaptation, and generation techniques, as well as their applications to recommender systems. (Li et al., 2023c) reviewed the recent progress of LLM-based generative recommendation and provided a general formulation for each generative recommendation task according to relevant research. (Wang et al., 2024d) reviewed existing LLM-based recommendation works and discussed the gap from academic research to industrial application.

However, despite their comprehensive scope, these surveys primarily reflect the state of the field up until 2024, and many of them

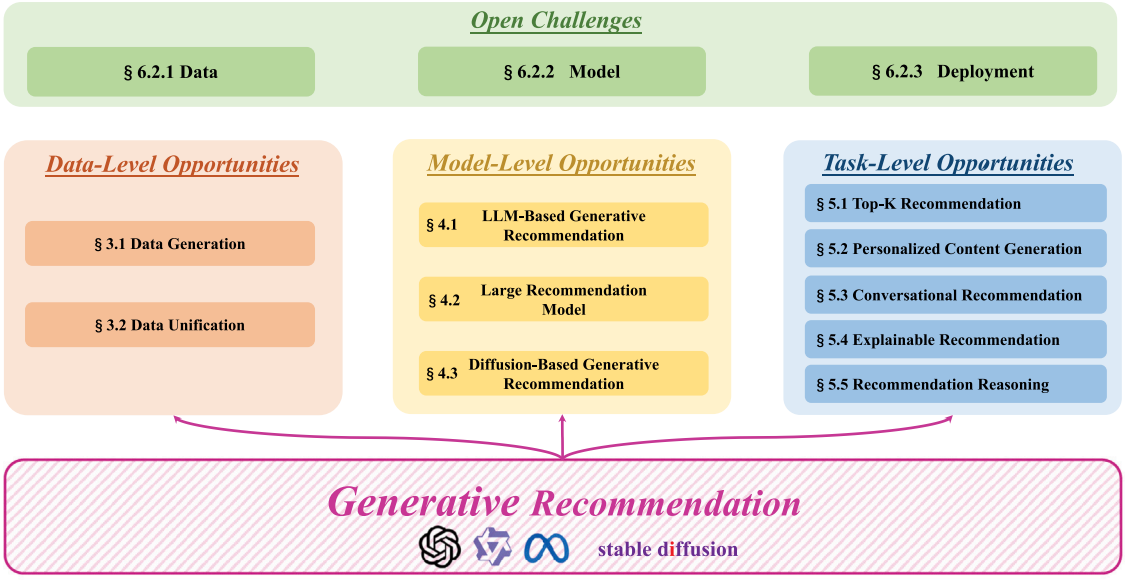


Fig. 2. Overview of this survey.

have overlooked the latest research that emerged in 2025 and beyond. Specifically, they fail to fully address the new research on agent-based recommender systems using LLMs, the rapidly growing exploration of LLM-based recommendation methods beyond supervised fine-tuning (SFT), and so on.

In contrast, our survey provides three key contributions. First, we offer a broader coverage of generative paradigms, categorizing research into LLM-based generative recommendation, large recommendation models (LRMs), and diffusion-based generative recommendation, ensuring we incorporate the most recent advancements that may have been overlooked in earlier surveys. Second, we introduce a data–model–task framework, which allows us to analyze the contributions of generative models at different levels, from data preparation to model architecture to task-specific innovations. This framework provides a more comprehensive understanding of the evolving role of generative models in recommendation systems. Third, we dedicate a section to task-level innovations, examining how generative models unlock novel recommendation scenarios such as interactive recommendation, conversational recommendation, and personalized content generation—an area that has been underexplored in prior works. Finally, our survey concludes with a discussion on current challenges and future research directions, offering an up-to-date roadmap for advancing the field of generative recommendation systems, especially in light of recent developments that have yet to be fully explored in the literature.

1.3. Scope and organization of this survey

The landscape of recommendation modeling can be broadly examined across three fundamental dimensions: **data**, **model**, and **task**. This survey centers on the transformative impact of generative models on each of these dimensions, aiming to provide a systematic and comprehensive review of the current state of research. The overview of this survey is shown in Fig. 2, and a detailed taxonomy of representative works along the data–model–task dimensions is presented in Fig. 3.

The remainder of this paper is organized as follows. Section 2 introduces the preliminary and background of recommendation, contrasting traditional discriminative approaches with generative paradigms, and highlighting their key concepts, differences, and distinctive characteristics. Section 3 surveys methods that incorporate generative models (particularly LLMs) at the data level, including feature generation, interaction synthesis, and data integration. Section 4 examines approaches that leverage generative models as the core recommendation engine. We categorize mainstream research into three directions: LLM-based generative recommendation, large recommendation models, and diffusion-based generative recommendation. Section 5 discusses task-level innovations enabled by generative models in recommendation. Section 6 outlines open challenges and future research opportunities in generative recommendation, proposing directions for advancing this emerging field. Finally, Section 7 concludes the survey by summarizing the contributions of generative models to RSs.

2. Preliminary and background

In this section, we begin with a foundational overview of traditional discriminative and generative recommendation models. We then explore the potential benefits that generative models can bring to recommender systems.

2.1. Preliminaries of discriminative & generative recommendation

Strictly speaking, discriminative models learn the conditional probability $P(y|x)$, or directly map input x to predict output y. In contrast, generative models learns the joint probability distribution $P(x, y)$, meaning it models how both the input x and the label y are generated together.

2.1.1. Discriminative recommendation models

For the recommendation task, discriminative recommendation models focus on learning a scoring or ranking function $f(u, i)$ that estimates the relevance or affinity between a user u and an item i. According to the recommendation system construction pipeline, we can divide it into three parts: data, model, and task.

Data Preparation. Given training data consisting of tuples $D = \{(u, i, y_{ui})\}$, where u represents a user, i represents an item, and y_{ui} denotes the observed interaction (ratings for explicit feedback or binary values $\{0, 1\}$ for implicit feedback). The inputs u and i can be one-hot IDs in collaborative filtering methods. Auxiliary content data are commonly utilized to enrich u and i, including but not limited to user social networks and profiles for users, and multimedia descriptions (images, videos, texts, audio) and knowledge graphs for items.

Model Construction. At the training time, discriminative recommendation methods start by utilizing embedding layers to map each user and item to a dense embedding vector: $\mathbf{e}_u = \phi_u(u), \mathbf{e}_i =$

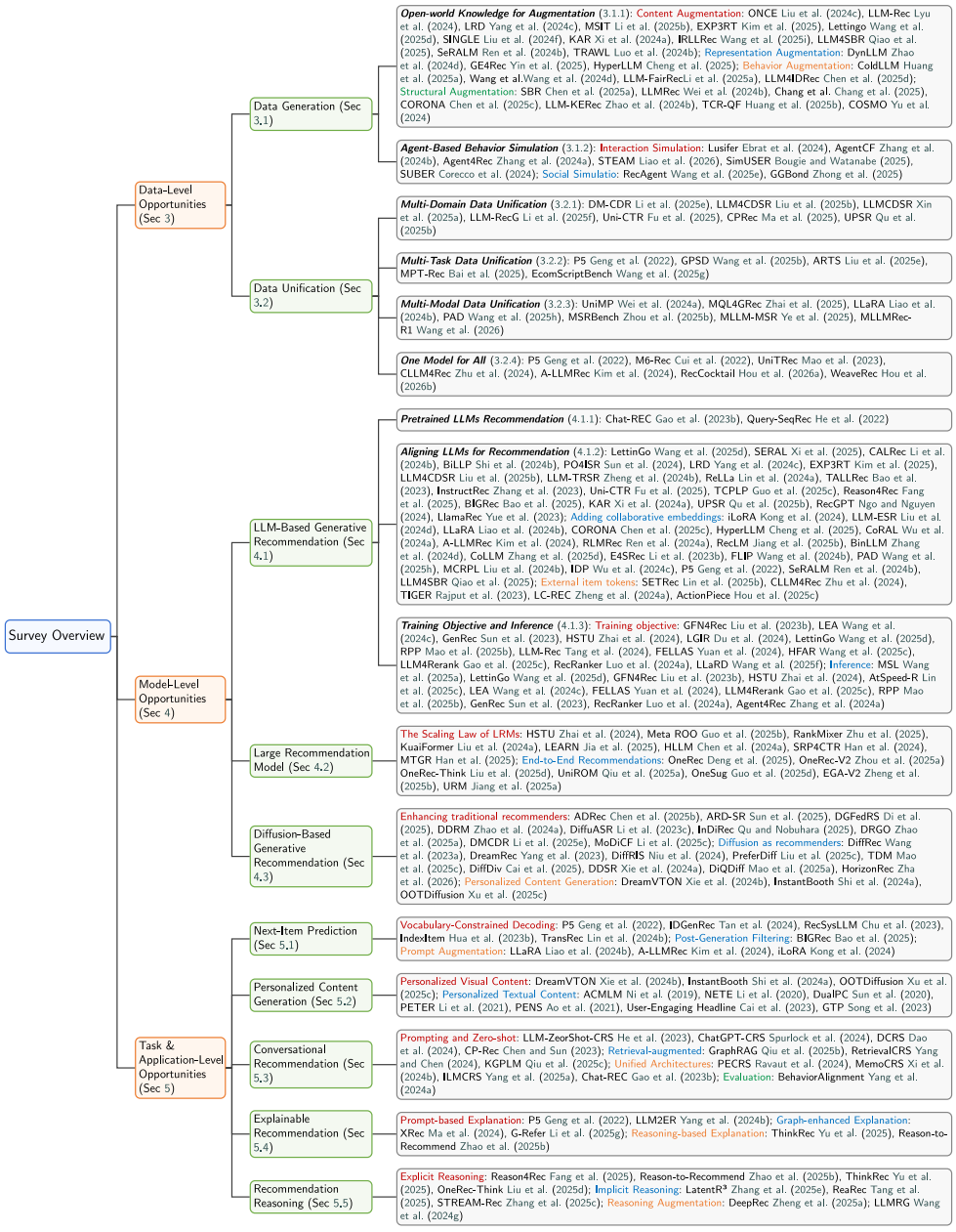


Fig. 3. Taxonomy of research on generative recommendation.

$\phi_i(i)$. $\phi_u(u)$ and $\phi_i(i)$ are embedding layers for users and items, often simple lookup tables, but can include Multi-Layer Perceptrons (MLPs), Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), transformers, etc. Then, discriminative recommendation models computes a matching score between the user and item embeddings: $f_{ui} = \text{Score}(\mathbf{e}_u, \mathbf{e}_i)$. Common scoring functions include inner product, distance-based metrics, and neural network-based metrics (e.g., MLP and CNN). The discriminative recommendation models are usually trained to discriminate between positive and negative interactions. Commonly used loss functions include Mean Squared Error (MSE) loss $\mathcal{L}_{\text{rating}}$ and Binary Cross Entropy (BCE) loss $\mathcal{L}_{\text{point}}$ for explicit feedback, and Bayesian Personalized Ranking (BPR) loss $\mathcal{L}_{\text{pair}}$ for implicit feedback.

$$\mathcal{L}_{\text{rating}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{ui} - f(u,i)\right)^2, \tag{1}$$

$$\mathcal{L}_{\text{point}} = -\sum_{(u,i)\in D}[y_{ui}\log\sigma(f_{ui}) + (1-y_{ui})\log(1-\sigma(f_{ui}))], \tag{2}$$

$$\mathcal{L}_{\text{pair}} = -\sum_{(u,i^+,i^-)\in D}\log\sigma(f_{ui^+} - f_{ui^-}). \tag{3}$$

Recommendation Task. For discriminative recommendation systems, the final recommendation task is mostly to select the top k items that users might like from a candidate list, which we refer to as top-k recommendation. At inference time, given a user u and candidate item set $\mathcal{I}$, the discriminative recommendation models compute scores and ranks the items.

$$\hat{i} = \arg\max_{i\in\mathcal{I}} f(u,i), \qquad \text{TopK}_u = \text{Top-K}_{i\in\mathcal{I}} f(u,i). \tag{4}$$

This process requires calculating the user's matching score for each item in the candidate list, then ranking and selecting the top-k items for recommendation.

2.1.2. Generative recommendation models

In this survey, we systematically categorize all recommendation approaches that utilize generative models within the overarching paradigm

of generative recommendation. We define generative recommendation as a broad recommendation paradigm in which generative models (such as LLMs, diffusion models) are utilized in stages of the recommendation pipeline. We can still categorize the recommendation pipeline into three major phases: data, model, and tasks. We identify three paradigms of generative recommendation in current research: data-level synthesis, model-level recommendation, and task-level generation.

Data-Level Synthesis. Generative models are used to synthesize training data, including both user/item features and interaction records. This is especially useful in scenarios like cold-start problems or sparsity in the dataset. Formally, let D represents the original dataset consisting of user set $\mathcal{U}$, item set $\mathcal{I}$, and interaction set $\mathcal{Y}$, generative models produce:

$$\mathcal{U}', \mathcal{I}', \mathcal{Y}' = G_{\text{data}}(\mathcal{U}, \mathcal{I}, \mathcal{Y} | \theta_g),$$

where G_{data} generates synthetic user features, item features, and interaction records to create an enriched training dataset for recommendation models.

Model-Level Recommendation. At the model level, generative models serve as the core recommendation engine to directly learn user preferences and generate personalized recommendations. Current mainstream work in this paradigm can be categorized into three major approaches: LLM-based approaches, large recommendation models, and diffusion model-based approaches. (i) LLM-based approaches leverage pre-trained language models as recommendation backbones. These methods convert recommendation data into samples with textual input and output, and model the recommendation task as the neural language generation process. (ii) Large recommendation models scale up traditional recommendation architectures with generative components, employing massive parameters to model complex user–item interactions and generate high-quality recommendations. (iii) Diffusion model-based approaches treat recommendation as a denoising process, learning to generate user preferences or item rankings through iterative refinement from noise to meaningful recommendation signals.

Task-Level Generation. At the task level, generative models reformulate recommendation as a generation task, producing recommendation outputs in natural language or structured formats. This paradigm not only addresses traditional recommendation tasks such as sequential recommendation and click-through rate prediction through generative approaches, but also enables novel recommendation capabilities including generating personalized explanations for recommendations, creating conversational recommendation dialogues, producing item reviews and descriptions, creating virtual items, and creating multi-modal recommendation content that combines text, images, and other media formats. This paradigm opens up new possibilities for interpretable and interactive recommendation experiences.

2.2. Advantages of generative models

In this subsection, we summarize the key advantages that generative models bring to recommender systems by addressing fundamental challenges in understanding users and items.

World Knowledge Integration. Effective recommendations require understanding not only user–item interaction patterns but also the rich semantic information and real-world context of items themselves. Traditionally, incorporating world knowledge into recommendations required explicit content enrichment methods, such as extracting knowledge graphs and leveraging auxiliary content information. These approaches often involved multi-stage pipelines with separate modules for knowledge extraction and integration. Generative models, particularly LLMs pre-trained on vast and diverse datasets, inherently encode extensive world knowledge about entities, events, relationships, and cultural contexts. By adopting a generative paradigm for recommendation, systems can directly leverage this embedded knowledge without requiring step-by-step knowledge extraction pipelines. For example, when recommending movies, a generative model can naturally incorporate information about directors, actors, cultural trends, or even recent events, all without explicit knowledge base construction. This seamless integration of world knowledge enables more contextually aware and informed recommendations.

Natural Language Understanding. Personalized recommendation fundamentally relies on understanding users. While implicit behavioral signals (clicks, purchases, ratings) have been the primary data source, users naturally express their preferences, needs, and intentions through language, including search queries, reviews, conversations, and feedback. Traditional recommendation systems often struggle to effectively process and understand these rich natural language signals. Generative models, particularly LLMs, with their advanced natural language understanding capabilities, can directly interpret user expressions in free-form text, understanding nuances, context, and intent. This enables recommendation systems to capture user preferences expressed in natural language, support conversational recommendation interfaces, and understand complex, multi-faceted queries. For instance, when a user asks "I want something relaxing but not boring for a Friday night", a generative model can parse this nuanced request and generate appropriate recommendations, bridging the gap between how users naturally communicate and how recommendation systems operate.

Reasoning Capabilities. User decision-making in recommendation scenarios often involves more than simple preference matching; it requires reasoning. Particularly in complex decision contexts (such as selecting financial products, planning travel, or making healthcare choices), users engage in multi-step reasoning processes that consider trade-offs, constraints, and causal relationships. Traditional recommendation models typically rely on pattern matching and correlation, lacking explicit reasoning abilities. Generative models, with their emergent reasoning capabilities, can model the logical processes behind user decisions. They can understand "why" a user might prefer one item over another by considering feature relationships, temporal sequences, and contextual factors. This reasoning ability allows generative recommendation systems to provide not just item suggestions but also explanations and justifications, making recommendations more transparent and trustworthy.

Scaling Law. The scaling law observed in large language models demonstrates that model performance improves significantly as both model size and training data increase. This principle offers a promising pathway for building more capable recommendation systems. By scaling up generative models with larger parameters and more diverse training data, recommendation systems can capture deeper patterns in user behavior and item characteristics. The scaling law suggests that as we increase the scale of generative models for recommendation, they can exhibit emergent capabilities that were not explicitly programmed, such as better understanding of user intent, more nuanced preference modeling, and improved handling of complex recommendation scenarios. This provides a systematic approach to advancing recommendation quality: instead of manually engineering features or designing intricate model architectures for each specific scenario, we can leverage the scaling law to achieve better performance through increased scale. Furthermore, larger-scale generative models trained on massive amounts of recommendation data across multiple platforms and contexts can learn universal patterns in user–item interactions, enabling more sophisticated and accurate personalized recommendations.

Generative Capabilities for Novel Recommendations. Unlike discriminative recommendation models that rank or select from existing item candidates, generative models can create novel content and recommendations. This capability is particularly valuable in cold-start scenarios, where new users or new items lack historical interaction data. Generative models can synthesize recommendations by drawing on broader patterns, user archetypes, and item similarities, ensuring meaningful suggestions even without explicit past behavior. Moreover, they can generate diverse and creative recommendations that break free from the filter bubble problem common in traditional systems. For

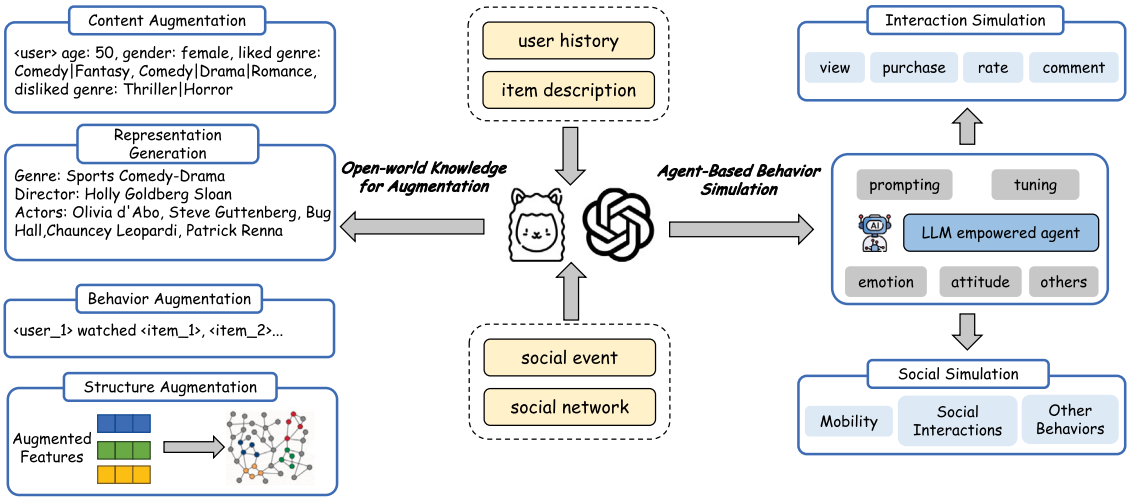


Fig. 4. Outline of key techniques in LLM-empowered data generation.

instance, instead of merely ranking a fixed set of products, a generative model could suggest customized bundles, personalized content variations, or even generate entirely new item descriptions tailored to individual user preferences.

2.3. When does generative recommendation outperform discriminative approaches?

While the advantages outlined above highlight the transformative potential of generative models, it is important to recognize that their superiority over discriminative approaches is not universal. It depends critically on the task, data, and deployment context. Based on existing empirical evidence, we identify the following three main conditions under which generative recommendation demonstrates genuine and reliable effectiveness gains. (1) Data-sparse and cross-domain scenarios. When interaction data is scarce or distributed across domains, LLMs' embedded world knowledge compensates for insufficient behavioral signals, leading to consistent improvements in cold-start and zero-shot settings that discriminative models struggle to match. (2) Inherently generative tasks. For tasks such as conversational recommendation, explainable recommendation, and personalized content creation, generative models hold a fundamental advantage, as these tasks require open-ended generation capabilities that discriminative scoring functions cannot provide. (3) Large-scale training regimes. As demonstrated by HSTU (Zhai et al., 2024) and related large recommendation models, generative architectures benefit from scaling laws that allow performance to improve predictably with increased model size and data, a property that discriminative models plateau on beyond a certain scale.

