

# A Survey of Foundation Model-Powered Recommender Systems: From Feature-Based, Generative to Agentic Paradigms

Chengkai Huang, Hongtao Huang, Tong Yu, Kaige Xie, Junda Wu, Shuai Zhang, Julian Mcauley, Dietmar Jannach and Lina Yao

**Abstract**—Recommender systems (RS) have become essential in filtering information and personalizing content for users. RS techniques have traditionally relied on modeling interactions between users and items as well as the features of content using models specific to each task. The emergence of foundation models (FMs), large scale models trained on vast amounts of data such as GPT, LLaMA and CLIP, is reshaping the recommendation paradigm. This survey provides a comprehensive overview of the Foundation Models for Recommender Systems (FM4RecSys), covering their integration in three paradigms: (1) *Feature-Based* augmentation of representations, (2) *Generative recommendation* approaches, and (3) *Agentic interactive systems*. We first review the data foundations of RS, from traditional explicit or implicit feedback to multimodal content sources. We then introduce FMs and their capabilities for representation learning, natural language understanding, and multi-modal reasoning in RS contexts. The core of the survey discusses how FMs enhance RS under the feature-based paradigm (improving feature representations), the generative paradigm (directly generating recommendations or related content), and the agentic paradigm (enabling autonomous recommendation agents and simulators). Afterward, we examine FM applications in various recommendation tasks: Top-N recommendation, sequential recommendation, zero/few-shot scenarios, conversational recommendation, and novel item/content generation. Through an analysis of recent research, we highlight key opportunities that have been realized (e.g., improved generalization, better explanations, reasoning ability) as well as challenges encountered (e.g., cross-domain generality, interpretability, fairness, and multimodal integration). Finally, we outline open research directions and technical challenges for next-generation FM4RecSys, such as multimodal recommender agents, retrieval-augmented frameworks, lifelong learning for long user sequences, efficiency and cost issues, etc. This survey not only reviews the state-of-the-art methods but also provides a critical analysis of the trade-offs among the feature-based, the generative, and the agentic paradigms, outlining key open issues and future research directions.

**Index Terms**—Foundation models, Recommender Systems, Multi-modal Representation, Survey.

## 1 INTRODUCTION

Recommender Systems (RSs) have become critical in a wide range of domains, from e-commerce and social media to healthcare and education [1], [2]. They aim to deliver personalized content by capturing user preferences, item characteristics, and contextual signals. Over the past decade, the field has witnessed remarkable progress, driven by advancements in deep learning architectures and the increasing availability of large-scale user behavior data. Despite these achievements, traditional RSs still face persistent challenges in capturing subtle user preferences, handling

cold-start scenarios, and providing transparent, context-rich explanations. These challenges limit the effectiveness of purely domain-specific or small-scale models in providing accurate and diverse recommendations.

In parallel, Foundation Models (FMs) have made significant strides in areas such as natural language processing, computer vision, and multi-modal tasks [3]. Recently, FMs have been reshaping recommender system architectures—boosting performance, enabling novel modes of user interaction, and demonstrating strong potential in capturing complex user-item relationships while generalizing across a broader spectrum of recommendation tasks. To be specific, Foundation Models for Recommender Systems (**FM4RecSys**) refer to leveraging the knowledge from pre-training and recommendation datasets to capture rich representations of user preferences, item features, and contextual variables for improving personalization and prediction accuracy in recommendation tasks. Meanwhile, Foundation Models (FMs)—large-scale, pre-trained models with strong task generalization capabilities that provide a unified and flexible modeling paradigm for various downstream recommendation tasks [4]. Unlike conventional methods that rely on meticulously engineered features or narrow architectures, FMs leverage broad pre-training over massive corpora, enabling stronger generalization and the ability to incorporate a wide variety of signals (text, images, audio,