These observations suggest that the effectiveness of generative recommendation is deeply intertwined with how generative models are leveraged across the entire recommendation pipeline, from enriching sparse data and scaling model capacity to enabling novel task formulations, which motivates the data-model-task framework adopted in this survey.

3. Data-level opportunities

LLMs have unlocked new possibilities for data-centric advancements in recommender systems by enabling effective data generation and unification. Unlike traditional methods that depend solely on existing datasets, LLMs empower systems to actively generate rich user and item data, as well as simulate realistic interaction scenarios. Additionally, LLMs facilitate the unification of heterogeneous data sources across tasks, domains, and modalities, addressing longstanding challenges of data sparsity and fragmentation. This section reviews recent progress in two primary areas: (1) **data generation**, encompassing open-world knowledge augmentation and agent-based behavior simulation, as shown in Fig. 4, and (2) **data unification**, which integrates diverse data types to form coherent inputs for recommendation models. Together, these advances form a foundational basis for building more adaptive, generalizable, and generative recommendation systems.

3.1. Data generation

3.1.1. Open-world knowledge for augmentation

LLMs unlock unprecedented opportunities at the data foundation of recommender systems by leveraging their vast open-world knowledge, natural language understanding, and generative capabilities. Moving beyond passive data consumption, LLMs actively enrich, synthesize, and unify recommendation data, driving a new wave of data-centric innovations. This section reorganizes recent advances into four intuitive dimensions of augmentation: **Content**, **Representation**, **Behavior**, and **Structure**, which collectively enable more adaptive, generalizable, and knowledge-aware recommendation models. In Table 1, we summarize the four types of data augmentation methods.

Content Augmentation. LLMs have been extensively used to generate natural-language representations that enrich sparse behavioral signals and construct expressive user and item profiles. Current recommendation performance is constrained by the inherent limitations of incomplete or insufficient information in item descriptions. The world knowledge stored within LLMs can help overcome these limitations, enabling the extraction of more informative text and enriching available training data (Liu et al., 2024b). LLM-REC (Lyu et al., 2024) employs diverse prompting strategies to extract key insights from the rich world knowledge stored within LLMs, thereby enriching input texts. LRD (Yang et al., 2024b) utilizes variational reasoning to uncover latent inter-item relationships. MSIT (Li et al., 2025b) leverages multimodal large language models (MLLMs) to mine latent item attributes from images and text via self-corrective instruction tuning. Some approaches prompt large models to extract multifaceted, diverse user profiles from interaction logs and review information (Kim et al., 2025; Wang et al., 2025k). SINGLE (Liu et al., 2024c) extracts users' constant preferences and instant interests from read texts. KAR (Xi et al., 2024a) employs decomposable prompts to parse user intent, while IRLLRec (Wang et al., 2025f) aligns cross-modal signals to construct multimodal intent representations. LLM4SBR (Qiao et al., 2025) integrates behavioral and semantic clues for multi-perspective intent modeling. However, raw external knowledge remains poorly aligned with recommendation objectives. SeRALM (Ren et al., 2024a) designs

Table 1
Open-world Knowledge Data Augmentation for Recommender Systems.

Category	Representative Works	Description/Focus
Content Augmentation	ONCE (Liu et al., 2024b), LLM-Rec (Lyu et al., 2024), LRD (Yang et al., 2024b), MSIT (Li et al., 2025b), EXP3RT (Kim et al., 2025), Lettingo (Wang et al., 2025k), SINGLE (Liu et al., 2024c), KAR (Xi et al., 2024a), IRLLRec (Wang et al., 2025f), LLM4SBR (Qiao et al., 2025), SeRALM (Ren et al., 2024a), TRAWL (Luo et al., 2024b)	Generate natural-language user/item profiles, summarize histories, enrich sparse metadata, and align textual semantics with feedback.
Representation Augmentation	DynLLM (Zhao et al., 2024b), HyperLLM (Cheng et al., 2025)	Automated feature construction, multimodal attribute extraction, external knowledge distillation, and hierarchical category generation.
Behavior Augmentation	ColdLLM (Huang et al., 2025a), LLM-FairRec (Li et al., 2025c), LLM4IDRec (Chen et al., 2025a)	Generate synthetic user–item interactions, simulate cold-start preferences, ensure fairness, and integrate pseudo-interactions into ID-based pipelines.
Structure Augmentation	SBR (Chen et al., 2025b), CORONA (Chen et al., 2025d), LLM-KERec (Zhao et al., 2024c), TCR-QF (Huang et al., 2025b) , COSMO (Yu et al., 2024)	Relation discovery, graph completion, social network generation, subgraph retrieval, knowledge graph construction & distillation.

alignment-targeted prompts to guide LLMs in generating item descriptions aligned with recommendation goals, filtering out irrelevant noise generated by the model. Lettingo (Wang et al., 2025k) employs Direct Preference Optimization (DPO) based on the impact of generated user profiles on recommendation outcomes, ensuring profile adaptability and effectiveness. TRAWL (Luo et al., 2024b) encodes generated textual information into embeddings and utilizes adapters to align them with the recommendation task space.

Together, these approaches demonstrate how textual augmentation transforms unstructured data into structured, interpretable, and adaptive profiles that improve recommendation quality.

Representation Augmentation. Traditional feature engineering is increasingly replaced by LLM-driven generative approaches, which automate the construction of semantically rich and task-specific features. DynLLM (Zhao et al., 2024b) employs an LLM as a content encoder to extract content representations, then refines and integrates the model for recommendation tasks, thereby enhancing user modeling and content understanding. HyperLLM (Cheng et al., 2025) generates hierarchical category labels for hyperbolic embeddings.GE4Rec (Yin et al., 2025) proposes a generative "feature generation" paradigm, which predicts each feature embedding based on the concatenation of all feature embeddings. These advances highlight LLMs' role as automated feature generators that bridge unstructured inputs and structured model representations.

Behavior Augmentation. Data sparsity and cold-start challenges are key constraints limiting recommendation systems, and LLMs offer opportunities to address these challenges. Through appropriate prompting, LLMs can understand user behavior and generate context for content of interest to users. Thus, we can leverage the reasoning and generalization capabilities of LLMs to generate pseudo-interactions aligned with the preferences of cold-start users (Wang et al., 2024g; Huang et al., 2025a) and users unfairly treated due to insufficient interactions (Li et al., 2025c) by generating pseudo-interactions aligned with their preferences, addressing cold-start and fairness issues from a data perspective. Specifically, ColdLLM (Huang et al., 2025a) employs a coupled-funnel architecture of filters and refiners to screen cold-start users and simulate interactions. LLM-FairRec (Li et al., 2025c) employs fairness-aware prompts to generate fair pseudo-interactions aligned with the preferences of unfairly treated minority users in recommendation systems. Additionally, some approaches explore the potential of LLMs in pure ID-based recommendation systems. LLM4IDRec (Chen et al., 2025a) leverages fine-tuned LLMs to augment interaction data in ID format, thereby enhancing traditional ID-based recommendation systems.

Structure Augmentation. Beyond individual features or interactions, LLMs are increasingly leveraged to induce higher-level semantic structures (e.g., relations and graphs), which support structured reasoning. At a structural level, SBR (Chen et al., 2025b) aligns item features with hierarchical intents, LLMRec (Wei et al., 2024b) infers missing nodes and edges in graphs, and (Chang et al., 2025) simulate realistic social networks. Knowledge graph construction and refinement are also explored: CORONA (Chen et al., 2025d) retrieves intent-aware subgraphs, LLM-KERec (Zhao et al., 2024c) infers new triples, TCR-QF (Huang et al., 2025b) improves completion with query-aware generation, and COSMO (Yu et al., 2024) distills e-commerce knowledge graphs for recommendation. Collectively, structural augmentation enables knowledge-infused representations that enhance explainability, robustness, and generalization.

3.1.2. Agent-based behavior simulation

With recent advancements in LLMs, researchers are leveraging their three fundamental capabilities to design LLM-driven agents: (i) the ability to perceive and understand the environment, (ii) reasoning techniques that integrate task requirements with corresponding rewards to design task planning, and (iii) the generation of text resembling human language. These agents can comprehend complex user preferences, generate contextual recommendations through fine-grained dynamic simulations of user behavior and interactions, and achieve more nuanced decision-making beyond simple feature matching. Agents can simulate cognition, memory, emotions, decision-making, and reflection to generate more authentic user profiles, providing rich signals for training and evaluating recommendation systems. We divide this research direction into two sub-directions: interaction simulation and social simulation.

Interaction Simulation. With recent advancements in LLMs, LLM-based agents have demonstrated impressive capabilities in autonomous interaction and decision-making. Applications in the recommendation domain that utilize agents to simulate users (Ebrat et al., 2024) typically combine memory modules, role modeling, and reflective loops to generate adaptive and interpretable user behavior simulations. Specifically, Agent4Rec (Zhang et al., 2024b) scales this approach to user agents with factual and emotional memories, simulating diverse behaviors, enabling causal behavior analysis. AgentCF (Zhang et al., 2024c) simultaneously simulates user agents and item agents, modeling the collaborative filtering concept in traditional recommendation systems. STEAM (Liao et al., 2026) advances this line of research by introducing structured and evolving agent memory to capture the implicit, diverse, and dynamic preferences reflected in user interactions. By moving beyond a single overwritten preference summary, it enables agents to preserve multi-faceted interests, track preference evolution, and make decisions more consistent with real user choices. Other works like SimUSER (Bougie and Watanabe, 2025) and SUBER (Corecco et al., 2024) design cognitive agents with episodic memory, persona grounding, and MDP-based interaction planning, providing realistic behavior logs to support offline evaluation and reinforcement learning policy learning. Collectively, these efforts focus on enabling agents to simulate

users based on their interaction history and reviews, thereby making decisions that align with real-world user choices.

Social Simulation. Beyond individual interactions, agent-based frameworks also simulate large-scale social dynamics. Social simulation aims to leverage user characteristics, contextual information within social networks, and complex mechanisms governing user cognition and decision-making. It employs simulation models capable of reasonably replicating the dynamic properties of historical social behaviors to simulate group-level dynamics, such as the propagation processes of information and emotions (Gao et al., 2023a; Piao et al., 2025; Jiang et al., 2024). Regarding explorations of social simulation in recommendation systems, GGBond (Zhong et al., 2025) integrates human-like cognitive agents and dynamic social interactions. It effectively models users' evolving social ties and trust dynamics based on interest similarity, personality compatibility, and structural homogeneity. By responding to recommendations and dynamically updating internal states and social connections, it forms stable multi-round feedback loops. Meanwhile, RecAgent (Wang et al., 2025j) constructs an interactive sandbox to simulate and study two social scenarios in recommendation systems: information silos and conformity. These social simulations contribute rich, multi-agent interaction data that extend beyond isolated user behaviors, providing valuable inputs for building more socially aware and robust recommender systems.

3.2. Data unification

LLMs provide powerful tools for unifying heterogeneous data across tasks, domains, and modalities, addressing long-standing challenges such as behavior sparsity, domain shift, and representational inconsistency. By leveraging LLMs' ability to encode diverse inputs into shared semantic spaces and perform prompt-driven generalization, recommender systems can move beyond siloed data processing toward unified architectures. This section highlights four major data-level opportunities: (1) **multi-domain unification**, enabling cross-domain transfer and cold-start adaptation; (2) **multi-task unification**, consolidating different recommendation objectives into a single learning framework; (3) **multi-modal unification**, integrating textual, visual, and behavioral signals into cohesive representations; (4) **one-model for all**, which builds general-purpose recommender models across scenarios. Together, these approaches reflect the growing trend of using LLMs to unify fragmented data sources and create more holistic, scalable recommendation systems.

3.2.1. Multi-domain data unification

Cross-domain recommendation faces challenges of behavioral sparsity, domain gaps, and representation misalignment. Traditional cross-domain recommendation approaches either model domain overlap from an item perspective or model cross-domain interaction sequences from a user perspective (Liu et al., 2024d; Guo et al., 2025c). However, due to the scarcity of overlapping users in practical scenarios, both methods encounter significant challenges. Diffusion-based DMCDR (Li et al., 2025d) employs a preference encoder to establish preference-guided signals based on source domain interaction histories. These signals are progressively and explicitly injected into user representations to guide the reverse process, enabling cross-domain transfer of user preferences. The powerful representation and reasoning capabilities of LLMs hold promise for bridging semantic relationships between items and deriving users' fine-grained preferences from cross-domain mixed interaction sequences. LLM4CDSR (Liu et al., 2025d) employs LLMs to extract semantic representations of items at the semantic level, learning connections between cross-domain items and modeling users' cross-domain interaction sequences. LLMCDSR (Xin et al., 2025a) employs LLMs to comprehend cross-domain information, generating pseudo-interactions for non-overlapping users regarding overlapping items. LLM-RecG (Li et al., 2025e) tackles zero-shot CDSR by aligning items and sequences without target data, showing promising results.

In multi-domain recommendation systems, models are prone to being dominated by domains possessing vast datasets. Several approaches have leveraged LLMs to learn domain-general knowledge and inject it into domain-specific recommendation tasks, thereby mitigating this challenge. Uni-CTR (Fu et al., 2025) employs LLMs to learn hierarchical semantic representations, capturing commonalities across domains. Other approaches train pre-trained language models (PLMs) to combine general textual information with personalized behavioral information from user history sequences, enhancing multi-domain recommendation tasks (Ma et al., 2025; Qu et al., 2025b). CPRec (Ma et al., 2025) mitigates domain distribution gaps by holistically aligning LLMs with general user behavior through a continuous pre-training paradigm.

In summary, LLM-powered multi-domain unification advances semantic bridging through pretraining, prompt tuning, and modular design, though challenges in scalability and domain interference remain.

3.2.2. Multi-task data unification

Modern recommender systems face diverse tasks — rating, ranking, explanation, intent recognition — often modeled separately, limiting efficiency and generalization. Multi-task unification integrates these objectives into a single framework for better transfer and scalability. A common strategy reformulates tasks as language generation. P5 (Geng et al., 2022) pioneered this, unifying recommendation tasks via text-to-text modeling with personalized prompts. GPSD (Wang et al., 2025g) further combines generative pretraining with discriminative fine-tuning to improve ranking accuracy. ARTS (Liu et al., 2025b) uses self-prompting for joint prediction and explanation, enhancing interpretability. Efficient Multi-task Prompt Tuning (Bai et al., 2025) reduces training cost by attaching lightweight prompts to a shared model, enabling scalable deployment. EcomScriptBench (Wang et al., 2025a) offers a multi-task benchmark simulating real-world shopping flows with intent recognition and explanation, supporting realistic evaluation. In short, multi-task unification boosts performance and flexibility but faces challenges in task interference and parameter efficiency. Future work may focus on modular prompts and unified multimodal training.

3.2.3. Multi-modal data unification

Recommendation involves multiple modalities like text, images, and behavior logs. Traditional methods fuse these separately, but large vision-language models (LVLMs) now enable unified multimodal representations, improving accuracy and interpretability. UniMP (Wei et al., 2024a) and MQL4GRec (Zhai et al., 2025) unify multimodal inputs into shared semantic spaces, boosting cross-modal generalization. LLaRA (Liao et al., 2024b) integrates item IDs and text via hybrid prompting for sequential recommendation, while PAD (Wang et al., 2025e) aligns modalities through a three-stage pretrain-align-disentangle process to enhance cold-start performance. MSRBench (Zhou et al., 2025b) benchmarks LVLM strategies, showing reranking models like GPT-4o balance performance and cost best, though real-time use remains challenging. MLLM-MSR (Ye et al., 2025) designed an item-summarizer based on multimodal large language models (MLLMs) to extract image features of a given item and convert them into text. Supervised fine-tuning (SFT) enables the MLLMs to be employed for multimodal recommendation tasks. Recent advances explore generative and agentic multimodal recommendation, such as Rec-GPT4V's zero-shot, multimodal interactions and frameworks for serendipitous recommendations. Some work improves explainability by mapping latent embeddings to readable text. Future systems aim for dynamic, interactive recommendations via multimodal LLMs. In short, multimodal unification harnesses foundation models to merge diverse signals into cohesive inputs, advancing personalization and generalization despite challenges in efficiency and deployment.

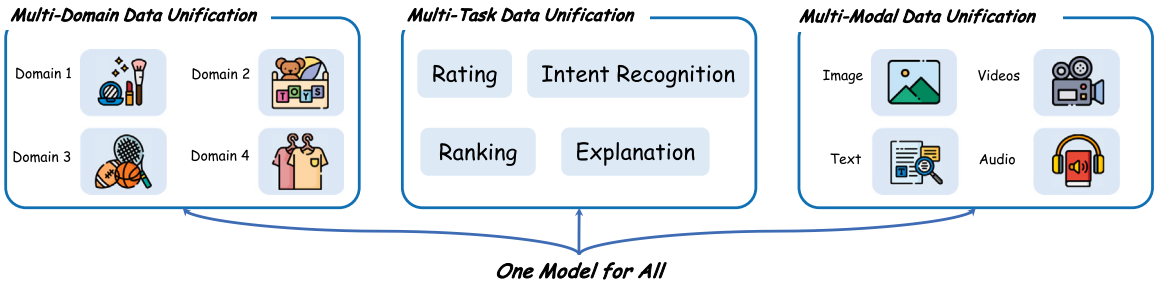


Fig. 5. LLM empowered data unification.

3.2.4. One Model for All

The rapid rise of LLMs is shifting recommender systems from task-specific designs to unified, general-purpose models capable of handling diverse tasks, domains, and modalities, as shown in Fig. 5. This "One Model for All" paradigm leverages foundation models' capacity for encoding heterogeneous data and performing multi-task inference via prompt-driven frameworks. Early work like P5 (Geng et al., 2022) reformulates recommendation as a text-to-text generation problem, unifying tasks such as rating prediction and sequential recommendation with personalized prompting. Building on this, M6-Rec (Cui et al., 2022)removes fixed candidate sets, enabling open-ended multimodal generation combining user behavior, text, and images. UniTRec (Mao et al., 2023) enhances representation by integrating generative modeling with contrastive learning, improving user intent and item semantic understanding. Similarly, CLLM4Rec (Zhu et al., 2024) incorporates user/item IDs into LLM vocabularies to unify generation and ranking, bridging collaborative filtering with language models. Beyond recommendations, joint modeling of search and recommendation as sequence generation has shown promising transfer benefits (Penha et al., 2024). Multimodal frameworks like vision-language model combine images and text to deliver personalized recommendations and explanations, expanding recommendation toward expressive generation (Rocamonde et al., 2023). From a practical view, A-LLMRec (Kim et al., 2024) integrates pretrained LLMs with collaborative filtering embeddings, using frozen language models and multi-task training to support cold-start, cross-domain, and explainable recommendations in scalable systems. On representation learning, works like LLM-Rec repurpose pretrained LLMs to generate semantic user and item embeddings, replacing manual features and demonstrating strong cross-domain generalization without fine-tuning. Beyond unifying tasks and modalities at the data or prompt level, a recent line of work tackles the "One Model for All" challenge from the perspective of *model merging*, i.e., composing a single deployable LLM-based recommender by fusing multiple domain/task specialized LoRA modules in the weight space. RecCocktail (Hou et al., 2026a) trains a reusable "base spirit" LoRA on multi-domain data and merges it with lightweight domain-specific "ingredient" LoRAs, yielding a unified model that inherits both generalization and specialization at no extra inference cost. WeaveRec (Hou et al., 2026b) extends model merging to multi-domain fusion in cross-domain sequential recommendation, resolving the conflicts that arise when merging multiple domain-specific LoRAs, and demonstrating that its merging-based approach consistently outperforms naive data-level fusion that simply mixes multi-domain data for joint training. Together, these two works establish model merging as a principled and deployment-friendly route to the "One Model for All" vision.

4. Model-level opportunities

Traditional recommender systems primarily use a discriminative framework based on passive user–item interactions for recommendation tasks. However, this approach has notable limitations in both recommendation accuracy and user engagement, as it lacks the ability to incorporate interactive, multi-turn, language-based feedback for refining recommendations.

The development of generative models offers significant opportunities to address the limitations of traditional recommender systems. LLMs, with their extensive world knowledge, hold substantial potential to enhance recommendation performance. At the same time, the rise of AI-generated content introduces new paradigms, such as the automatic editing of existing items and the generation of new ones. Additionally, diffusion models have gained considerable attention in the recommendation community due to their impressive generative capabilities. This section explores model-level opportunities by categorizing existing generative recommendation systems into three main approaches: (1) **LLM-Based Generative Recommendation**, (2) **Large Recommendation Models**, and (3) **Diffusion-Based Generative Recommendation**.

4.1. LLM-based generative recommendation

With the rapid progress of LLMs, applying them to recommendation has become a major research line. LLM-based generative recommendation leverages pretrained LLMs to produce personalized suggestions via natural-language prompts or lightweight fine-tuning, tapping the broad world knowledge encoded in the models. Existing studies can be grouped into three strands: (i) Pretrained LLMs for recommendation, which explore prompt design and tool use to solve recommendation tasks without heavy retraining; (ii) Aligning LLMs to recommendation, which tackles the mismatch between generic language modeling objectives and the goals of ranking, personalization, and domain constraints; and (iii) Training objectives and inference, which develop fine-tuning losses, preference-optimization methods, and efficient decoding to improve accuracy and efficiency.

4.1.1. Pretrained LLMs recommendation

The utilization of pretrained LLMs for recommendation relies on prompt design and in-context learning, so the models can be applied without extra fine-tuning. A key benefit is that LLMs bring broad world knowledge and strong semantic understanding, which helps in cold-start and cross-domain settings. Research in this line follows two complementary routes: (i) LLM-as-Enhancer (Sileo et al., 2022; Hou et al., 2024; Kang et al., 2023; Liu et al., 2023b; Harrison et al., 2023). LLMs are used to rewrite user/item profiles and interaction histories into natural-language features, which are then fed to or combined with collaborative filtering, sequential models, or re-rankers. This improves explainability, user interaction, and sometimes long-tail coverage. (ii) LLM-as-Recommender (Gao et al., 2023b). With task-specific prompts or templates, a pretrained LLM directly generates recommendations (e.g., item titles or IDs) and can operate in zero-shot mode across scenarios. This idea extends to multimodal settings (Wang et al., 2025i, 2026), where MLLMs take both text and images (e.g., posters) as input to produce zero-shot recommendations. These approaches show how pretrained LLMs can either enhance existing pipelines or serve as generators, offering flexible ways to deploy LLM capability in recommender systems (Harrison et al., 2023).

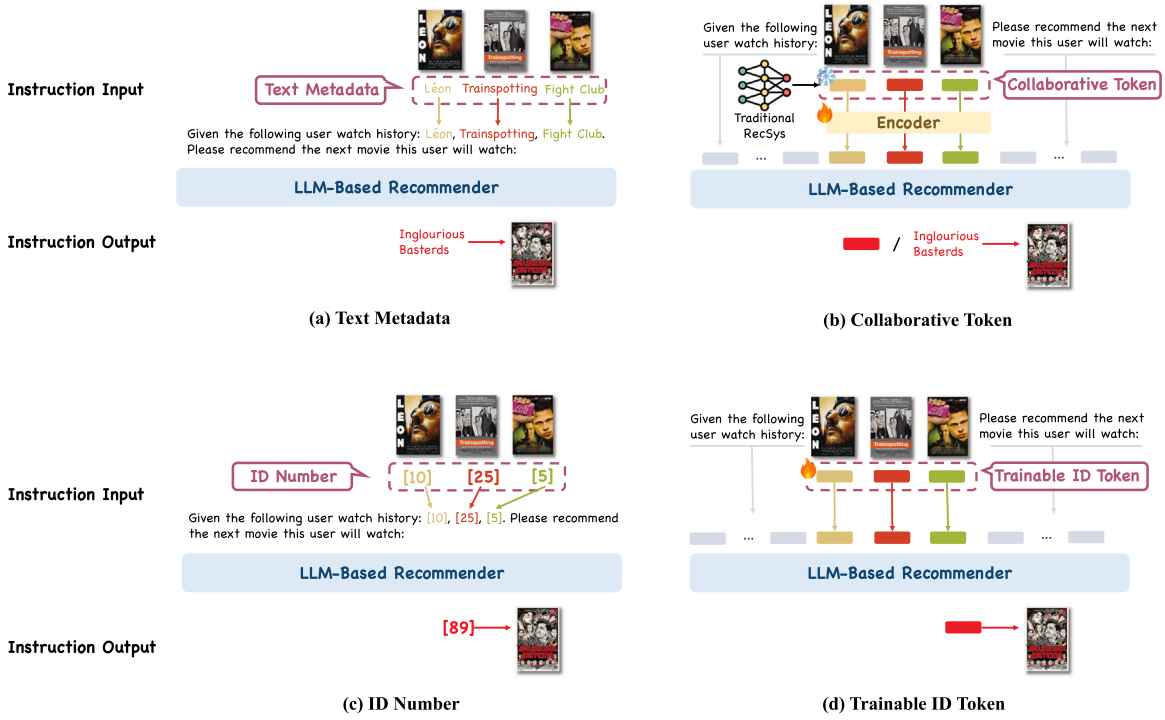


Fig. 6. The paradigms of aligning LLMs for recommendation. Inspired by the figure in Yue et al. (2023).

4.1.2. *Aligning LLMs for recommendation*

Aligning LLMs for recommendation adapts a pretrained model to recommender objectives and evaluation. While LLMs offer strong semantics and zero-/few-shot use, their pretraining ignores click/engagement signals, top-k ranking, long-tail coverage, and exposure bias, so direct use often trails specialized models. To bridge this gap, fine-tuning on recommendation-specific data has emerged as a promising approach to align the training signal with language models. During fine-tuning, these works generate user/item profiles to present the LLM with a structured, compact, and consistent collaborative signal. By how the user/item profile injects structure, approaches include: text prompting, collaborative signal, and item tokenization. In Tables 2, 3, 4, we summarize representative works of these methods, respectively.

Text prompting based methods. Text prompting based methods build the user profile entirely in natural language by combining the task description with the user's chronological interaction history. The framework is shown in Fig. 6(a). Early work (Li et al., 2024a; Shi et al., 2024a; Sun et al., 2024) mainly feeds the sequence of consumed items, but such inputs may miss fine-grained preference signals. To enrich the profile, later studies explicitly model preferences within the history: TALLRec (Bao et al., 2023) inserts explicit preference statements into prompt templates and uses LoRA for lightweight adaptation; LlamaRec (Yue et al., 2023) first contracts the candidate set with a sequential recommender to provide a focused context; LRD (Yang et al., 2024b) couples LLM knowledge with a variational autoencoder to reveal latent item relations. Some works also adopt preference optimization to keep profiles flexible—LettinGo (Wang et al., 2025k) uses direct preference optimization to adapt the model to user preferences without rigid supervised targets. Beyond static histories, recent studies incorporate dynamic feedback to track evolving interests (Ngo and Nguyen, 2024; Kim et al., 2025; Fang et al., 2025); for instance, Reason4Rec (Fang et al., 2025) leverages user reviews to extract preferences and salient item attributes, helping the model judge the match between user interests and candidates. While text prompting makes recommendations possible with textual descriptions alone, it lacks explicit collaborative signals. As a result, performance can fall short in settings where inter-item dependencies and collaborative patterns are critical for ranking quality.

Collaborative signal based methods. To capture inter-item relations that plain text misses, these methods inject collaborative signals into the user/item profile so the LLM sees both semantics and relational knowledge (Fig. 6(b)). Existing studies can be broadly categorized into three directions: (i) LLM-augmented representation for CF models, which map Collaborative signal and semantic representations into a shared space and fuse them to form enriched profiles. Methods such as (Kong et al., 2024; Liu et al., 2024e; Liao et al., 2024b; Wang et al., 2025e; Zhang et al., 2024a, 2025a; Li et al., 2023a; Ren et al., 2024a; Kim et al., 2024; Ren et al., 2024b; Jiang et al., 2025b) concatenate the two sources of information to construct enriched user profiles, leveraging their complementary strengths. (ii) LLM-assisted summarization for CF models, which distills user preferences into compact textual summaries that serve as auxiliary inputs to traditional recommenders. For example, CORONA (Chen et al., 2025d) combines LLM reasoning with a GNN in a coarse-to-fine pipeline, and HyperLLM (Cheng et al., 2025) uses LLM-generated summaries to enhance collaborative models. CoRAL (Wu et al., 2024a) reformulates collaborative signals as explicit sentences (e.g., "User A also prefers X, Y, Z"), making them more interpretable and directly usable by LLMs. While collaborative-signal methods enrich user profiles with interaction-derived information, dense embeddings are not natively interpretable for LLMs. In practice, they often require projection or verbalization to bridge representation spaces, and this mapping may introduce a potential gap between collaborative and textual semantics.

Item tokenization based methods. To narrow the gap between collaborative and textual semantics, this line of work maps items into the LLM's vocabulary by assigning identifiable tokens that the model can generate autoregressively (Fig. 6(c) and (d)). The key challenge is to design identifiers that are both efficient and semantically meaningful, balancing textual semantics with collaborative knowledge. Existing work can be grouped into five directions: (i) ID-based tokenization: The simplest approach assigns a special token to each user or item (Geng et al., 2022; Zhu et al., 2024). While conceptually straightforward, this strategy is infeasible at scale due to vocabulary explosion and its inability to encode semantics, making cold-start generalization difficult. (ii) Text-based tokenization: To introduce semantics, some

Table 2
Summary of the representative text prompting-based recommendation methods.

Methods	User formulation				Backbone
	Task description	Historical interactions	Profile	Feedback	
Chat-Rec (Gao et al., 2023b)	ranking	history interactions	✓	–	GPT-3.5
TALLRec (Bao et al., 2023)	preference classification	user preference	–	–	LLaMA-7B
LlamaRec (Yue et al., 2023)	retrieval, ranking	history interactions	–	–	LLaMA2-7B
LRD (Yang et al., 2024b)	ranking	history interactions	–	–	GPT-3.5
ReLLa (Lin et al., 2024a)	ranking	history interactions	✓	–	Vicuna-7B
CALRec (Li et al., 2024a)	ranking	history interactions	–	–	PaLM-2 XXS
BiLLP (Shi et al., 2024a)	long-term Interactive	history interactions, reward model	–	–	GPT-3.5, GPT-4, LLaMA2-7B
PO4ISR (Sun et al., 2024)	Ranking	history interactions	–	–	LLaMA2-7B
LLM-TRSR (Zheng et al., 2024a)	Ranking	history interactions	✓	–	LLaMA2-7B
RecGPT (Ngo and Nguyen, 2024)	Ranking	history interactions, user preference	–	✓	RecGPT-7B
KAR (Xi et al., 2024a)	Ranking	history interactions, user preference	✓	–	GPT-3.5
LLM4CDSR (Liu et al., 2025d)	Ranking	history interactions	✓	✓	GPT-3.5, GLM4-Flash
EXP3RT (Kim et al., 2025)	rating prediction	history interactions	✓	✓	LLaMA3-8B
SERAL (Xi et al., 2025)	retrieval, ranking	history interactions	✓	–	Qwen2-0.5B
LettinGo (Wang et al., 2025k)	Ranking	history interactions	✓	–	LLaMA3-8B
Reason4Rec (Fang et al., 2025)	Rating Prediction	history interactions, user preference	–	✓	LLaMA3-8B
InstructRec (Zhang et al., 2023)	Ranking	history interactions	✓	–	Flan-T5-XL
Uni-CTR (Fu et al., 2025)	Rating Prediction	history interactions, user preference	✓	–	DeBERTaV3-large
BIGRec (Bao et al., 2025)	Ranking	history interactions	–	–	LLaMA-7B
UPSR (Qu et al., 2025b)	Ranking	history interactions	–	–	T5, FLAN-T5

Table 3
Summary of the representative collaborative signal-based methods.

Methods	User Formulation				Combining Method	Backbone
	Task Description	Historical Interactions	Profile	Feedback		
iLoRA (Kong et al., 2024)	Ranking	history interactions	–	–	Concatenation	GPT-3.5
LLM-ESR (Liu et al., 2024e)	Ranking	history interactions	–	–	Concatenation	LLaMA2-7B
LLaRA (Liao et al., 2024b)	Ranking	history interactions	–	–	Concatenation	LLaMA2-7B
A-LLMRec (Kim et al., 2024)	Ranking	history interactions	–	–	Concatenation	OPT-6.7B
RLMRec (Ren et al., 2024b)	Ranking	history interactions, user preference	–	✓	Concatenation	GPT-3.5
CoRAL (Wu et al., 2024a)	Ranking	history interactions, user preference	–	✓	Retrieval-Augmented	GPT-4
BinLLM (Zhang et al., 2024a)	Ranking	history interactions, user preference	–	✓	Concatenation	Vicuna-7B
E4SRec (Li et al., 2023a)	Ranking	history interactions	–	–	Concatenation	Vicuna-7B
SeRALM (Ren et al., 2024a)	Ranking	history interactions	–	–	Concatenation	LLaMA2-7b
CORONA (Chen et al., 2025d)	Ranking	history interactions	✓	–	Pipeline Integration	GPT-4o-mini
HyperLLM (Cheng et al., 2025)	Ranking	history interactions	–	–	Pipeline Integration	LLaMA3-8B
RecLM (Jiang et al., 2025b)	Ranking	history interactions	✓	–	Concatenation	LLaMA2-7b
CoLLM (Zhang et al., 2025a)	Ranking	history interactions	–	–	Concatenation	Vicuna-7B
PAD (Wang et al., 2025e)	Ranking	history interactions, user preference	–	–	Concatenation	LLaMA3-8B
IDP (Wu et al., 2024b)	Ranking	history interactions, user preference	–	–	Concatenation	T5

Table 4
Summary of the representative item tokenization-based methods.

Methods	User formulation			Backbone
	Task description	Historical interactions	Token types	
P5 (Geng et al., 2022)	Ranking	historical interactions, user preference	ID-based tokenization	Transformer
CLLM4Rec (Zhu et al., 2024)	Ranking	historical interactions	ID-based tokenization	GPT-2
BIGRec (Bao et al., 2025)	Ranking	historical interactions	Text-based tokenization	LLaMA-7B
M6 (Cui et al., 2022)	Retrieval, Ranking	historical interactions	Text-based tokenization	M6
IDGenRec (Tan et al., 2024)	Ranking	historical interactions	Text-based tokenization	BERT4Rec
TIGER (Rajput et al., 2023)	Ranking	historical interactions	Codebook-based tokenization	T5
RPG (Hou et al., 2025a)	Ranking	historical interactions	Codebook-based tokenization	LLaMA-2-7B
LC-Rec (Zheng et al., 2024b)	Ranking	historical interactions	Codebook-based tokenization	LLaMA-2-7B
ActionPiece (Hou et al., 2025b)	Retrieval	historical interactions	Codebook-based tokenization	LLaMA-2-7B
LETTER (Wang et al., 2024a)	Ranking	historical interactions	Codebooks with collaborative signals	LLaMA-7B
TokenRec (Qu et al., 2025a)	Retrieval	historical interactions	Codebooks with collaborative signals	T5-small
SETRec (Lin et al., 2025b)	Ranking	historical interactions	Codebooks with collaborative signals	T5, Qwen
CCFRec (Liu et al., 2025e)	Ranking	historical interactions	Codebooks with collaborative signals	LLaMA-2-7B
LLM2Rec (He et al., 2025)	Ranking	historical interactions	Codebooks with collaborative signals	LLaMA-2-7B
SIIT (Chen et al., 2024b)	Retrieval	historical interactions	Self-adaptive tokenization	LLaMA-2-7B

methods (Bao et al., 2025; Cui et al., 2022; Tan et al., 2024; Zhang et al., 2026) use textual attributes such as titles and descriptions. However, these are often too lengthy for autoregressive generation and fail to capture collaborative knowledge, limiting effectiveness in practice. (iii) Codebook-based tokenization: A more compact alternative represents items as sequences of discrete tokens from a shared vocabulary (Rajput et al., 2023; Hou et al., 2025a; Zheng et al., 2024b). This reduces vocabulary size and avoids overly long sequences while retaining partial semantics. For instance, Semantic IDs cluster items with similar attributes to share tokens, and ActionPiece (Hou et al., 2025b) compresses feature patterns into tags. Yet, these methods struggle to balance textual and collaborative semantics. (iv) Codebooks with collaborative signals: To bridge this gap, recent studies integrate CF signals directly into tokenization. LETTER (Wang et al., 2024a)

Table 5
Unified training objectives for LLM-based generative recommendation.

Category	Representative Works	Formula
Supervised Fine-Tuning	P5 (Geng et al., 2022) LGIR (Du et al., 2024) LLM-Rec (Tang et al., 2024) RecRanker (Luo et al., 2024a)	$-\log \pi_\theta(y^+ \mid x)$
Self-Supervised Learning	FELLAS (Yuan et al., 2024) HFAR (Wang et al., 2025h)	$-\log \frac{\exp\left(\mathrm{sim}(y^+, y^-)/\tau\right)}{\sum_{y \in \mathcal{N}} \exp\left(\mathrm{sim}(y^+, y)/\tau\right)}$
Reinforcement Learning	LEA (Wang et al., 2024c) RPP (Mao et al., 2025b)	$-\left[r_\phi(x, y^+) - \beta\, D_{\mathrm{KL}}\left(\pi_\theta(y \mid x) \,\|\, \pi_{\mathrm{ref}}(y \mid x)\right)\right]$
Direct Preference Optimization	LettinGo (Wang et al., 2025k) RosePO (Liao et al., 2024a) SPRec (Gao et al., 2025a)	$-\log \sigma\left(\beta \log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\mathrm{ref}}(y^- \mid x)}\right)$

Notation: x is user profile/context; y^+/y^- is preferred/rejected item; π_θ is policy model; π_{ref} is reference model; $\mathrm{sim}(\cdot,\cdot)$ is similarity; τ is temperature; $\mathcal{N}$ is negative set; r_ϕ is reward; β is penalty/scale; D_{KL} is KL divergence; σ is sigmoid.

combines RQ-VAE with contrastive alignment to regulate semantics and collaboration; TokenRec (Qu et al., 2025a) quantizes masked user/item embeddings into discrete tokens, smoothly incorporating high-order collaborative knowledge; SETRec (Lin et al., 2025b) encodes histories as unordered sets; CCFRec (Liu et al., 2025e) enhances modeling with code masking and alignment; and LLM2Rec (He et al., 2025) injects CF signals via supervised fine-tuning and embedding modeling. (v) Self-adaptive tokenization via LLMs: The latest direction allows LLMs themselves to refine identifiers during training. SIIT (Chen et al., 2024b), for example, enables self-tuning of item tokens, aligning them with internal representations and mitigating inconsistencies caused by external tokenizers. Item tokenization provides a scalable solution to unify textual and collaborative semantics, but designing tokens that balance both efficiently remains an open challenge.

4.1.3. Training objective & inference

While aligning LLMs with recommendation data improves their adaptability, model performance still hinges on how training and inference are formulated. To this end, researchers have proposed a series of objectives and strategies specifically designed for recommendation scenarios.

Training objective. In recommender systems, the training objective is typically next-item prediction, where models infer the most probable subsequent item given a user's interaction history. To align large language models (LLMs) with this objective, prior work explores four training paradigms, each addressing a distinct limitation. In Table 5, we list representative works and formulas for these four training paradigms.

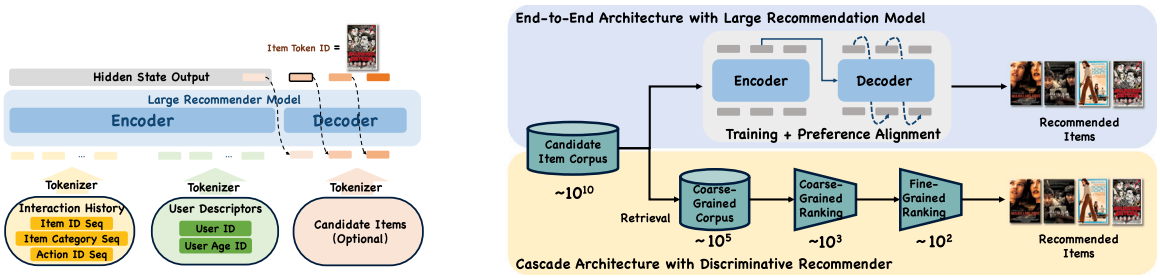
- Supervised Fine-Tuning (SFT). To learn next-item prediction, LLMs are fine-tuned with predefined templates that specify the task, encode interaction histories, and standardize outputs (Geng et al., 2022; Tang et al., 2024; Luo et al., 2024a). For example, P5 (Geng et al., 2022) designs prompts for five representative tasks, and LGIR (Du et al., 2024) augments SFT with a GAN-based module to improve robustness in few-shot settings. Despite strong results, SFT mainly learns from positive pairs and lacks explicit negatives, making it harder to learn ranking margins and to adapt to evolving user preferences.
- Self-Supervised Learning (SSL). SSL-based methods (Yuan et al., 2024; Wang et al., 2025h; Ren and Huang, 2024) reduces reliance on manual templates by generating auxiliary training signals. For example, EasyRec (Ren and Huang, 2024) builds a text–behavior alignment objective that combines contrastive learning with collaborative language-model tuning, enabling the LLM to separate candidates without explicit labels and improving zero-shot generalization.
- Reinforcement learning (RL). RL-based methods introduces reward-driven optimization over ranked sessions to model negatives and handle non-differentiable metrics. LEA (Wang et al., 2024c) learns user states with task-specific rewards, and RPP (Mao et al., 2025b) shapes rewards from downstream metrics to directly optimize recommendation goals. While effective, RL often needs large-scale feedback and can be costly and unstable to train.
- Preference Optimization (PO). To avoid training a reward model and reduce the instability of RL, PO-base (Wang et al., 2025k; Xu et al., 2026a,b; Liao et al., 2024a; Gao et al., 2025a) methods directly optimizes on preference pairs. In recommendation, RosePO (Liao et al., 2024a) tailors preference construction and applies a DPO-style objective to align the model with ranking goals without explicit rewards. SPRec (Gao et al., 2025a) adopts a self-play pipeline to stabilize training and strengthen preference alignment without a separate reward model.