- C. Huang is with the School of Computer Science and Engineering, The University of New South Wales. E-mail: chengkai.huang1@unsw.edu.au.
- H. Huang is with the School of Computer Science and Engineering, The University of New South Wales. E-mail: hongtao.huang@unsw.edu.au
- T. Yu, is with Adobe Research. E-mail: tyu@adobe.com
- K. Xie is with the School of Computer Science, Georgia Institute of Technology. E-mail: kaigexie@gatech.edu.
- J. Wu is with the University of California, San Diego. E-mail: juw069@ucsd.edu.
- S. Zhang is with ETH Zurich. E-mail: cheungshuai@outlook.com.
- J. Mcauley is with the University of California, San Diego. E-mail: jmcauley@eng.ucsd.edu.
- D. Jannach is with the University of Klagenfurt. E-mail: Dietmar.Jannach@aau.at.
- L. Yao is with The University of New South Wales and CSIRO's Data61. E-mail: lina.yao@data61.csiro.au.

(Corresponding authors: Chengkai Huang.)

knowledge graphs, etc.). This flexibility can yield richer user/item representations and help overcome the data sparsity and cold-start issues that plague traditional collaborative filtering. Beyond boosting predictive accuracy, foundation models (FMs) unlock novel capabilities—including natural language explanations, interactive conversational interfaces, and even agentic decision-making. In particular, agentic frameworks leverage FMs to autonomously plan, reason, and adapt within dynamic environments by incorporating iterative user feedback and real-time contextual understanding. Next, we dive into the motivations behind existing works that incorporate foundation models into recommender systems, aiming to deepen our understanding of how FMs are applied and what impact they have across different recommendation tasks.

## 1.1 Motivation

We enumerate the primary motivations driving the research in the evolving landscape of FM4RecSys, aiming to provide a comprehensive understanding of the factors that contribute to the development and adoption of FM-powered RSs.

**Enhanced Generalization Capabilities.** Foundation Models are designed to learn from large-scale data, enabling them to understand complex patterns. FMs can generalize better to new, unseen data [5]. In the context of RSs, this means that FMs can more accurately predict user preferences and behaviors, especially in scenarios with sparse data or novel items (defined as zero-shot/few-shot recommendations in some papers [6]–[8]). Through zero-shot/few-shot inference of user preferences and item attributes, FMs are able to deliver effective recommendations, even in the absence of extensive interaction history.

**Elevated Recommendation Experience.** Foundation Models foster a transformative interface paradigm for recommendation systems, significantly altering the user interaction experience. For instance, conversational RS is a classic use case, the previous Conversational (CRSs) [9], [10] predominantly rely on pre-established dialogue templates, a dependency that often constrains the breadth and adaptability of user engagements. In contrast, FMs introduce a paradigm shift towards more dynamic and unstructured conversational interactions, offering enhanced interactivity and flexibility. The interactive design allows for more engaging and natural user interactions with the system. Users can conversationally communicate their preferences, ask questions, and receive customized recommendations.

**Improved Explanation and Reasoning Capabilities.** Foundation Models augment the explanation and reasoning capabilities of RS. Whereas traditional recommender systems predominantly derive explanations from rudimentary sources such as user reviews or elementary user behaviors, including co-purchased items or peer purchases, these explanations are often bereft of in-depth logic and context [11]. In contrast, Foundation Models possess the ability to formulate explanations that are enriched with a comprehensive grasp of commonsense and user-specific context [12]. These models leverage an array of data, encompassing user preferences, historical interactions, and distinctive item characteristics, to generate explanations that are both more