Inference. In recommender systems, direct inference lets the LLM directly generate the Top-K items or score candidates one by one, yielding the simplest pipeline and requiring no re-ranker or reward model. However, it is prompt-sensitive, hard to control for stability and diversity, lacks explicit ranking signals, and long histories inflate prompts and memory overhead. These limitations motivate two complementary lines of work: inference reranking and inference acceleration.

- Reranking. Improve output quality by injecting stronger ranking signals at inference: RecRanker (Luo et al., 2024a) adopts a two-stage pipeline (candidate retrieval followed by LLM-based reranking) with a hybrid objective that integrates pairwise and applies position shifting to mitigate input bias: LLM4Rerank (Gao et al., 2025b) frames inference as multi-node, multi-hop reasoning with goal guidance and history pools to trade off accuracy, diversity, and fairness; GFN4Rec (Liu et al., 2023a) employs GFlowNets to autoregressively generate lists, improving coverage and diversity.
- Acceleration. Reduce latency and memory by cutting LLM usage, shortening inputs, and speeding decoding (Wang et al., 2025d): FELLAS limits the LLM to produce item/sequence embeddings while a lightweight model makes final predictions (Yuan et al., 2024); Prompt Distillation (GenRec (Sun et al., 2023)) compresses long histories and distills to a small model so the LLM is triggered only when needed; AtSpeed (Lin et al., 2025c) applies speculative decoding and tree-based attention to achieve 2–2.5× speedups while preserving Top-K consistency.

4.2. Large recommendation model

The success of LLMs has established compelling scaling laws: model performance improves predictably with increased compute, data, and parameters following power-law relationships. LLM-based recommenders discussed above inherit the benefits of pre-trained language models whose scaling laws have been well-established, but this comes with the requirement of reformulating user behavior data as natural language sequences. Recently, a parallel research direction has emerged that forgoes LLM adaptation in favor of establishing native scaling laws tailored specifically for recommendation tasks (Zhai et al., 2024; Deng et al., 2025; Guo et al., 2025a; Rajput et al., 2023; Huang et al., 2025c; Zhou et al., 2025a; Guo et al., 2025b; Zhu et al., 2025). Rather than borrowing from language models, these approaches design specialized



(a) The Architecture of LRMs (b) End-to-End Recommendation

Fig. 7. Illustration of two research directions of large recommendation models. (a) The Architecture of LRMs, and (b) End-to-End Recommendation.

Table 6
Summary of the representative large recommendation model.

Methods	User Formulation		Architectures	Backbone
	Task	Historical Interactions		
LEARN (Jia et al., 2025)	Ranking	history interactions, user preference	Cascaded	Baichuan2-7B, Transformer
HLLM (Chen et al., 2024a)	Retrieval, Ranking	history interactions	Cascaded	TinyLlama-1.1B, Baichuan2-7B
KuaiFormer (Liu et al., 2024a)	Retrieval	history interactions	Cascaded	Stacked Transformer
SRP4CTR (Han et al., 2024)	Ranking	history interactions, user preference	Cascaded	FG-BERT
HSTU (Zhai et al., 2024)	Ranking	history interactions	Cascaded	Transformer
MTGR (Han et al., 2025)	Ranking	history interactions	Cascaded	Transformer
UniROM (Qiu et al., 2025c)	Ranking	history interactions	End-to-End	RecFormer
URM (Jiang et al., 2025a)	Ranking	history interactions, user preference	End-to-End	BERT
OneRec (Deng et al., 2025)	Generative Retrieval and Ranking	history interactions, user preference	End-to-End	Transformer
OneSug (Guo et al., 2025a)	Generative Retrieval and Ranking	history interactions, user preference	End-to-End	Transformer
EGA-V2 (Zheng et al., 2025b)	Generative Retrieval and Ranking	history interactions, user preference	End-to-End	Transformer

architectures optimized directly for user behavior data. We refer to this emerging paradigm as Large Recommendation Models (LRMs).

LRMs have emerged as an active research area in industry in recent years, enabled by the massive scale of user behavior data available in industrial settings. In recent few years, industrial recommendation teams have faces two critical bottlenecks: First, diminishing returns from discriminative recommendation models' complexity. Increasingly sophisticated architectures no longer yield significant performance improvements. For instance, while Alibaba's Deep Interest Network (DIN) (Zhou et al., 2018) achieved substantial gains, subsequent models like MIMN (Pi et al., 2019) showed progressively marginal improvements—a pattern observed across all major industry teams. Second, escalating costs of cascaded architectures. Multi-stage recommendation pipelines in industry, typically consisting of retrieval, coarse ranking, fine ranking, and re-ranking stages, have become increasingly complex, resulting in higher maintenance overhead and growing communication and caching costs between stages. To address these bottlenecks, recent industrial research has increasingly turned to LRMs along two primary directions as shown in Fig. 7. First, to overcome the diminishing returns from model complexity, researchers are developing novel generative training paradigms that fundamentally rethink how recommendation models learn from data. Second, to mitigate the costs of cascaded architectures, efforts focus on leveraging LRMs to construct end-to-end recommendation frameworks that bypass the traditional multi-stage pipeline altogether. In Table 6, we summarize some representative large recommendation models.

The Scaling Law of LRMs. Meta's proposed HSTU (Zhai et al., 2024) is groundbreaking work in large recommendation models, validating that the scaling law of LLM is equally applicable to recommendation systems. HSTU transforms the traditional discriminative CTR prediction task into a generative sequence modeling task. For each user, it unified multiple pointwise user–item interaction samples into a single user behavior sequence containing interacted items, interaction behaviors, user and items' categorical features. HSTU adopts a causal autoregressive modeling approach that takes ultra-long user sequences as input, unifying the retrieval and ranking tasks into a sequence generation problem. The output is a probability distribution over candidate items, supporting multi-task joint training. HSTU employs sequence lengths ranging from 1024 to 8192, which far exceeds the length that traditional discriminative models can handle. Moreover, from a performance perspective, there is a trend showing that longer sequences and more complex models lead to better results. HSTU validates that as model scale increases, its performance continues to improve, whereas traditional discriminative recommendation models plateau. Ultimately, HSTU reaches a parameter scale of 1.5 trillion, while discriminative recommendation models stagnate in effectiveness at around 200 billion parameters. Based on HSTU, several major tech companies have successively proposed LRMs. For example, Meituan proposed MTGR (Han et al., 2025), a generative ranking framework. MTGR incorporates cross features commonly used in discriminative recommendation systems to prevent information loss in the generative architecture. For sequence encoding, it combines Group LayerNorm and a dynamic hybrid masking strategy to enhance HSTU's learning effectiveness. Redbook proposed GenRank (Huang et al., 2025c), applied to their fine-grained ranking applications. HSTU concatenates user-interacted items and their corresponding actions sequentially in the input sequence, which doubles the sequence length. GenRank treats items as positional information and focuses on iteratively predicting actions associated with items, significantly reducing the input sequence length, making it suitable for resource-sensitive ranking scenarios.

End-to-End Recommendations. The new paradigm of large recommendation models also makes it possible to unify the recommendation framework. Simplifying the recommendation framework can reduce engineering costs, and merging cascade structures can unify model optimization objectives, bringing performance gains. Kuaishou's large recommendation model OneRec (Deng et al., 2025; Zhou et al., 2025a) has made an excellent attempt in this direction. OneRec uses an end-to-end generative recommendation model to replace the traditional retrieval-coarse ranking-fine ranking cascade architecture. The entire OneRec solution achieved a significant 1.68% improvement in the key online metric of total watch time. The system's computational resource utilization increased from 11% to 28.8%. Since there is no need for inter-layer communication and data caching, OneRec's runtime cost is only 10.6% of the cascade architecture, demonstrating that generative

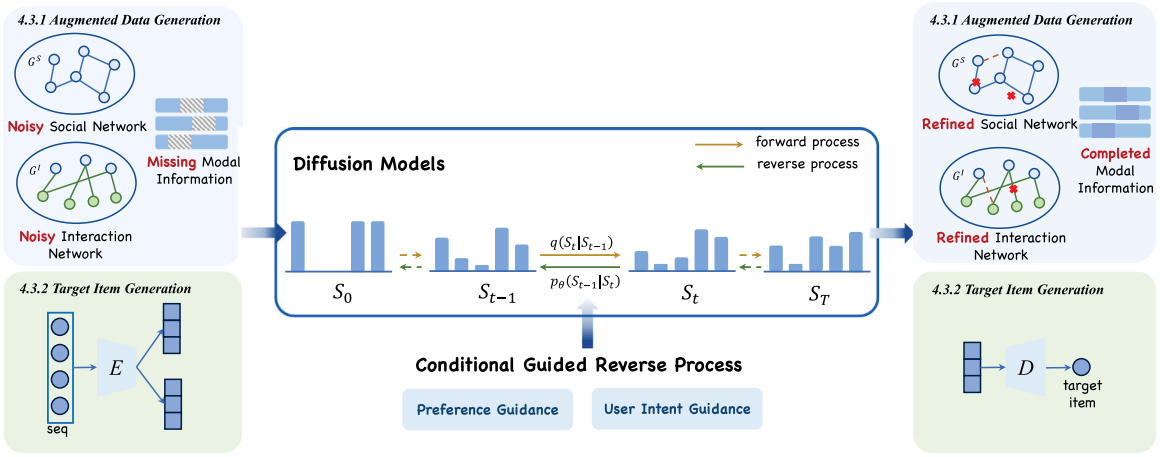


Fig. 8. Illustration of two types of diffusion-based generative recommendation. (a) Augmented Data Generation, and (b) Target Item Generation.

models have found new paradigms for both performance improvement and engineering optimization in recommendation systems. In terms of model architecture, OneRec adopts an encoder–decoder structure and uses the MoE architecture to expand model capacity and enhance user interest modeling capabilities. For generation methods, unlike traditional pointwise prediction, OneRec proposes a session-based generation approach that generates entire recommendation lists to better capture contextual information. In terms of training, OneRec adds a preference alignment phase, using Direct Preference Optimization (DPO) with a reward model to generate preference data and optimize generation results. OneSug (Guo et al., 2025a) extends this idea to query recommendation, unifying prefix, context, and history with RQ-VAE–based semantic IDs and Reward-Weighted Ranking to align with graded feedback. EGA-V2 (Zheng et al., 2025b) pushes this direction further by introducing hierarchical tokenization and multi-token prediction, integrating user interest modeling, POI generation, creative selection, ad allocation, and payment computation into a single generative framework. Together, these advances illustrate how end-to-end models are evolving from unifying retrieval and ranking to fully generative paradigms that integrate diverse recommendation tasks.

4.3. Diffusion-based generative recommendation

With the remarkable success of diffusion models (DM) in image synthesis, recent research has explored extending diffusion models to various recommendation tasks. DM-based methods can be categorized into two main types: approaches for augmented data generation and approaches for target item generation. (see Fig. 8)

4.3.1. Augmented data generation

The denoising characteristics and generative nature of diffusion models align well with obtaining high-quality user interaction data for robust recommendation systems, while their flexible conditional generation framework can effectively integrate user intentions and preferences for accurate personalized modeling. Based on these advantages, the ways to enhance traditional recommendations using diffusion models can be primarily categorized into three paradigms: generate high-quality interaction data, generate robust representations and preference injected conditional generation.

Generate high-quality interaction data. Some works (Di et al., 2025; Li et al., 2025f, 2023b) leverage the generation capabilities of the diffusion model to generate high-quality interaction data, alleviating the data sparsity problem of the recommendation system. DGFedRS (Di et al., 2025) pretrains diffusion model to capture latent personalized user information, generating high-quality interactions. MoDiCF (Li et al., 2025f) and TDM (Mao et al., 2025c) focus on data missing scenario. MoDiCF (Li et al., 2025f) designs a modality-aware diffusion module to generate and iteratively optimize missing data from the learned modality-specific distribution space. TDM (Mao et al., 2025c) simulates extra missing data in the guidance signals and allows diffusion models to handle existing missing data through extrapolation. Diffurec (Li et al., 2023b) considers user/item representation as distribution and adds Gaussian noise into embedding generation in the diffusion phase to inject uncertainty and diversity.

Generate robust representations. (Sun et al., 2025; Zhao et al., 2024d, 2025b) utilize the denoising capability of the diffusion model to inject controlled noise into the user and item representations, and remove the noise in the reverse iterative process. ARD (Sun et al., 2025) employs the diffusion process to refine social networks while DDRM (Zhao et al., 2024d) and DRGO (Zhao et al., 2025b) utilize diffusion model to learn robust representation.

Preference injected conditional generation. (Qu and Nobuhara, 2025; Li et al., 2025d,a) harness the information injection capabilities of diffusion models to formulate conditional probability generation tasks, thereby incorporating user intentions and preferences into the user data generation pipeline. DMCDR (Li et al., 2025d) leverages the preference guidance signals from the source domain to guide the reverse process and generate the personalized user representations in the target domain. In InDiRec (Qu and Nobuhara, 2025), conditional diffusion model is guided to generate forward views with same intent.

4.3.2. Target item generation

The objective of recommendation aligns well with diffusion model since it infers the future interaction probabilities based on noisy historical interactions. Therefore, several works (Wang et al., 2023b; Yang et al., 2023; Niu et al., 2024; Mao et al., 2025a; Cai et al., 2025; Xie et al., 2024b; Chen et al., 2025c; Liu et al., 2025c) explore diffusion models as recommenders.

Diffusion recommender model. DiffRec (Wang et al., 2023b) treats the prediction of users' interaction as a denoising process, it further build L-DiffRec and T-DiffRec to handle large-scale item prediction and temporal modeling challenges in generative recommendations. DreamRec (Yang et al., 2023) noises the target item to explore the underlying distribution of item space, and generate recommended items directly, thereby eliminating the need for negative sampling and enabling exploration of the entire item space. DiffRIS (Niu et al., 2024)uses both local and global implicit features from user historical sequences as conditional guidance, directing the denoising process of the diffusion model to generate more personalized recommendation results. DiQDiff (Mao et al., 2025a) enhances the robustness of guidance information for heterogeneous and noisy sequences through semantic vector quantization, and introduces contrastive discrepancy

maximization to distinguish the denoising trajectories of different users for personalized recommendation generation. HorizonRec (Zha et al., 2026) proposes an align-for-fusion framework for cross-domain sequential recommendation, replacing the conventional align-then-fusion paradigm with dual-oriented diffusion models that harmonize triple-domain preferences at a fine-grained level.

Diversity and uncertainty modeling. Although the above models have achieved competitive results, they encode user preference in the form of deterministic embeddings and rely on a homogeneous diffusion inference mechanism. This conveys a limited range of information and is insufficient to capture the diversity of user preferences. Inspired by this, DiffDiv (Cai et al., 2025) designs diversity-aware guided learning mechanism to guide the training of the diffusion model, enabling it to effectively capture users' diverse preferences.

Tailored Optimization for DM-based recommendation. DDSR (Xie et al., 2024b) employs discrete diffusion to construct fuzzy sets of interaction sequences to capture the evolution of users' interests. Some works focus on the embedding collapse problem of applying diffusion models to recommender systems, ADRec (Chen et al., 2025c) and PreferDiff (Liu et al., 2025c) argues that traditional objectives fail to leverage limits the full utilization of the generative potential of diffusion models, and propose a tailored optimization objective for diffusion based recommendation models.

5. Task-level opportunities

5.1. Top-K recommendation

Traditional discriminative recommendations calculate a preference score for each candidate item one-by-one, and then rank them to generate recommendations. In contrast, generative recommenders can directly generate item. However, to ensuring that the recommendation items are mapped to valid items, the generative models perform generation grounding during the inference stage. Several typical strategies have been developed: (i) Vocabulary-Constrained Decoding restricts the decoding space of the generative model to a predefined vocabulary of item identifiers or tokens. This guarantees that every generated token corresponds to an actual item. For example, P5 (Geng et al., 2022) utilized constrained decoding with a predefined item vocabulary and applied beam-search, ensuring outputs correspond to valid catalog items. IDGenRec (Tan et al., 2024) utilized a prefix tree to store all generated candidate IDs. Each newly generated token is constrained by the previously generated tokens, ensuring that the generation process only considers tokens that can potentially form an existing candidate ID in the dataset. Chu et al. (2023) and Hua et al. (2023b) proposed to utilize the Trie algorithm for constrained generation, where the generated identifier is guaranteed to be a valid identifier. However, Trie strictly generates the valid identifier from the first token, where the accuracy of the recommendation depends highly on the accuracy of the first several generated tokens. To combat this issue, TransRec (Lin et al., 2024b) utilized FM-index to achieve position-free constrained generation, which allows the generated token from any position of the valid identifier. The generated valid tokens will then be grounded to the valid identifiers through aggregations from different views. TransRec also introduced multi-facet identifiers (ID, title, attributes) and an Aggregated Grounding Module to map the generated identifiers back to in-corpus items. (ii) Post-Generation Filtering. Post-generation filtering lets an LLM freely generate text (IDs, titles, or semantic tokens) and then maps/reranks those outputs to in-catalog items via exact/semantic matching or reranking. For example, BIGRec (Bao et al., 2025) proposed to ground the generated identifier to the valid items via L2 distance between the generated token sequences' representations and the item representation. These approaches are efficient and scalable, but heavily depend on the quality of item embeddings. (iii) Prompt Augmentation. For LLM-based recommenders, this strategy injects the candidate item set into text prompts and asks the model to recommend items from the candidate set. Many LLM-based recommenders adopt this strategy, such as LLaRA (Liao et al., 2024b), A-LLMRec (Kim et al., 2024), iLoRA (Kong et al., 2024).

5.2. Personalized content generation

In addition to the typical recommendation that requires valid generation to recommend existing items to users, another research direction capitalizes on the generative capabilities of models to create entirely new content tailored to individual users. It is important to clarify how this direction differs from general AI-generated content (AIGC): while general AIGC focuses on open-ended content creation (e.g., text-to-image synthesis from arbitrary prompts), personalized content generation in recommendation is fundamentally preference-conditioned and task-situated. Specifically, the generation process is guided by the personalized user preferences, and its primary objective is to enhance the recommendation experience, rather than producing standalone creative works. We organize existing efforts into personalized visual content generation and personalized textual content generation. (i) Personalized visual content generation. Many Recent works leverage diffusion models to generate personalized recommendation content. For example, DiFashion (Xu et al., 2024) generates personalized outfit combinations based on a user's style preferences, which can serve both as direct recommendations and as guidance for fashion production. With the advancement of large-scale model technologies, recent research work focuses on leveraging diffusion models to generate virtual visualizations of recommendation results within the recommendation process. DreamVTON (Xie et al., 2024a) addresses the multi-view consistency problem in personalized diffusion models for 3D generation through a template-driven optimization mechanism and normal-style LoRA, achieving high-quality 3D virtual try-on based solely on person images, garment images, and text prompts. InstantBooth (Shi et al., 2024b) introduces a concept encoder and patch encoder to learn global concept embeddings and rich local patch features respectively, and incorporates novel adapter layers and concept token normalization techniques to enable personalized image generation. OOTDiffusion (Xu et al., 2025a) designs an outfitting UNet to learn detailed garment features and employs outfitting fusion to precisely align garment features with the human body in self-attention layers, while introducing outfitting dropout to implement classifier-free guidance, thereby generating high-quality, controllable virtual try-on images without requiring explicit deformation processes. (ii) Personalized textual content generation (Xu et al., 2025d). Some works leveraged realistic user interaction to explore personalization for review generation (Ni et al., 2019; Li et al., 2020; Sun et al., 2020; Li et al., 2021) and news headline generation (Ao et al., 2021; Cai et al., 2023; Song et al., 2023). For example, Ao et al. (2021) presents a personalized headline generation benchmark by collecting user's click history on Microsoft News.

As generative models become increasingly capable, the boundary between selecting existing items and creating new personalized content is progressively blurring. This convergence positions personalized content generation as a distinctive frontier of generative recommendation.

5.3. Conversational recommendation

Conversational recommendation is capable of eliciting the dynamic preferences of users and taking actions based on their current needs through real-time multi-turn interactions using natural language. The previous works mostly focus on single-turn LLM recommendation personalized based on pre-existing user history and item text. Nowadays, LLMs provide new opportunities for multi-turn conversational recommender systems (CRSs) where each turn presents a chance for the user to clarify or revise their preferences, critique and ask questions about recommended items, or convey a variety of other real-time intents.

In this subsection, we provide a taxonomy of recent research on LLM-based conversational recommendation, categorizing them into five methodological directions. (i) Prompting and Zero-shot Methods. One of the earliest applications of LLMs to CRS is through prompting, where task-specific templates or demonstration examples are crafted to steer LLMs toward generating recommendations. These methods often

exploit the zero-shot or few-shot capabilities of LLMs. For instance, He et al. (2023) showed that off-the-shelf LLMs can outperform supervised CRS baselines without fine-tuning. Later studies such as Spurlock et al. (2024) incorporated user feedback iteratively into prompts, improving personalization over multiple turns. Dao et al. (2024) highlighted the effectiveness of combining demonstrations or structured contexts to guide the model. Chen and Sun (2023) structured dialogue context into prompts for better recommendations. (ii) Retrieval-augmented and Knowledge-enhanced Approaches. A major limitation of purely prompt-based methods is the lack of grounding in the item space, which can lead to hallucinations. To address this, recent works combine LLMs with retrieval modules or knowledge graphs. For example, Qiu et al. (2025a) enhanced recommendations by retrieving relevant items and entities. Yang and Chen (2024) integrated collaborative filtering signals into the retrieval process, while Qiu et al. (2025b) focused on jointly modeling semantic and structural information. These retrieval-enhanced approaches improve factual grounding and personalization but come at the cost of higher latency and knowledge maintenance overhead. (iii) Beyond prompting and retrieval, researchers have developed unified and parameter-efficient architectures. Ravaut et al. (2024) reformulated CRS as a single NLP task. MemoCRS (Xi et al., 2024b) introduced memory modules to capture sequential coherence, while Yang et al. (2025a) represented items directly in natural language. In parallel, Chat-REC (Gao et al., 2023b) enhanced interaction and explainability. (iv) Evaluation has also received attention. Yang et al. (2024a) proposed assessing whether system strategies align with human expectations rather than relying only on accuracy metrics. This line of work points to the need for new evaluation protocols, bias mitigation, and human-in-the-loop assessments.

5.4. Explainable recommendation

Explainable recommendation has demonstrated significant advantages in informing users about the logic behind recommendations, thereby increasing system transparency, effectiveness, and trustworthiness. Early works focus on generating explanations with predefined templates or extracting logic reasoning rules from recommendation models. With the advance of LLMs, a burgeoning body of research started to build LLM-based explainable recommendation models. (i) P5 (Geng et al., 2022) mainly focus on designing prompts to guide LLMs to directly generate explanations. However, LLMs face the hallucination problem, resulting in generating low-quality explanations. To address these challenges, LLM2ER (Yang et al., 2024c) fine-tuned LLM-based explainable recommendation backbone in a reinforcement learning paradigm with two explainable quality reward models. (ii) Recently, researchers have tried to combine the advantages of graphs to enhance the explanations generated by LLMs. XRec (Ma et al., 2024) adopts a graph neural network (GNN) to model the graph structure and generate embeddings. Then, the embeddings are fed into LLMs to generate explanations. G-Refer (Li et al., 2025g) leveraged a hybrid graph retrieval to extract explicit CF signals, employing knowledge pruning to filter out less relevant samples, and utilizing retrieval-augmented fine-tuning to integrate retrieved knowledge into explanation generation. (iii) Some existing works (Yu et al., 2025; Zhao et al., 2025a) on large model-based recommendation leverage the reasoning abilities of thinking models and use the thought process as an explanation for recommendations. The challenge in this type of work lies in how to construct the ground truth for the explanation.

5.5. Recommendation reasoning

Recent advances in LLMs have led to remarkable progress in reasoning capabilities, with representative models including DeepSeek-R1 (Guo et al., 2025d), GPT o-series, etc. LLMs achieve complex reasoning through various prompting techniques, with CoT prompting (Wei et al., 2022) serving as the foundational approach that decomposes problems into intermediate reasoning steps. This has inspired numerous extensions, including zero-shot CoT (Kojima et al., 2022), self-consistency decoding (Wang et al., 2023a), and tree-of-thoughts (Yao et al., 2023). These techniques enable test-time scaling where additional computational budget during inference improves performance. Recent work has shifted focus from prompting to post-training enhancement of reasoning capabilities by using Reinforcement Learning techniques. Reasoning capability has recently garnered significant attention in RSs research. Reasoning-based RSs aim to perform multi-step deduction to deliver more accurate and explainable recommendations. Existing studies can be broadly grouped into three categories. (i) Explicit reasoning methods generate explicit and human-readable reasoning processes. Reason4Rec (Fang et al., 2025) introduced the task of deliberative recommendation, which requires LLMs to explicitly reason before predicting user feedback. It further develops a multi-expert framework consisting of preference distillation, preference matching, and feedback prediction. Reason-to-Recommend (Zhao et al., 2025a) proposed Interaction-of-Thought (IoT) reasoning, which organized user–item interaction chains into structured reasoning paths and trains LLMs via a two-stage SFT and RL paradigm. Similarly, ThinkRec (Yu et al., 2025) leveraged specially designed thinking prompts to induce LLMs to generate reasoning chains that enhance both recommendation quality and explainability. These works highlight the benefits of making the reasoning process explicit and transparent. OneRec-Think (Liu et al., 2025a) activated the LLM-based recommender's reasoning capabilities through CoT-based SFT. To generate effective CoT examples, the framework first extracted coherent reasoning trajectories from pruned user contexts, then leveraged these trajectories to guide CoT generation over raw behavioral data, enabling effective contextual distillation for noisy industrial settings. Finally, OneRec-Think optimized reasoning quality using recommendation-specific RL reward functions. (ii) Implicit reasoning methods perform latent reasoning without textual interpretability. LatentR^3 (Zhang et al., 2025c) introduced reinforced latent reasoning, which encodes reasoning processes into a compact sequence of latent tokens rather than generating lengthy textual CoTs. Through supervised fine-tuning followed by RL optimization, LatentR^3 achieves both efficiency and effectiveness, offering a practical paradigm for large-scale online recommendation scenarios. ReaRec (Tang et al., 2025) introduced an inference-time computing framework that enabled traditional sequential recommenders to perform multi-step latent reasoning by autoregressively processing hidden states with reasoning position embedding. STREAM-Rec (Zhang et al., 2025d) introduced a slow thinking paradigm for sequential recommendation, enabling traditional lightweight recommenders to perform multi-step deliberative reasoning rather than one-step direct matching. The core innovation is an iterative residual-based reasoning method, where the model progressively refines predictions by computing and fitting the residual between its current output and the target behavior representation, with these intermediate residual corrections serving as reasoning paths. (iii) LLM reasoning augmentation methods leveraged LLMs to generate reasoning steps that enhance the training of traditional RSs. For example, DeepRec (Zheng et al., 2025a) proposed an autonomous interaction paradigm where the LLM hypothesizes user preferences, queries a traditional RS for candidates, and iteratively refines its reasoning before final recommendation. LLMRG (Wang et al., 2024b) constructed reasoning graphs through LLM-driven chain reasoning, divergent extension, and self-verification, and then integrated them into recommendation models via graph neural networks.

6. Discussion and open challenges

6.1. Discussion

In the previous sections, we summarized current generative recommendation methods from the aspects of data, model, and task. We can observe that the emergence of generative models has led to a paradigm

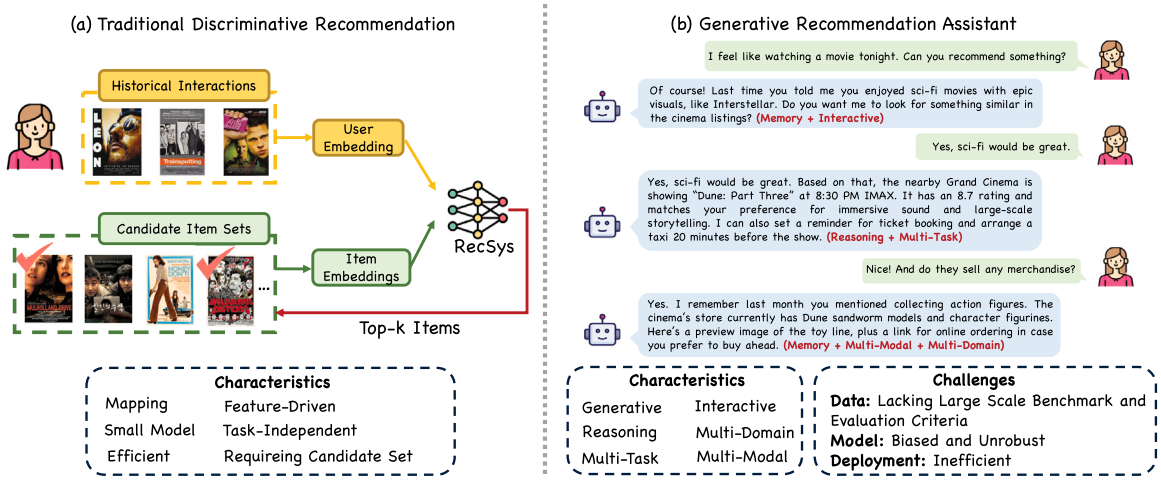


Fig. 9. Illustration of traditional discriminative recommendation and generative recommendation assistant.

shift in RSs. This shift makes the field more open and flexible across all three dimensions: data, model, and task.

Specifically, data-wise, traditional discriminative RSs relied on handcrafted features from domain-specific datasets. These features were manually curated for one specific and fixed task. Generative models, particularly LLMs, bring world knowledge into the recommendation process. Additionally, LLMs' ability to perform real-time internet searches introduces new opportunities for bringing massive, up-to-date world knowledge. Moreover, LLMs are naturally suited for integrating recommendation knowledge from multi-domain and tasks. In summary, generative models bring more knowledge to the data side.

On the model side, generative models have demonstrated the power of scaling laws. As model size, training data, and computational resources increase, recommendation performance improves significantly. Larger models also exhibit emergent capabilities. For example, instruction fine-tuning, RLHF, and in-context learning changes the way of training and inference. Reasoning ability enables the model to understand and infer complex relationships between users, items, and contexts in ways that smaller traditional models cannot. Moreover, unlike traditional discriminative recommenders that restrict a candidate set, generative models are capable of generating recommendations directly. In summary, large generative models bring stronger capabilities to the model side.

On the task side, traditional RSs were often designed to perform a single task, such as CTR prediction or rating prediction. While LLMs are task-agnostic by design, meaning they can handle a variety of recommendation tasks within a single framework. The same model can rank items, generate explanations, and produce personalized content in an interactive way. This flexibility allows the system to adapt dynamically to different user needs. In short, generative recommendation supports more diverse and flexible functionalities at the task dimension.

Generative RSs have broken away from the traditional discriminative paradigm, which were mapping-based, feature-driven, small-model, task-independent, and reliant on pre-defined candidate sets. Nowadays, generative recommenders are reshaping the recommendation paradigm. They are moving towards recommendation assistants that are open, capable of handling a wide range of tasks, and able to adapt to changing user needs. These systems can interact with users more flexibly, supporting multiple tasks such as ranking, explanation, content generation, and conversation. This shift represents a fundamental change in how recommendations are made, making the process more dynamic and responsive to diverse user requirements. The difference between the two paradigms is shown in Fig. 9.

While the paradigm shift from traditional recommendation systems to generative models opens up new possibilities, there is still a long way to go before realizing the full potential of recommendation assistants. The journey from the current state to the future vision requires overcoming several challenges. In the following sections, we will explore these challenges across three key dimensions: data, model, and deployment.

6.2. Open challenges

6.2.1. Data

Dataset. As a data-driven and user-centric research direction, the development of RS has always been closely tied to the availability and evolution of datasets. Key datasets have driven significant advancements in the field, providing benchmarks for training and evaluating algorithms and spurring new research directions. The MovieLens dataset played a foundational role in the early stage, providing large-scale rating data that enabled the development and evaluation of collaborative filtering and matrix factorization methods. Later, the release of the Netflix Prize dataset became a milestone that pushed forward latent factor models and matrix factorization techniques, demonstrating their effectiveness in large-scale recommendation tasks. The Amazon Review dataset further shifted attention from explicit ratings to implicit feedback such as clicks and purchases, and its combination of ratings, text, and product metadata fostered research on hybrid recommendation that integrates multiple signals. The Yelp dataset, with its rich user reviews and business information, advanced the use of deep learning-based recommendation, where natural language processing was leveraged for sentiment analysis, representation learning, and context-aware suggestions.

These datasets have greatly advanced the development of recommendation systems, serving as the foundation for many breakthroughs over the past two decades. However, for generative recommendation, they are no longer fully suitable. Most of them are non-interactive,offline, and static, capturing only point-in-time user preferences rather than the dynamic feedback loops and multi-round interactions that characterize real-world scenarios. On top of these datasets, most existing studies rely on fixed task-specific prompt templates to generate recommendations and use traditional metrics for evaluation. These datasets are more suitable for assessing the accuracy performance of traditional RSs, but they restrict the assessment of generative models as personalized assistants, which in reality need to operate across multiple scenarios and handle diverse tasks in interactive settings. This gap highlights the urgent need for new benchmarks that can better support the next stage of recommendation research.

Evaluation Metrics. To provide a clearer picture of the current evaluation landscape, Table 7 summarizes representative metrics used across different generative recommendation scenarios. These metrics

Table 7
Representative evaluation metrics for generative recommender systems, grouped by evaluation scenario.

Category	Metric	Formulation	Notes
Recommendation	NDCG@K	$\frac{\sum_{i=1}^{K} rel_i / \log_2(i+1)}{\sum_{i=1}^{K} rel_i^* / \log_2(i+1)}$	Position-aware ranking metric; $rel_i \in \{0,1\}$, rel_i^* is the ideal ranking. Penalizes relevant items ranked lower.
	Recall@K	$\frac{\sum_{i=1}^{K} rel_i}{\lvert\mathcal{G}\rvert}$	Fraction of ground-truth items $\mathcal{G}$ retrieved within top-K.
	Precision@K	$\frac{\sum_{i=1}^{K} rel_i}{K}$	Fraction of top-K items that are relevant. Complements Recall@K by penalizing irrelevant recommendations.
	MRR@K	$\frac{1}{\lvert\mathcal{U}\rvert} \sum_{u \in \mathcal{U}} \frac{1}{\text{rank}_u^*}$	Mean Reciprocal Rank; rank_u^* is the rank of the first relevant item for user u within top-K.
	HR@K	$\mathbf{1}\left(\sum_{i=1}^{K} rel_i > 0\right)$	Binary indicator of whether at least one relevant item appears in top-K; $\mathbf{1}(\cdot)$ is the indicator function.
	AUC	$\frac{\sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \mathbf{1}[s_i > s_j]}{\lvert\mathcal{P}\rvert \lvert\mathcal{N}\rvert}$	$\mathcal{P}/\mathcal{N}$: positive/negative item sets; s_i, s_j: predicted scores. Threshold-independent and robust to class imbalance.
	CTR	$\frac{\#\ \text{Clicks}}{\#\ \text{Impressions}}$	Click-through rate; most direct signal of short-term user engagement.
	CVR	$\frac{\#\ \text{Conversions}}{\#\ \text{Clicks}}$	Conversion rate from click to target action (e.g., purchase); captures downstream business value.
Content Gen. Quality	BLEU	$\begin{cases} \frac{\lvert S \cap R\rvert}{\lvert S\rvert} \cdot \exp\left(1 - \frac{r}{c}\right) & c < r \\ \frac{\lvert S \cap R\rvert}{\lvert S\rvert} & c \geq r \end{cases}$	S: generated text, R: reference, c/r: their lengths. Includes a brevity penalty when $c < r$.
	ROUGE-L	$\frac{LCS(S,R)}{\lvert R\rvert}$	Longest common subsequence between S and R normalized by $\lvert R\rvert$. Captures structural similarity beyond n-gram matching.
	SBERT	$\frac{\mathbf{v}_S \cdot \mathbf{v}_R}{\lVert\mathbf{v}_S\rVert \cdot \lVert\mathbf{v}_R\rVert}$	Cosine similarity of sentence embeddings $\mathbf{v}_S$, $\mathbf{v}_R$. Captures semantic similarity without requiring lexical overlap.
	LLM-E	$\text{LLM}(S, R)$	An LLM judge scores S against R on relevance, fluency, and factuality.
	FID	$\lVert\mu_r - \mu_g\rVert^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right)$	Fréchet Inception Distance for visual content; $(\mu_r, \Sigma_r)/(\mu_g, \Sigma_g)$: Inception feature statistics of real/generated images. Lower is better.
Diversity	ILD	$\frac{1}{K(K-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{K} d(\mathbf{v}_i, \mathbf{v}_j)$	Intra-List Diversity; $d(\cdot,\cdot)$ is embedding distance between items in the list. Higher values indicate more diverse recommendations.
	Coverage	$\frac{\lvert\bigcup_{u \in \mathcal{U}} \mathcal{L}_u\rvert}{\lvert\mathcal{I}\rvert}$	Fraction of catalog $\mathcal{I}$ covered by recommendation lists $\{\mathcal{L}_u\}$. Measures ability to surface long-tail items.
	Novelty	$\frac{1}{K} \sum_{i=1}^{K} -\log_2 p(i)$	$p(i)$: popularity of item i in training data. Rewards for recommending less-known items.
Fairness	DP	$\max_{g,g' \in \mathcal{G}} \left\lvert P_g(\hat{y}{=}1) - P_{g'}(\hat{y}{=}1)\right\rvert$	Demographic Parity gap across sensitive groups $\mathcal{G}$ (e.g., gender, age). Zero indicates equal recommendation rates across groups.
	EO	$\max_{g,g' \in \mathcal{G}} \left\lvert P_g(\hat{y}{=}1 \mid y{=}1) - P_{g'}(\hat{y}{=}1 \mid y{=}1)\right\rvert$	Equal Opportunity gap; compares true positive rates across sensitive groups. Zero indicates relevant items are surfaced equally regardless of group.
Conv. Rec.	SR	$\frac{1}{\lvert\mathcal{D}\rvert} \sum_{d \in \mathcal{D}} \mathbf{1}(\text{success}_d)$	Success Rate over dialogues $\mathcal{D}$; success_d indicates a relevant item was recommended within the allotted turns.
	AT	$\frac{1}{\lvert\mathcal{D}\rvert} \sum_{d \in \mathcal{D}} T_d$	Average Turns to success; T_d is the turn count of session d. Fewer turns indicate more efficient preference elicitation.

can be broadly grouped into five categories: ranking-oriented metrics (e.g., NDCG@K, Recall@K, HR@K) that assess the accuracy of top-K item retrieval; content generation quality metrics (e.g., BLEU, ROUGE-L, SBERT, LLM-E, FID) that evaluate the fluency, semantic fidelity, and visual realism of generated content; diversity metrics (e.g., ILD, Coverage, Novelty) that measure the breadth and freshness of recommendation lists; fairness metrics (e.g., DP, EO) that quantify demographic parity across sensitive user groups; and conversational metrics (e.g., Success Rate, Average Turns) that capture the efficiency of preference elicitation in multi-turn dialogues. While these metrics collectively cover a wide range of evaluation dimensions, they each carry inherent limitations when applied to generative recommendation systems. Ranking metrics such as NDCG@K and HR@K presuppose a fixed candidate set and binary relevance labels, making them ill-suited for open-ended generation scenarios where the output is not constrained to a predefined item pool. Content generation metrics like BLEU and ROUGE-L rely on lexical overlap with reference texts, which poorly reflects semantic quality in free-form personalized generation. Although LLM-based evaluation (LLM-E) offers a more flexible alternative, it introduces non-determinism and incurs substantial computational cost at scale. Diversity and fairness metrics, while important, are rarely jointly optimized or reported alongside accuracy metrics in existing benchmarks, leaving a gap in holistic evaluation. Conversational metrics such as Success Rate and Average Turns depend heavily on the design of dialogue simulators, whose fidelity to real user behavior remains questionable.