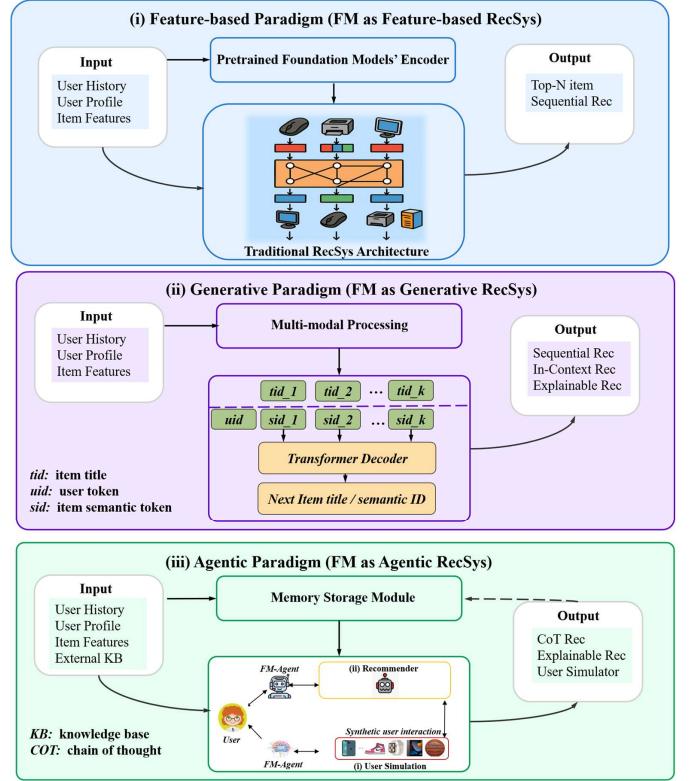


Fig. 1. Three Paradigms of FM-Powered Recommender Systems

coherent and logically sound. Utilizing Foundation Models to deeply interpret user behavior sequences and interests can significantly enhance the effectiveness of future recommender systems in complex scenarios [13], promising to advance informed and responsible decision-making processes in areas like medicine and healthcare, e.g., treatment and diagnosis recommendations.

Given these benefits, a wave of research has begun exploring FM4RecSys. As traditional RSs struggle with issues such as data sparsity and rigid feature extraction, the emergence of FMs promises broader generalization capabilities. However, realizing this potential in real-world applications brings new challenges, such as real-time adaptation, computational efficiency, and interoperability that remain underexplored. To better understand both the opportunities and limitations, we provide a comprehensive and critical assessment of FM4RecSys, organized around three core paradigms and a range of recommendation tasks.

## 1.2 Paradigms of FM-Powered Recommender Systems

How can FMs be integrated into recommender systems? We identify three paradigms of integration in current research: Feature-Based, Generative, and Agentic. These paradigms differ in the role the foundation model plays in the RS pipeline (from a passive feature provider to an active decision-maker). Figure 1 provides a high-level comparison of these paradigms with examples.

**Feature-Based Paradigm:** This approach treats foundation models as feature extractors to generate high-quality embeddings for users, items, or interactions. For example, text-based FMs (e.g., BERT) [14] encode item descriptions or

user reviews into semantic vectors, while vision-language models (e.g., CLIP) [15] align multimodal features (text, images) for cross-domain recommendations. While effective, this paradigm often limits FMs to auxiliary roles, decoupled from the core recommendation logic.

**Generative Paradigm:** This approach leverages the generative capabilities of FMs (e.g., GPT), this paradigm directly synthesizes recommendations as generated outputs [16]. Examples include generating personalized explanations [17], creating virtual items (e.g., advertising slogans, product designs), or predicting user preferences through autoregressive token prediction. However, such systems face challenges in controllability and alignment with user intent, as generation may prioritize fluency over relevance.

**Agentic Paradigm:** The emerging agentic paradigm reimagines RSs as autonomous agents powered by FMs [18]. These agents dynamically interact with users (e.g., via natural language), reason about long-term preferences, and even take actions (e.g., probing questions, multi-step planning) to refine recommendations. Unlike static models, agentic systems exhibit goal-driven behavior, leveraging tools (e.g., search engines, databases) and feedback loops to adapt to evolving contexts.