Beyond these metrics, the agent-based simulation works reviewed in Section 3.1.2 open up new possibilities for recommendation evaluation: by generating diverse and controllable interaction signals, they offer a promising alternative to static offline benchmarks, enabling

dynamic, multi-round, and cost-effective evaluation that better reflects real-world recommendation scenarios. However, despite this promise, their validity as behavioral proxies remains an open question. On one hand, the behavioral fidelity of LLM-based agents is inherently constrained by a chicken-and-egg dilemma: building accurate user simulators requires large-scale real user behavior data, yet the very motivation for simulation is to compensate for the lack of such data. As a result, simulated behaviors risk reflecting the distributional biases of pretraining corpora rather than the true diversity of real user preferences. On the other hand, most existing simulation frameworks model users as static agents with fixed personas, failing to capture the dynamic nature of real user behavior, such as preference drift, mood-driven decisions, and evolving interests over time. Taken together, the fragmented and task-specific nature of existing evaluation protocols makes it difficult to assess generative recommendation systems as unified, multi-capable assistants. This underscores the urgent need for new benchmarks that support dynamic, multi-task, and interactive evaluation aligned with the full potential of generative models.

6.2.2. Model

At the model level, generative RSs face significant challenges in terms of bias and robustness.

Bias. We categorize the main biases in existing generative RSs into three types. (i) The first is popularity bias. The training data bias in generative recommendation primarily arises from two aspects: the biases present in large-scale pretraining corpora and those found in sparse user interaction. The ranking results of LLMs are influenced by the popularity of items in user interaction data, and popular items that are frequently discussed and mentioned in the LLM's pretraining corpus tend to be ranked higher. This may lead to a lack of diversity in the responses and potentially marginalize less popular or minority viewpoints. Existing methods (Li et al., 2025c; Gao et al., 2025c; Xi et al., 2025) mitigate the fairness issue from the perspective of adjusting or generating training data distribution. Although effective, further research is needed in utilizing the generative capabilities of large models to design prompts that extract or generate fair user interactions. Additionally, existing training methods may exacerbate popularity bias (Zhang et al., 2025b). SFT maximizes the likelihood function, is prone to overfitting on popular items, thereby amplifying popularity bias and reducing the diversity of recommendation results. DPO, on the other hand, tends to suppress the probability of non-preferred responses, reinforcing existing patterns in the data while being highly sensitive to the quality of preference pairs. (ii) The second is fairness. Generative model exhibits fairness issues related to sensitive attributes (Hou et al., 2025c), as the model implicitly utilizes demographic characteristics of individuals involved in certain task annotations from user interaction data or pretraining data. This may lead the model to make recommendations based on assumptions that users belong to specific groups (Hua et al., 2023a; Xu et al., 2025b; Xin et al., 2025b), such as biases in recommendations due to gender or race. Addressing these fairness issues is critical and necessary to ensure fair and unbiased recommendations. Determining whether the model can utilize or infer sensitive attributes of users, and reducing the overlap between user preference modeling and the representation of sensitive user information, is key to mitigating fairness issues on the user side. (iii) The third is position bias. In the generative recommendation modeling paradigm, user behavior sequences and recommended candidate items are input into the language model in the form of text sequences, which may introduce certain positional biases inherent to the language model itself (Kweon et al., 2025; Li et al., 2024b). LLMs are highly sensitive to the structure and content of input prompts, with even minor changes potentially leading to significant differences in the output. Variations in prompt design, such as the complexity of product descriptions, the number of candidate items, or their order, can influence the model's attention distribution, thereby increasing the uncertainty of the prompt. For instance, LLMs typically prioritize higher-ranked items.

Robustness. For robustness, we categorize the main challenges into robustness against natural noise and robustness against malicious attack. (i) For robustness against natural noise, recommendation tasks have long been plagued by noise issues, such as user interactions being influenced by clickbait or disrupted by unintended interactions. The challenge of denoising lies in the lack of clear noise signals, with existing denoising methods largely relying on joint training guided by the recommendation objective (Liao et al., 2025; Yang et al., 2025b). With the rise of generative recommendation, we aim to leverage the built-in robustness of LLMs, including their extensive open knowledge and advanced semantic reasoning capabilities, to identify noisy interactions and provide more reliable suggestions (Wang et al., 2024f, 2025l; Peng et al., 2025; Song et al., 2024; Wang et al., 2025c). However, directly applying LLMs to denoising presents several challenges: there is a significant gap between the pretraining objectives of LLMs and the specific requirements of recommendation denoising, and direct denoising prompts often result in completely incorrect noise classification outcomes. Even with fine-tuning, LLMs can summarize user preferences from interaction items, but the presence of false positive samples in historical interactions can cause biases in the summary of user preferences, thus reducing the effectiveness of noise sample identification (Song et al., 2024; Wang et al., 2025c). Additionally, LLMs are prone to the notorious hallucination phenomenon, where they may incorrectly label non-existent interactions as noise. Therefore, it is essential to consider how to constrain the knowledge generated by LLMs to align with the specific prediction objectives in recommendation tasks. (ii) For robustness against malicious attack (Yang et al., 2026; Liu et al., 2026), injection attacks represent a competitive approach to attacking traditional recommender systems (Nguyen et al., 2024; Gunes et al., 2014). While effective, it is costly due to the need to inject a substantial number of malicious profiles. Such large-scale injections are infeasible due to limited attack budgets (Wang et al., 2025b). In contrast, due to the sensitivity of LLMs to text description, textual simulation attack is a potential direction. This attack directly rephrases the original item description of target items into an untrue crafted version. The adversarial texts are semantically similar to the originals, and align with the stylistic patterns of the system, all while exerting minimal impact on overall recommendation performance (Wang et al., 2025b; Zhang et al., 2024d; Ning et al., 2024, 2025). Compared to traditional recommendation attacks, textual attacks are remarkably low-cost, and can be executed under black-box settings, lowering the barrier to attack and raising the practical risk. Moreover, the malicious texts generated through such attacks exhibit transferability across different LLM-based recommendation models and tasks, meaning that adversarial prompts tailored to a single model may effectively compromise multiple systems or applications. As researchers increasingly focus on multi-modal LLM-based generative recommendations (Ye et al., 2025; Zhang et al., 2025e), the robustness of these models against attacks has emerged as a potential research direction. Furthermore, although studies on attacks targeting LLM-based generative recommendations are still in their preliminary stages, existing defense strategies exhibit limited effectiveness against data poisoning attacks (Wang et al., 2025b).

6.2.3. Deployment

Beyond the challenges at the data and the model level, deploying generative RSs in real-world scenarios introduces difficulties of efficiency. Unlike traditional recommender models, generative models are usually large and computationally intensive, which makes their deployment far from trivial.

Training Efficiency. Although parameter-efficient fine-tuning (PEFT) methods reduce the training cost of LLM-based generative recommendations by updating only a subset of model parameters, they remain insufficient to address the challenges posed by the rapidly increasing scale of recommendation datasets. The key to improving fine-tuning efficiency lies in rapidly adapting large language models to perform effectively in recommendation tasks using fewer data and

computational resources. One intuitive approach is to select a small, representative subset of data (Lin et al., 2024c; Mei et al., 2025). However, such methods typically require computing the influence or gradient information for each sample, which incurs significant computational complexity. Moreover, selections based on individual interactions often overlook the correlations between interaction records in recommendation tasks, leading to suboptimal results.

Inference Efficiency. Generative recommendation faces a significant issue of inefficient inference: its inference process, due to the nature of autoregressive decoding, requires multiple serial calls to the LLM, resulting in excessive time consumption, which hinders its practical application in real-time recommendation scenarios. How can we leverage the superior performance of LLM-based recommenders while maintaining inference latency as low as traditional recommenders? Unlike natural language processing (NLP) tasks, recommendation tasks require the generation of the top-K distinct sequences (recommended item lists), which necessitates beam search during the decoding process to ensure the generation of diverse and high-quality recommendation results. This makes traditional language model inference acceleration methods, such as speculative decoding, difficult to apply directly (Lin et al., 2025c). Knowledge distillation (Cui et al., 2024; Xu et al., 2025c) alleviates this issue to some extent by transferring knowledge from a complex LLM-based recommendation model to a lightweight traditional recommender or a smaller language model.

7. Conclusion

This survey has systematically examined how generative models are revolutionizing recommender systems, marking a paradigm shift from discriminative matching to intelligent synthesis. Our investigation reveals a tripartite transformation. At the data level, generative models transcend traditional boundaries through knowledge augmentation and behavioral simulation. At the model level, LLM-based methods, large-scale architectures, and diffusion approaches establish powerful alternatives to conventional paradigms. At the task level, new capabilities emerge including conversational dynamics, transparent reasoning, and personalized content creation, fundamentally redefining user-system interaction. Critical challenges remain: developing dynamic benchmarks that capture real-world complexity, fortifying models against biases and attacks, and achieving deployment efficiency. We envision recommendation systems evolving into cognitive assistants that are transparent, contextually aware, and seamlessly integrate reasoning and generation through natural language. This represents not incremental improvement but a fundamental reconceptualization of how intelligent systems serve human information needs. As we stand at this inflection point, this survey serves as both a comprehensive synthesis of current achievements and a strategic roadmap for future research, guiding the community toward realizing the full potential of generative intelligence in recommendation systems.

CRediT authorship contribution statement

Min Hou: Writing – original draft, Data curation, Conceptualization. **Le Wu:** Writing – review & editing, Supervision, Conceptualization. **Yuxin Liao:** Writing – original draft, Data curation. **Yonghui Yang:** Writing – original draft, Conceptualization. **Zhen Zhang:** Writing – original draft, Data curation. **Yu Wang:** Writing – original draft. **Changlong Zheng:** Visualization, Data curation. **Han Wu:** Writing – review & editing. **Richang Hong:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. U23B2031, 62436003, and 62402159), the Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGPB0248).

References

Ao, X., Wang, X., Luo, L., Qiao, Y., He, Q., Xie, X., 2021. PENS: A dataset and generic framework for personalized news headline generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 82–92.

Bai, T., Huang, L., Yu, Y., Yang, C., Hou, C., Zhao, Z., Shi, C., 2025. Efficient multi-task prompt tuning for recommendation. ACM Trans. Inf. Syst..

Bao, K., Zhang, J., Wang, W., Zhang, Y., Yang, Z., Luo, Y., Chen, C., Feng, F., Tian, Q., 2025. A bi-step grounding paradigm for large language models in recommendation systems. ACM Trans. Recomm. Syst. 3 (4), 1–27.

Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., He, X., 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 1007–1014.

Bougie, N., Watanabe, N., 2025. Simuser: Simulating user behavior with large language models for recommender system evaluation. arXiv preprint arXiv:2504.12722.

Cai, P., Song, K., Cho, S., Wang, H., Wang, X., Yu, H., Liu, F., Yu, D., 2023. Generating user-engaging news headlines. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp. 3265–3280. http://dx.doi.org/10.18653/v1/2023.acl-long.183, URL https://aclanthology.org/2023.acl-long.183/.

Cai, Z., Wang, S., Chu, V.W., Naseem, U., Wang, Y., Chen, F., 2025. Unleashing the potential of diffusion models towards diversified sequential recommendations. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1476–1486.

Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., Leskovec, J., 2025. LLMs generate structurally realistic social networks but overestimate political homophily. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 19, pp. 341–371.

Chen, J., Chi, L., Peng, B., Yuan, Z., 2024a. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. arXiv preprint arXiv:2409.12740.

Chen, L., Gao, C., Du, X., Luo, H., Jin, D., Li, Y., Wang, M., 2025a. Enhancing ID-based recommendation with large language models. ACM Trans. Inf. Syst. 43 (5), 1–30.

Chen, J., He, J., Li, H., Wang, S., Cao, Y., Wei, K., Yang, Z., Ji, Y., 2025b. Hierarchical intent-guided optimization with pluggable LLM-driven semantics for session-based recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1655–1665.

Chen, R., Ju, M., Bui, N., Antypas, D., Cai, S., Wu, X., Neves, L., Wang, Z., Shah, N., Zhao, T., 2024b. Enhancing item tokenization for generative recommendation through self-improvement. arXiv preprint arXiv:2412.17171.

Chen, K., Sun, S., 2023. CP-Rec: Contextual prompting for conversational recommender systems. Proc. AAAI Conf. Artif. Intell. 37 (11), 12635–12643. http://dx.doi.org/10.1609/aaai.v37i11.26487, URL https://ojs.aaai.org/index.php/AAAI/article/view/26487.

Chen, J., Xu, Y., Jiang, Y., 2025c. Unlocking the power of diffusion models in sequential recommendation: A simple and effective approach. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 155–166.

Chen, J., Yang, X., Yang, C., Bao, J., Guo, Z., Li, Y., Shi, C., 2025d. CORONA: A coarse-to-fine framework for graph-based recommendation with large language models. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2048–2058.

Cheng, W., Qin, Z., Wu, Z., Zhou, P., Huang, T., 2025. Large language models enhanced hyperbolic space recommender systems. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1944–1953.

Chu, Z., Hao, H., Ouyang, X., Wang, S., Wang, Y., Shen, Y., Gu, J., Cui, Q., Li, L., Xue, S., Zhang, J.Y., Li, S., 2023. Leveraging large language models for pre-trained recommender systems. arXiv:2308.10837. URL https://arxiv.org/abs/2308.10837.

Corecco, N., Piatti, G., Lanzendörfer, L.A., Fan, F.X., Wattenhofer, R., 2024. Suber: An rl environment with simulated human behavior for recommender systems. arXiv preprint arXiv:2406.01631.

Cui, Y., Liu, F., Wang, P., Wang, B., Tang, H., Wan, Y., Wang, J., Chen, J., 2024. Distillation matters: empowering sequential recommenders to match the performance of large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 507–517.

Cui, Z., Ma, J., Zhou, C., Zhou, J., Yang, H., 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. arXiv preprint arXiv:2205.08084.

Dao, H., Deng, Y., Le, D.D., Liao, L., 2024. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, Association for Computing Machinery, New York, NY, USA, pp. 785–795. http://dx.doi.org/10.1145/3626772.3657755.

Deldjoo, Y., He, Z., McAuley, J., Korikov, A., Sanner, S., Ramisa, A., Vidal, R., Sathiamoorthy, M., Kasirzadeh, A., Milano, S., 2024. A review of modern recommender systems using generative models (gen-recsys). In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6448–6458.

Deng, J., Wang, S., Cai, K., Ren, L., Hu, Q., Ding, W., Luo, Q., Zhou, G., 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. arXiv preprint arXiv:2502.18965.

Di, Y., Shi, H., Wang, X., Ma, R., Liu, Y., 2025. Federated recommender system based on diffusion augmentation and guided denoising. ACM Trans. Inf. Syst. 43 (2), 1–36.

Du, Y., Luo, D., Yan, R., Wang, X., Liu, H., Zhu, H., Song, Y., Zhang, J., 2024. Enhancing job recommendation through llm-based generative adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (8), pp. 8363–8371.

Ebrat, D., Paradalis, E., Rueda, L., 2024. Lusifer: LLM-based user simulated feedback environment for online recommender systems. arXiv preprint arXiv:2405.13362.

Fang, Y., Wang, W., Zhang, Y., Zhu, F., Wang, Q., Feng, F., He, X., 2025. Reason4Rec: Large language models for recommendation with deliberative user preference alignment. arXiv preprint arXiv:2502.02061.

Fu, Z., Li, X., Wu, C., Wang, Y., Dong, K., Zhao, X., Zhao, M., Guo, H., Tang, R., 2025. A unified framework for multi-domain CTR prediction via large language models. ACM Trans. Inf. Syst..

Gao, C., Chen, R., Yuan, S., Huang, K., Yu, Y., He, X., 2025a. Sprec: Self-play to debias llm-based recommendation. In: Proceedings of the ACM on Web Conference 2025. pp. 5075–5084.

Gao, J., Chen, B., Zhao, X., Liu, W., Li, X., Wang, Y., Wang, W., Guo, H., Tang, R., 2025b. Llm4rerank: Llm-based auto-reranking framework for recommendations. In: Proceedings of the ACM on Web Conference 2025. pp. 228–239.

Gao, C., Gao, M., Fan, C., Yuan, S., Shi, W., He, X., 2025c. Process-supervised llm recommenders via flow-guided tuning. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1934–1943.

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y., 2023a. S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J., 2023b. Chat-rec: Towards interactive and explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524.

Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y., 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In: Proceedings of the 16th ACM Conference on Recommender Systems. pp. 299–315.

Gunes, I., Kaleli, C., Bilge, A., Polat, H., 2014. Shilling attacks against recommender systems: a comprehensive survey. Artif. Intell. Rev. 42 (4), 767–799.

Guo, X., Chen, B., Wang, S., Yang, Y., Lei, C., Ding, Y., Li, H., 2025a. OneSug: The unified end-to-end generative framework for E-commerce query suggestion. arXiv preprint arXiv:2506.06913.

Guo, L., Li, W., Liao, L., Cheng, H., Zhang, R., Shi, Y., Wang, Y., Huang, Y., Zhai, K., Wang, P., Shi, T., Cao, X., Wang, S., Cai, R., Gong, Z., Vichare, O., Jian, R., Gao, L., Deng, S., Liu, X., Zhang, X., Li, F., Xie, W., Wen, B., Li, R., Fang, L., Liu, X., Zhai, J., 2025b. Request-only optimization for recommendation systems. arXiv:2508.05640. URL https://arxiv.org/abs/2508.05640.

Guo, L., Song, C., Guo, F., Han, X., Chang, X., Zhu, L., 2025c. Semantic-enhanced co-attention prompt learning for non-overlapping cross-domain recommendation. ACM Trans. Inf. Syst..

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Liang, W., et al., 2025d. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Nature 645 (8081), 633–638. http://dx.doi.org/10.1038/s41586-025-09422-z, URL https://www.nature.com/articles/s41586-025-09422-z.

Han, R., Li, Q., Jiang, H., Li, R., Zhao, Y., Li, X., Lin, W., 2024. Enhancing CTR prediction through sequential recommendation pre-training: Introducing the SRP4CTR framework. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 3777–3781.

Han, R., Yin, B., Chen, S., Jiang, H., Jiang, F., Li, X., Ma, C., Huang, M., Li, X., Jing, C., et al., 2025. MTGR: Industrial-scale generative recommendation framework in meituan. arXiv preprint arXiv:2505.18654.

Harrison, R.M., Dereventsov, A., Bibin, A., 2023. Zero-shot recommendations with pre-trained large language models for multimodal nudging. In: 2023 IEEE International Conference on Data Mining Workshops. ICDMW, IEEE, pp. 1535–1542.

He, Y., Liu, X., Zhang, A., Ma, Y., Chua, T.S., 2025. LLM2rec: Large language models are powerful embedding models for sequential recommendation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 896–907.

He, Z., Xie, Z., Jha, R., Steck, H., Liang, D., Feng, Y., Majumder, B.P., Kallus, N., Mcauley, J., 2023. Large language models as zero-shot conversational recommenders. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. CIKM '23, Association for Computing Machinery, New York, NY, USA, pp. 720–730. http://dx.doi.org/10.1145/3583780.3614949.

He, Z., Zhao, H., Wang, Z., Lin, Z., Kale, A., Mcauley, J., 2022. Query-aware sequential recommendation. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4019–4023.

Hou, M., Bai, C., Wu, L., Liu, H., Zhang, K., Liu, W., Hong, R., Tang, R., Wang, M., 2026a. RecCocktail: A generalizable and efficient framework for LLM-based recommendation. Proc. AAAI Conf. Artif. Intell. 40 (17), 14838–14847. http://dx.doi.org/10.1609/aaai.v40i17.38504, URL https://ojs.aaai.org/index.php/AAAI/article/view/38504.

Hou, Y., Li, J., Shin, A., Jeon, J., Santhanam, A., Shao, W., Hassani, K., Yao, N., McAuley, J., 2025a. Generating long semantic IDs in parallel for recommendation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 956–966.

Hou, M., Liu, X., Wu, L., He, C., Liu, H., Li, Z., Li, X., Wei, S., 2026b. WeaveRec: An LLM-based cross-domain sequential recommendation framework with model merging. In: Proceedings of the ACM Web Conference 2026. WWW '26, Association for Computing Machinery, New York, NY, USA, pp. 6342–6353. http://dx.doi.org/10.1145/3774904.3792365.

Hou, Y., Ni, J., He, Z., Sachdeva, N., Kang, W.C., Chi, E.H., McAuley, J., Cheng, D.Z., 2025b. ActionPiece: Contextually tokenizing action sequences for generative recommendation. In: Forty-Second International Conference on Machine Learning.

Hou, M., Wu, Y., Xu, C., Huang, Y.H., Bai, C., Wu, L., Bian, J., 2025c. InvDiff: Invariant guidance for bias mitigation in diffusion models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1. KDD '25, Association for Computing Machinery, New York, NY, USA, pp. 472–483. http://dx.doi.org/10.1145/3690624.3709165.

Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., Zhao, W.X., 2024. Large language models are zero-shot rankers for recommender systems. In: European Conference on Information Retrieval. Springer, pp. 364–381.

Hua, W., Ge, Y., Xu, S., Ji, J., Zhang, Y., 2023a. Up5: Unbiased foundation model for fairness-aware recommendation. arXiv preprint arXiv:2305.12090.

Hua, W., Xu, S., Ge, Y., Zhang, Y., 2023b. How to index item IDs for recommendation foundation models. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. In: SIGIR-AP '23, Association for Computing Machinery, New York, NY, USA, pp. 195–204. http://dx.doi.org/10.1145/3624918.3625339.

Huang, F., Bei, Y., Yang, Z., Jiang, J., Chen, H., Shen, Q., Wang, S., Karray, F., Yu, P.S., 2025a. Large language model simulator for cold-start recommendation. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. pp. 261–270.

Huang, M., Bu, C., He, Y., Wu, X., 2025b. How to mitigate information loss in knowledge graphs for GraphRAG: Leveraging triple context restoration and query-driven feedback. arXiv preprint arXiv:2501.15378.

Huang, Y., Chen, Y., Cao, X., Yang, R., Qi, M., Zhu, Y., Han, Q., Liu, Y., Liu, Z., Yao, X., et al., 2025c. Towards large-scale generative ranking. arXiv preprint arXiv:2505.04180.

Jia, J., Wang, Y., Li, Y., Chen, H., Bai, X., Liu, J., Chen, Q., Li, H., Jiang, P., et al., 2025. LEARN: Knowledge adaptation from large language model to recommendation for practical industrial application. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11861–11869.

Jiang, J., Huang, Y., Liu, B., Kong, X., Li, X., Xu, Z., Zhu, H., Xu, J., Zheng, B., 2025a. Large language model as universal retriever in industrial-scale recommender system. arXiv preprint arXiv:2502.03041.

Jiang, Z., Shi, Y., Li, M., Xiao, H., Qin, Y., Wei, Q., Wang, Y., Zhang, Y., 2024. Casevo: A cognitive agents and social evolution simulator. arXiv preprint arXiv:2412.19498.

Jiang, Y., Yang, Y., Xia, L., Luo, D., Lin, K., Huang, C., 2025b. RecLM: Recommendation instruction tuning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15443–15459.

Kang, W.C., Ni, J., Mehta, N., Sathiamoorthy, M., Hong, L., Chi, E., Cheng, D.Z., 2023. Do llms understand user preferences? evaluating llms on user rating prediction. arXiv preprint arXiv:2305.06474.

Kim, S., Kang, H., Choi, S., Kim, D., Yang, M., Park, C., 2024. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1395–1406.

Kim, J., Kim, H., Cho, H., Kang, S., Chang, B., Yeo, J., Lee, D., 2025. Review-driven personalized preference reasoning with large language models for recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1697–1706.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA.

Kong, X., Wu, J., Zhang, A., Sheng, L., Lin, H., Wang, X., He, X., 2024. Customizing language models with instance-wise lora for sequential recommendation. Adv. Neural Inf. Process. Syst. 37, 113072–113095.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30–37.

Kweon, W., Jang, S., Kang, S., Yu, H., 2025. Uncertainty quantification and decomposition for LLM-based recommendation. In: Proceedings of the ACM on Web Conference 2025. pp. 4889–4901.

Li, X., Chen, C., Zhao, X., Zhang, Y., Xing, C., 2023a. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation. arXiv preprint arXiv:2312.02443.

Li, W., Huang, R., Zhao, H., Liu, C., Zheng, K., Liu, Q., Mou, N., Zhou, G., Lian, D., Song, Y., et al., 2025a. DimeRec: a unified framework for enhanced sequential recommendation via generative diffusion models. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. pp. 726–734.

Li, J., Li, Y., Shen, X., Zhang, C., Qi, G., Sheng, B., 2025b. Open-world attribute mining for E-commerce products with multimodal self-correction instruction tuning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1702–1714.

Li, H., Shen, D., Wang, C., Liu, Y., Gu, J., 2025c. Can LLMs enhance fairness in recommendation systems? A data augmentation approach. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 570–580.

Li, Z., Sun, A., Li, C., 2023b. Diffurec: A diffusion model for sequential recommendation. ACM Trans. Inf. Syst. 42 (3), 1–28.

Li, X., Tang, H., Sheng, J., Zhang, X., Gao, L., Cheng, S., Yin, D., Liu, T., 2025d. Exploring preference-guided diffusion model for cross-domain recommendation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 719–728.

Li, Y., Wang, J., Sundaram, H., Liu, Z., 2025e. LLM-RecG: A semantic bias-aware framework for zero-shot sequential recommendation. In: Proceedings of the Nineteenth ACM Conference on Recommender Systems. pp. 237–246.

Li, J., Wang, S., Zhang, Q., Yu, S., Chen, F., 2025f. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In: Proceedings of the ACM on Web Conference 2025. pp. 2787–2798.

Li, Y., Zhai, X., Alzantot, M., Yu, K., Vulić, I., Korhonen, A., Hammad, M., 2024a. CALRec: Contrastive alignment of generative LLMs for sequential recommendation. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 422–432.

Li, L., Zhang, Y., Chen, L., 2020. Generate neural template explanations for recommendation. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20, Association for Computing Machinery, New York, NY, USA, pp. 755–764. http://dx.doi.org/10.1145/3340531.3411992.

Li, L., Zhang, Y., Chen, L., 2021. Personalized transformer for explainable recommendation. In: ACL.

Li, L., Zhang, Y., Liu, D., Chen, L., 2023c. Large language models for generative recommendation: A survey and visionary discussions. arXiv preprint arXiv:2309.01157.

Li, Y., Zhang, X., Luo, L., Chang, H., Ren, Y., King, I., Li, J., 2025g. G-refer: Graph retrieval-augmented large language model for explainable recommendation. In: Proceedings of the ACM on Web Conference 2025. WWW '25, Association for Computing Machinery, New York, NY, USA, pp. 240–251. http://dx.doi.org/10.1145/3696410.3714727.

Li, X., Zhao, C., Zhao, H., Wu, L., He, M., 2024b. Ganprompt: Enhancing robustness in llm-based recommendations with gan-enhanced diversity prompts. arXiv preprint arXiv:2408.09671.

Liao, J., He, X., Xie, R., Wu, J., Yuan, Y., Sun, X., Kang, Z., Wang, X., 2024a. Rosepo: Aligning llm-based recommenders with human values. arXiv preprint arXiv:2410.12519.

Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y., Wang, X., He, X., 2024b. Llara: Large language-recommendation assistant. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1785–1795.

Liao, Y., Wu, L., Hou, M., Wang, Y., Wu, H., Wang, M., 2026. From atom to community: Structured and evolving agent memory for user behavior modeling. arXiv preprint arXiv:2601.16872.

Liao, Y., Yang, Y., Hou, M., Wu, L., Xu, H., Liu, H., 2025. Mitigating distribution shifts in sequential recommendation: An invariance perspective. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1603–1613.

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., Liu, Y., Wu, C., Li, X., Zhu, C., et al., 2025a. How can recommender systems benefit from large language models: A survey. ACM Trans. Inf. Syst. 43 (2), 1–47.

Lin, J., Shan, R., Zhu, C., Du, K., Chen, B., Quan, S., Tang, R., Yu, Y., Zhang, W., 2024a. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. In: Proceedings of the ACM Web Conference 2024. pp. 3497–3508.

Lin, X., Shi, H., Wang, W., Feng, F., Wang, Q., Ng, S.K., Chua, T.S., 2025b. Order-agnostic identifier for large language model-based generative recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1923–1933.

Lin, X., Wang, W., Li, Y., Feng, F., Ng, S.-K., Chua, T.-S., 2024b. Bridging items and language: A transition paradigm for large language model-based recommendation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1816–1826.

Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., Chua, T.S., 2024c. Data-efficient fine-tuning for LLM-based recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 365–374.

Lin, X., Yang, C., Wang, W., Li, Y., Du, C., Feng, F., Ng, S.K., Chua, T.S., 2025c. Efficient inference for large language model-based generative recommendation. In: The Thirteenth International Conference on Learning Representations.

Liu, S., Cai, Q., He, Z., Sun, B., McAuley, J., Zheng, D., Jiang, P., Gai, K., 2023a. Generative flow network for listwise recommendation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1524–1534.

Liu, C., Cao, J., Huang, R., Zheng, K., Luo, Q., Gai, K., Zhou, G., 2024a. KuaiFormer: Transformer-based retrieval at kuaishou. arXiv preprint arXiv:2411.10057.

Liu, Q., Chen, N., Sakai, T., Wu, X.M., 2024b. ONCE: Boosting content-based recommendation with both open- and closed-source large language models. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. pp. 452–461.

Liu, Z., Chen, Z., Zhang, M., Duan, S., Wen, H., Li, L., Li, N., Gu, Y., Yu, G., 2024c. Modeling user viewing flow using large language models for article recommendation. In: Companion Proceedings of the ACM Web Conference 2024. pp. 83–92.

Liu, H., Guo, L., Zhu, L., Jiang, Y., Gao, M., Yin, H., 2024d. MCRPL: A pretrain, prompt, and fine-tune paradigm for non-overlapping many-to-one cross-domain recommendation. ACM Trans. Inf. Syst. 42 (4), 1–24.

Liu, J., Liu, C., Zhou, P., Lv, R., Zhou, K., Zhang, Y., 2023b. Is chatgpt a good recommender? a preliminary study. arXiv preprint arXiv:2304.10149.

Liu, Z., Wang, S., Wang, X., Zhang, R., Deng, J., Bao, H., Zhang, J., Li, W., Zheng, P., Wu, X., Hu, Y., Hu, Q., Luo, X., Ren, L., Zhang, Z., Wang, Q., Cai, K., Wu, Y., Cheng, H., Cheng, Z., Ren, L., Wang, H., Su, Y., Tang, R., Gai, K., Zhou, G., 2025a. OneRec-think: In-text reasoning for generative recommendation. arXiv:2510.11639. URL https://arxiv.org/abs/2510.11639.

Liu, Q., Wu, X., Wang, Y., Zhang, Z., Tian, F., Zheng, Y., Zhao, X., 2024e. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. Adv. Neural Inf. Process. Syst. 37, 26701–26727.

Liu, Z., Xu, Y., Cong, G., Zhu, L., Qiu, Q., Zhang, H., 2025b. ARTS: A general and efficient multi-task self-prompt framework for explainable sequential recommendation. ACM Trans. Inf. Syst. 43 (3), 1–30.

Liu, J., Yang, Y., Shao, P., Ma, H., Qin, W., Hong, R., 2026. wDPO: Winsorized direct preference optimization for robust LLM alignment. arXiv preprint arXiv:2603.07211.

Liu, S., Zhang, A., Hu, G., Qian, H., Chua, T.-s., 2025c. Preference diffusion for recommendation. In: The Thirteenth International Conference on Learning Representations.

Liu, Q., Zhao, X., Wang, Y., Zhang, Z., Zhong, H., Chen, C., Li, X., Huang, W., Tian, F., 2025d. Bridge the domains: Large language models enhanced cross-domain sequential recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1582–1592.

Liu, E., Zheng, B., Zhao, W.X., Wen, J.R., 2025e. Bridging textual-collaborative gap through semantic codes for sequential recommendation. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1788–1798.

Liu, Q., Zhu, J., Yang, Y., Dai, Q., Du, Z., Wu, X.M., Zhao, Z., Zhang, R., Dong, Z., 2024f. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6566–6576.

Luo, S., He, B., Zhao, H., Shao, W., Qi, Y., Huang, Y., Zhou, A., Yao, Y., Li, Z., Xiao, Y., et al., 2024a. Recranker: Instruction tuning large language model as ranker for top-k recommendation. ACM Trans. Inf. Syst..

Luo, W., Song, C., Yi, L., Cheng, G., 2024b. Trawl: External knowledge-enhanced recommendation with llm assistance. arXiv preprint arXiv:2403.06642.

Lyu, H., Jiang, S., Zeng, H., Xia, Y., Wang, Q., Zhang, S., Chen, R., Leung, C., Tang, J., Luo, J., 2024. LLM-rec: Personalized recommendation via prompting large language models. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 583–612.

Ma, H., Ma, Y., Xie, R., Meng, L., Shen, J., Sun, X., Kang, Z., Chua, T.-S., 2025. Large language model empowered recommendation meets all-domain continual pre-training. arXiv preprint arXiv:2504.08949.

Ma, Q., Ren, X., Huang, C., 2024. XRec: Large language models for explainable recommendation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp. 391–402, URL https://aclanthology.org/2024.findings-emnlp.22.

Mao, W., Liu, S., Liu, H., Liu, H., Li, X., Hu, L., 2025a. Distinguished quantized guidance for diffusion-based sequence recommendation. In: Proceedings of the ACM on Web Conference 2025. pp. 425–435.

Mao, Z., Wang, H., Du, Y., Wong, K.F., 2023. UniTRec: A unified text-to-text transformer and joint contrastive learning framework for text-based recommendation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1160–1170.

Mao, W., Wu, J., Chen, W., Gao, C., Wang, X., He, X., 2025b. Reinforced prompt personalization for recommendation with large language models. ACM Trans. Inf. Syst. 43 (3), 1–27.

Mao, W., Yang, Z., Wu, J., Liu, H., Yuan, Y., Wang, X., He, X., 2025c. Addressing missing data issue for diffusion-based recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2152–2161.

Mei, T., Chen, H., Yu, P., Liang, J., Yang, D., 2025. GORACS: Group-level optimal transport-guided coreset selection for LLM-based recommender systems. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2126–2137.

Ngo, H., Nguyen, D.Q., 2024. RecGPT: Generative pre-training for text-based recommendation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 302–313.

Nguyen, T.T., Quoc Viet Hung, N., Nguyen, T.T., Huynh, T.T., Nguyen, T.T., Weidlich, M., Yin, H., 2024. Manipulating recommender systems: A survey of poisoning attacks and countermeasures. ACM Comput. Surv. 57 (1), 1–39.

Ni, J., Li, J., McAuley, J., 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 188–197. http://dx.doi.org/10.18653/v1/D19-1018, URL https://aclanthology.org/D19-1018/.

Ning, L., Fan, W., Li, Q., 2025. Retrieval-augmented purifier for robust LLM-empowered recommendation. arXiv preprint arXiv:2504.02458.

Ning, L.-b., Wang, S., Fan, W., Li, Q., Xu, X., Chen, H., Huang, F., 2024. Cheatagent: Attacking llm-empowered recommender systems via llm agent. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2284–2295.

Niu, Y., Xing, X., Jia, Z., Liu, R., Xin, M., Cui, J., 2024. Diffusion recommendation with implicit sequence influence. In: Companion Proceedings of the ACM Web Conference 2024. pp. 1719–1725.

Peng, Y., Gao, C., Zhang, Y., Dan, T., Du, X., Luo, H., Li, Y., Meng, X., 2025. Denoising alignment with large language model for recommendation. ACM Trans. Inf. Syst. 43 (2), 1–35.

Penha, G., Vardasbi, A., Palumbo, E., De Nadai, M., Bouchard, H., 2024. Bridging search and recommendation in generative retrieval: Does one task help the other? In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 340–349.

Pi, Q., Bian, W., Zhou, G., Zhu, X., Gai, K., 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2671–2679.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J.Y., Zhou, D., et al., 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. arXiv preprint arXiv:2502.08691.

Qiao, S., Zhou, W., Wen, J., Gao, C., Luo, Q., Chen, P., Li, Y., 2025. Multi-view intent learning and alignment with large language models for session-based recommendation. ACM Trans. Inf. Syst. 43 (4), 1–25.

Qiu, Z., Luo, L., Zhao, Z., Pan, S., Liew, A.W.C., 2025a. Graph retrieval-augmented LLM for conversational recommendation systems. In: Wu, X., Spiliopoulou, M., Wang, C., Kumar, V., Cao, L., Wu, Y., Yao, Y., Wu, Z. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Nature Singapore, Singapore, pp. 344–355.

Qiu, Z., Tao, Y., Pan, S., Liew, A.W.C., 2025b. Knowledge graphs and pretrained language models enhanced representation learning for conversational recommender systems. IEEE Trans. Neural Netw. Learn. Syst. 36 (4), 6107–6121. http://dx.doi.org/10.1109/TNNLS.2024.3395334.

Qiu, J., Wang, Z., Zhang, F., Zheng, Z., Zhu, J., Fan, J., Zhang, T., Wang, H., Wang, X., 2025c. EGA-V1: Unifying online advertising with end-to-end learning. arXiv preprint arXiv:2505.19755.

Qu, H., Fan, W., Zhao, Z., Li, Q., 2025a. TokenRec: Learning to tokenize ID for LLM-based generative recommendations. IEEE Trans. Knowl. Data Eng..

Qu, Y., Nobuhara, H., 2025. Intent-aware diffusion with contrastive learning for sequential recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1552–1561.

Qu, Z., Xie, R., Xiao, C., Yao, Y., Liu, Z., Lian, F., Kang, Z., Zhou, J., 2025b. Thoroughly modeling multi-domain pre-trained recommendation as language. ACM Trans. Inf. Syst. 43 (2), 1–28.

Rajput, S., Mehta, N., Singh, A., Keshavan, R., Vu, T., Heidt, L., Hong, L., Tay, Y., Tran, V.Q., Samost, J., et al., 2023. Recommender systems with generative retrieval. In: Advances in Neural Information Processing Systems. pp. 10299–10315.

Ravaut, M., Zhang, H., Xu, L., Sun, A., Liu, Y., 2024. Parameter-efficient conversational recommender system as a language processing task. In: Graham, Y., Purver, M. (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, St. Julian's, Malta, pp. 152–165. http://dx.doi.org/10.18653/v1/2024.eacl-long.9, URL https://aclanthology.org/2024.eacl-long.9/.

Ren, Y., Chen, Z., Yang, X., Li, L., Jiang, C., Cheng, L., Zhang, B., Mo, L., Zhou, J., 2024a. Enhancing sequential recommenders with augmented knowledge from aligned large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 345–354.

Ren, X., Huang, C., 2024. Easyrec: Simple yet effective language models for recommendation. arXiv preprint arXiv:2408.08821.

Ren, X., Wei, W., Xia, L., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C., 2024b. Representation learning with large language models for recommendation. In: Proceedings of the ACM Web Conference 2024. pp. 3464–3475.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. Grouplens: An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. pp. 175–186.

Rocamonde, J., Montesinos, V., Nava, E., Perez, E., Lindner, D., 2023. Vision-language models are zero-shot reward models for reinforcement learning. arXiv preprint arXiv:2310.12921.

Shi, W., He, X., Zhang, Y., Gao, C., Li, X., Zhang, J., Wang, Q., Feng, F., 2024a. Large language models are learnable planners for long-term recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1893–1903.

Shi, J., Xiong, W., Lin, Z., Jung, H.J., 2024b. Instantbooth: Personalized text-to-image generation without test-time finetuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8543–8552.

Sileo, D., Vossen, W., Raymaekers, R., 2022. Zero-shot recommendation as language modeling. In: European Conference on Information Retrieval. Springer, pp. 223–230.

Song, T., Chao, W., Liu, H., 2024. Large language model enhanced hard sample identification for denoising recommendation. arXiv preprint arXiv:2409.10343.

Song, Y.Z., Chen, Y.S., Wang, L., Shuai, H.H., 2023. General then personal: Decoupling and pre-training for personalized headline generation. Trans. Assoc. Comput. Linguist. 11, 1588–1607.

Spurlock, K.D., Acun, C., Saka, E., Nasraoui, O., 2024. ChatGPT for conversational recommendation: Refining recommendations by reprompting with feedback. arXiv:2401.03605. URL https://arxiv.org/abs/2401.03605.

Sun, Z., Liu, H., Qu, X., Feng, K., Wang, Y., Ong, Y.S., 2024. Large language models for intent-driven session recommendations. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 324–334.

Sun, Y., Sun, Z., Du, Y., Zhang, J., Ong, Y.S., 2025. Model-agnostic social network refinement with diffusion models for robust social recommendation. In: Proceedings of the ACM on Web Conference 2025. pp. 370–378.

Sun, P., Wu, L., Zhang, K., Fu, Y., Hong, R., Wang, M., 2020. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In: Proceedings of the Web Conference 2020. WWW '20, Association for Computing Machinery, New York, NY, USA, pp. 837–847. http://dx.doi.org/10.1145/3366423.3380164.

Sun, W., Xie, R., Zhang, J., Zhao, W.X., Lin, L., Wen, J.R., 2023. Generative next-basket recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 737–743.

Tan, J., Xu, S., Hua, W., Ge, Y., Li, Z., Zhang, Y., 2024. IdGenRec: LLM-RecSys alignment with textual ID learning. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, Association for Computing Machinery, New York, NY, USA, pp. 355–364. http://dx.doi.org/10.1145/3626772.3657821.

Tang, J., Dai, S., Shi, T., Xu, J., Chen, X., Chen, W., Wu, J., Jiang, Y., 2025. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. arXiv:2503.22675. URL https://arxiv.org/abs/2503.22675.

Tang, Z., Huan, Z., Li, Z., Zhang, X., Hu, J., Fu, C., Zhou, J., Zou, L., Li, C., 2024. One model for all: Large language models are domain-agnostic recommendation systems. ACM Trans. Inf. Syst..

Wang, W., Bao, H., Lin, X., Zhang, J., Li, Y., Feng, F., Ng, S.K., Chua, T.S., 2024a. Learnable item tokenization for generative recommendation. In: Proceedings of the International Conference on Information and Knowledge Management. pp. 2400–2409.

Wang, Y., Chu, Z., Ouyang, X., Wang, S., Hao, H., Shen, Y., Gu, J., Xue, S., Zhang, J., Cui, Q., Li, L., Zhou, J., Li, S., 2024b. LLMRG: improving recommendations through large language model reasoning graphs. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. In: AAAI'24/IAAI'24/EAAI'24, AAAI Press, http://dx.doi.org/10.1609/aaai.v38i17.29887.

Wang, W., Cui, L., Liu, X., Nag, S., Xu, W., Luo, C., Sarwar, S.M., Li, Y., Gu, H., Liu, H., et al., 2025a. EcomScriptBench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. Proc. 63rd Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers) 1–22.

Wang, Z., Gao, M., Yu, J., Gao, X., Nguyen, Q.V.H., Sadiq, S., Yin, H., 2025b. Id-free not risk-free: Llm-powered agents unveil risks in id-free recommender systems. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1902–1911.

Wang, Z., Gao, M., Yu, J., Hou, Y., Sadiq, S., Yin, H., 2025c. RuleAgent: Discovering rules for recommendation denoising with autonomous language agents. arXiv preprint arXiv:2503.23374.

Wang, J., Karatzoglou, A., Arapakis, I., Jose, J.M., 2024c. Reinforcement learning-based recommender systems with large language models for state reward and action modeling. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 375–385.

Wang, Q., Li, J., Wang, S., Xing, Q., Niu, R., Kong, H., Li, R., Long, G., Chang, Y., Zhang, C., 2024d. Towards next-generation LLM-based recommender systems: A survey and beyond. arXiv preprint arXiv:2410.19744.

Wang, H., Lin, J., Li, X., Chen, B., Zhu, C., Tang, R., Zhang, W., Yu, Y., 2024e. FLIP: Fine-grained alignment between ID-based models and pretrained language models for CTR prediction. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 94–104.

Wang, B., Liu, F., Chen, J., Lou, X., Zhang, C., Wang, J., Sun, Y., Feng, Y., Chen, C., Wang, C., 2025d. MSL: Not all tokens are what you need for tuning LLM as a recommender. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1912–1922.

Wang, B., Liu, F., Zhang, C., Chen, J., Wu, Y., Zhou, S., Lou, X., Wang, J., Feng, Y., Chen, C., et al., 2024f. LLM4DSR: Leveraging large language model for denoising sequential recommendation. ACM Trans. Inf. Syst..

Wang, J., Lu, H., Caverlee, J., Chi, E.H., Chen, M., 2024g. Large language models as data augmenters for cold-start item recommendation. In: Companion Proceedings of the ACM Web Conference 2024. pp. 726–729.

Wang, Y., Pan, J., Jia, P., Wang, W., Wang, M., Feng, Z., Li, X., Jiang, J., Zhao, X., 2025e. Pre-train, align, and disentangle: Empowering sequential recommendation with large language models. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1455–1465.

Wang, Y., Sang, L., Zhang, Y., Zhang, Y., 2025f. Intent representation learning with large language model for recommendation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieva7962l. pp. 1870–1879.

Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D., 2023a. Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wang, C., Wu, B., Chen, Z., Shen, L., Wang, B., Zeng, X., 2025g. Scaling transformers for discriminative recommendation via generative pretraining. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2893–2903.

Wang, H., Wu, C., Huang, Y., Qi, T., 2025h. Learning human feedback from large language models for content quality-aware recommendation. ACM Trans. Inf. Syst. 43 (4), 1–28.

Wang, W., Xu, Y., Feng, F., Lin, X., He, X., Chua, T.S., 2023b. Diffusion recommender model. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 832–841.

Wang, Y., Yang, Y., Wu, L., Wu, J., Xu, H., Lin, H., 2026. MLLMRec-R1: Ince7962ntivizing reasoning capability in large language models for multimodal sequential recommendation. arXiv preprint arXiv:2603.06243.

Wang, Y., Yang, Y., Wu, L., Zhang, Y., Liu, F., Hong, R., 2025i. Multimodal large language models with adaptive preference optimization for sequential recommendation. arXiv preprint arXiv:2511.18740.

Wang, L., Zhang, J., Yang, H., Chen, Z.-Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., et al., 2025j. User behavior simulation with large language model-based agents. ACM Trans. Inf. Syst. 43 (2), 1–37.

Wang, L., Zhang, D., Yang, F., Zhao, P., Liu, J., Zhan, Y., Sun, H., Lin, Q., Deng, W., Zhang, D., et al., 2025k. LettinGo: Explore user profile generation for recommendation system. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2985–2995.

Wang, S., Zheng, Z., Sui, Y., Xiong, H., 2025l. Unleashing the power of large language model for denoising recommendation. In: Proceedings of the ACM on Web Conference 2025. pp. 252–263.

Wei, T., Jin, B., Li, R., Zeng, H., Wang, Z., Sun, J., Yin, Q., Lu, H., Wang, S., He, J., et al., 2024a. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. In: The Twelfth International Conference on Learning Representations.

Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C., 2024b. Llmrec: Large language models with graph augmentation for recommendation. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. pp. 806–815.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA.

Wu, J., Chang, C.C., Yu, T., He, Z., Wang, J., Hou, Y., McAuley, J., 2024a. Coral: collaborative retrieval-augmented large language models improve long-tail recommendation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3391–3401.

Wu, L., He, X., Wang, X., Zhang, K., Wang, M., 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. IEEE Trans. Knowl. Data Eng. 35 (5), 4425–4445.

Wu, Y., Xie, R., Zhang, Z., Zhang, X., Zhuang, F., Lin, L., Kang, Z., An, Z., Xu, Y., 2024b. ID-centric pre-training for recommendation. ACM Trans. Inf. Syst..

Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., et al., 2024c. A survey on large language models for recommendation. World Wide Web 27 (5), 60.

Xi, Y., Liu, W., Lin, J., Cai, X., Zhu, H., Zhu, J., Chen, B., Tang, R., Zhang, W., Yu, Y., 2024a. Towards open-world recommendation with knowledge augmentation from large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 12–22.

Xi, Y., Liu, W., Lin, J., Chen, B., Tang, R., Zhang, W., Yu, Y., 2024b. MemoCRS: Memory-enhanced sequential conversational recommender systems with large language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. CIKM '24, Association for Computing Machinery, New York, NY, USA, pp. 2585–2595. http://dx.doi.org/10.1145/3627673.3679599.

Xi, Y., Weng, M., Chen, W., Yi, C., Chen, D., Guo, G., Zhang, M., Wu, J., Jiang, Y., Liu, Q., et al., 2025. Bursting filter bubble: Enhancing serendipity recommendations with aligned large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 5059–5070.

Xie, Z., Dong, H., Gao, Y., Ma, Z., Liang, X., 2024a. Dreamvton: Customizing 3d virtual try-on with personalized diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10784–10793.

Xie, W., Wang, H., Zhang, L., Zhou, R., Lian, D., Chen, E., 2024b. Breaking determinism: Fuzzy modeling of sequential recommendation using discrete state space diffusion model. Adv. Neural Inf. Process. Syst. 37, 22720–22744.

Xin, H., Sun, Y., Wang, C., Xiong, H., 2025a. Llmcdsr: Enhancing cross-domain sequential recommendation with large language models. ACM Trans. Inf. Syst..

Xin, H., Sun, Y., Wang, C., Yu, Y., Zhang, W., Xiong, H., 2025b. Improving recommendation fairness without sensitive attributes using multi-persona LLMs. arXiv preprint arXiv:2505.19473.

Xu, Y., Gu, T., Chen, W., Chen, A., 2025a. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, (9), pp. 8996–9004.

Xu, H., Hou, M., Wu, L., Liu, F., Yang, Y., Bai, H., Hong, R., Wang, M., 2025b. Fair personalized learner modeling without sensitive attributes. In: Proceedings of the ACM on Web Conference 2025. pp. 4612–4624.

Xu, Y., Wang, W., Feng, F., Ma, Y., Zhang, J., He, X., 2024. Diffusion models for generative outfit recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1350–1359.

Xu, H., Wu, L., Cheng, C., Liu, H., 2026a. Multi-value alignment for LLMs via value decorrelation and extrapolation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, (40), pp. 34133–34141.

Xu, W., Wu, Q., Liang, Z., Han, J., Ning, X., Shi, Y., Lin, W., Zhang, Y., 2025c. SLMRec: Distilling large language models into small for sequential recommendation. In: The Thirteenth International Conference on Learning Representations.

Xu, H., Wu, L., Wang, Y., Hou, M., Wu, H., Zhang, Z., Wang, M., 2026b. VC-soup: Value-consistency guided multi-value alignment for large language models. In: Proceedings of the ACM Web Conference 2026. pp. 9699–9710.

Xu, Y., Zhang, J., Salemi, A., Hu, X., Wang, W., Feng, F., Zamani, H., He, X., Chua, T.S., 2025d. Personalized generation in large model era: A survey. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp. 24607–24649. http://dx.doi.org/10.18653/v1/2025.acl-long.1201, URL https://aclanthology.org/2025.acl-long.1201/.

Yang, T., Chen, L., 2024. Unleashing the retrieval potential of large language models in conversational recommender systems. In: Proceedings of the 18th ACM Conference on Recommender Systems. RecSys '24, Association for Computing Machinery, New York, NY, USA, pp. 43–52. http://dx.doi.org/10.1145/3640457.3688146.

Yang, D., Chen, F., Fang, H., 2024a. Behavior alignment: A new perspective of evaluating LLM-based conversational recommendation systems. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, Association for Computing Machinery, New York, NY, USA, pp. 2286–2290. http://dx.doi.org/10.1145/3626772.3657924.

Yang, S., Ma, W., Sun, P., Ai, Q., Liu, Y., Cai, M., Zhang, M., 2024b. Sequential recommendation with latent relations based on large language model. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 335–344.

Yang, L., Subbiah, A., Patel, H., Li, J.Y., Song, Y., Mirghaderi, R., Aggarwal, V., Wang, Q., 2025a. Item-language model for conversational recommendation. arXiv: 2406.02844. URL https://arxiv.org/abs/2406.02844.

Yang, Y., Tao, W., Liu, J., Zhu, X., Fang, J., Huang, W., Wu, L., Hong, R., Chua, T.-S., 2026. Revisiting robustness for LLM safety alignment via selective geometry control. arXiv preprint arXiv:2602.07340.

Yang, Y., Wu, L., Liao, Y., He, Z., Shao, P., Hong, R., Wang, M., 2025b. Invariance matters: Empowering social recommendation via graph invariant learning. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2038–2047.

Yang, Z., Wu, J., Wang, Z., Wang, X., Yuan, Y., He, X., 2023. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. Adv. Neural Inf. Process. Syst. 36, 24247–24261.

Yang, M., Zhu, M., Wang, Y., Chen, L., Zhao, Y., Wang, X., Han, B., Zheng, X., Yin, J., 2024c. Fine-tuning large language model based explainable recommendation with explainable quality reward. Proc. AAAI Conf. Artif. Intell. 38 (8), 9250–9259. http://dx.doi.org/10.1609/aaai.v38i8.28777, URL https://ojs.aaai.org/index.php/AAAI/article/view/28777.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.R., 2023. Tree of thoughts: Deliberate problem solving with large language models. In: Thirty-Seventh Conference on Neural Information Processing Systems. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Ye, Y., Zheng, Z., Shen, Y., Wang, T., Zhang, H., Zhu, P., Yu, R., Zhang, K., Xiong, H., 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, (12), pp. 13069–13077.

Yin, M., Pan, J., Wang, H., Wang, X., Zhang, S., Jiang, J., Lian, D., Chen, E., 2025. From feature interaction to feature generation: A generative paradigm of CTR prediction modelswang2024large-WWW-2024. In: Forty-Second International Conference on Machine Learning.

Yu, Q., Fu, K., Zhang, S., Lv, Z., Wu, F., Wu, F., 2025. ThinkRec: Thinking-based recommendation via LLM. arXiv:2505.15091. URL https://arxiv.org/abs/2505.15091.

Yu, C., Liu, X., Maia, J., Li, Y., Cao, T., Gao, Y., Song, Y., Goutam, R., Zhang, H., Yin, B., et al., 2024. COSMO: A large-scale e-commerce common sense knowledge generation and serving system at amazon. In: Companion of the 2024 International Conference on Management of Data. pp. 148–160.

Yuan, W., Yang, C., Ye, G., Chen, T., Hung, N.Q.V., Yin, H., 2024. Fellas: Enhancing federated sequential recommendation with llm as external services. ACM Trans. Inf. Syst..

Yue, Z., Rabhi, S., Moreira, G.d.P., Wang, D., Oldridge, E., 2023. Llamarec: Two-stage recommendation using large language models for ranking. arXiv preprint arXiv:2311.02089.

Zha, Y., Dong, X., Ma, H., Yang, Y., Wang, X., 2026. Align-for-fusion: Harmonizing triple preferences via dual-oriented diffusion for cross-domain sequential recommendation. In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. pp. 1868–1879.

Zhai, J., Liao, L., Liu, X., Wang, Y., Li, R., Cao, X., Gao, L., Gong, Z., Gu, F., He, J., et al., 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In: Proceedings of the 41st International Conference on Machine Learning. pp. 58484–58509.

Zhai, J., Mai, Z.F., Wang, C.D., Yang, F., Zheng, X., Li, H., Tian, Y., 2025. Multimodal quantitative language for generative recommendation. In: The Thirteenth International Conference on Learning Representations.

Zhang, Y., Bao, K., Yan, M., Wang, W., Feng, F., He, X., 2024a. Text-like encoding of collaborative information in large language models for recommendation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9181–9191.

Zhang, A., Chen, Y., Sheng, L., Wang, X., Chua, T.S., 2024b. On generative agents in recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1807–1817.

Zhang, Y., Feng, F., Zhang, J., Bao, K., Wang, Q., He, X., 2025a. Collm: Integrating collaborative embeddings into large language models for recommendation. IEEE Trans. Knowl. Data Eng..

Zhang, J., Hou, Y., Xie, R., Sun, W., McAuley, J., Zhao, W.X., Lin, L., Wen, J.R., 2024c. Agentcf: Collaborative learning with autonomous language agents for recommender systems. In: Proceedings of the ACM Web Conference 2024. pp. 3679–3689.

Zhang, J., Li, Y., Xu, Y., Zhang, L., Feng, X., Ren, Z., Chen, C., 2025b. Bifair: A fairness-aware training framework for LLM-enhanced recommender systems via bi-level optimization. arXiv preprint arXiv:2507.04294.

Zhang, J., Liu, Y., Liu, Q., Wu, S., Guo, G., Wang, L., 2024d. Stealthy attack on large language model based recommendation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5839–5857.

Zhang, J., Xie, R., Hou, Y., Zhao, X., Lin, L., Wen, J.-R., 2023. Recommendation as instruction following: A large language model empowered recommendation approach. ACM Trans. Inf. Syst..

Zhang, Y., Xu, W., Zhao, X., Wang, W., Feng, F., He, X., Chua, T.-S., 2025c. Reinforced latent reasoning for LLM-based recommendation. arXiv:2505.19092. URL https://arxiv.org/abs/2505.19092.

Zhang, J., Zhang, B., Sun, W., Lu, H., Zhao, W.X., Chen, Y., Wen, J.R., 2025d. Slow thinking for sequential recommendation. arXiv:2504.09627. URL https://arxiv.org/abs/2504.09627.

Zhang, C., Zhang, H., Wu, S., Wu, D., Xu, T., Zhao, X., Gao, Y., Hu, Y., Chen, E., 2025e. Notellm-2: Multimodal large representation models for recommendation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2815–2826.

Zhang, Y., Zhang, Y., Zheng, K., Chen, T., Yin, H., 2026. ProMax: Exploring the potential of LLM-derived profiles with distribution shaping for recommender systems. arXiv preprint arXiv:2604.26231.

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., et al., 2024a. Recommender systems in the era of large language models (llms). IEEE Trans. Knowl. Data Eng. 36 (11), 6889–6907.

Zhao, Z., Lin, F., Zhu, X., Zheng, Z., Xu, T., Shen, S., Li, X., Yin, Z., Chen, E., 2024b. DynLLM: when large language models meet dynamic graph recommendation. arXiv preprint arXiv:2405.07580.

Zhao, Q., Qian, H., Liu, Z., Zhang, G.D., Gu, L., 2024c. Breaking the barrier: utilizing large language models for industrial recommendation systems through an inferential knowledge graph. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 5086–5093.

Zhao, J., Wenjie, W., Xu, Y., Sun, T., Feng, F., Chua, T.S., 2024d. Denoising diffusion recommender model. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1370–1379.

Zhao, K., Xu, F., Li, Y., 2025a. Reason-to-recommend: Using interaction-of-thought reasoning to enhance LLM recommendation. arXiv:2506.05069. URL https://arxiv.org/abs/2506.05069.

Zhao, C., Yang, E., Liang, Y., Zhao, J., Guo, G., Wang, X., 2025b. Distributionally robust graph out-of-distribution recommendation via diffusion model. In: Proceedings of the ACM on Web Conference 2025. pp. 2018–2031.

Zheng, Z., Chao, W., Qiu, Z., Zhu, H., Xiong, H., 2024a. Harnessing large language models for text-rich sequential recommendation. In: Proceedings of the ACM Web Conference 2024. pp. 3207–3216.

Zheng, B., Hou, Y., Lu, H., Chen, Y., Zhao, W.X., Chen, M., Wen, J.-R., 2024b. Adapting large language models by integrating collaborative semantics for recommendation. In: 2024 IEEE 40th International Conference on Data Engineering. ICDE, IEEE, pp. 1435–1448.

Zheng, B., Wang, X., Liu, E., Wang, X., Hongyu, L., Chen, Y., Zhao, W.X., Wen, J.-R., 2025a. DeepRec: Towards a deep dive into the item space with large language model based recommendation. arXiv:2505.16810. URL https://arxiv.org/abs/2505.16810.

Zheng, Z., Wang, Z., Yang, F., Fan, J., Zhang, T., Wang, Y., Wang, X., 2025b. EGA-V2: An end-to-end generative framework for industrial advertising. arXiv preprint arXiv:2505.17549.

Zhong, H., Wang, H., Ye, Y., Zhang, M., Zhu, S., 2025. GGBond: Growing graph-based AI-agent society for socially-aware recommender simulation. arXiv preprint arXiv:2505.21154.

Zhou, G., Hu, H., Cheng, H., Wang, H., Deng, J., Zhang, J., Cai, K., Ren, L., Ren, L., Yu, L., et al., 2025a. Onerec-v2 technical report. arXiv preprint arXiv:2508.20900.

Zhou, P., Liu, C., Ren, J., Zhou, X., Xie, Y., Cao, M., Rao, Z., Huang, Y.-L., Chong, D., Liu, J., et al., 2025b. When large vision language models meet multimodal sequential recommendation: An empirical study. In: Proceedings of the ACM on Web Conference 2025. pp. 275–292.

Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., Gai, K., 2018. Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1059–1068.

Zhu, J., Fan, Z., Zhu, X., Jiang, Y., Wang, H., Han, X., Ding, H., Wang, X., Zhao, W., Gong, Z., Yang, H., Chai, Z., Chen, Z., Zheng, Y., Chen, Q., Zhang, F., Zhou, X., Xu, P., Yang, X., Wu, D., Liu, Z., 2025. RankMixer: Scaling up ranking models in industrial recommenders. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. CIKM '25, Association for Computing Machinery, New York, NY, USA, pp. 6309–6316. http://dx.doi.org/10.1145/3746252.3761507.

Zhu, Y., Wu, L., Guo, Q., Hong, L., Li, J., 2024. Collaborative large language model for recommender systems. In: Proceedings of the ACM Web Conference 2024. pp. 3162–3172.