While feature-based and generative paradigms have advanced recommendation accuracy and diversity, the agentic paradigm represents a transformative shift toward proactive, explainable, and human-aligned systems. By integrating reasoning, tool use, and multi-turn interaction, agentic FMs address critical limitations of traditional RS: (i) Dynamic Adaptation: Agents continuously update user profiles based on real-time feedback, mitigating the cold-start and data sparsity issues. (ii) Multimodal Contextualization: They unify text, voice, and visual inputs to capture nuanced preferences (e.g., interpreting a user's screenshot of a product). (iii) Ethical Alignment: To balance personalization with fairness and transparency, recent studies have explored constitutional AI techniques [19], which guide model behaviors by incorporating predefined ethical principles or human-aligned rules into the generation process. The rapid progress of LLM-based agents (e.g., AutoGPT<sup>1</sup>, Meta's CICERO [20]) and retrieval-augmented generation (RAG) frameworks further demonstrates the feasibility of this paradigm.

### 1.3 Distinguishing Features from Recent LLM-based RS Surveys

Research on FM4RecSys is accelerating, and several surveys have recently reviewed parts of this emerging intersection. Liu *et al.* [21] delved into the training strategies and learning objectives of language modeling paradigm adaptations for recommenders, while Wu *et al.* [22] provided insights from both discriminative and generative viewpoints on Language Model-based Recommender Systems (LLM4Rec). Lin *et al.* [4] introduces two orthogonal perspectives: where and how to adapt LLMs in recommender systems. Fan *et al.* [23] offered an overview of LLMs for recommender systems, concentrating on paradigms such as pre-training, fine-tuning, and prompting. Lin *et al.* [24] summarized the current progress in generative recommendations, organizing them across various recommendation tasks.

1. <https://github.com/microsoft/autogen>

**Differences and Key Contributions:** In contrast to previous surveys, our survey takes a broader view: we cover foundation models beyond just LLMs, including vision and multimodal models, and structure the discussion along a new taxonomy spanning data, integration paradigms, tasks, and open challenges. In particular, we emphasize a three-paradigm framework — feature-based, generative, and agentic — for understanding how FMs can be leveraged in recommender systems, alongside a range of downstream recommendation tasks that have been tackled with FMs. As shown in Figure 2, we systematically outline the framework for using Foundation Models (FMs) for recommender systems (FM4RecSys), covering everything from the characteristics of recommendation data to specific downstream tasks. We analyzed existing publications from various perspectives and introduced new insights. Finally, we further delve into the latest unresolved questions and potential opportunities in this area.

**Criteria for Collecting Papers:** We collected over 150 papers related to Foundation Models for Recommender Systems. Initially, we searched top-tier conferences and journals such as ICLR, NeurIPS, WWW, WSDM, SIGIR, KDD, ACL, EMNLP, NAACL, RecSys, CIKM, TOIS, TORS, and TKDE to identify recent work. The primary keywords used in our search included large language models for recommender systems, generative recommendation, large language models, multi-modal recommendation, and agents for recommendation.

**Contributions of This Survey:** The aim of this survey is to conduct a thorough review of the advancements in Foundation Models for Recommender Systems (FM4RecSys). It offers a comprehensive overview that enables readers to quickly grasp and engage with the field of foundation model-based recommendations. This survey establishes the groundwork to encourage innovation in RSs and explore the depth of this research domain. It is intended for researchers and practitioners interested in RSs, providing them with a guide for selecting FMs to address recommendation tasks. In summary, the key contributions of this survey are threefold: (1) it offers a detailed review of foundational models for recommendation and introduces a classification scheme to organize and position current work; (2) It provides an overview and summary of the state of the arts; and (3) it discusses the challenges, open issues, and identifies new trends and future directions in this research area to expand the horizons of FM4RecSys research.

The rest of this paper is structured as follows: **Section 2** explores the characteristics of RS data by contrasting the foundational aspects of traditional sources—such as user interactions, sequential behaviors, and network connections—with the emerging significance of multi-modal data. **Section 3** introduces the recent advances in FMs, highlighting the strengths and limitations of these models, from their scalability and generalization capabilities to their ability in different tasks. **Section 4** discusses representation learning within FMs for RSs, analyzing the techniques employed to derive representations that encapsulate the RS data. **Section 5** examines integration approaches for FM4RecSys, focusing on strategies that incorporate foundation models into RS pipelines. **Section 6** details the specific tasks FM4RecSys models are designed to address, including top-n ranking,

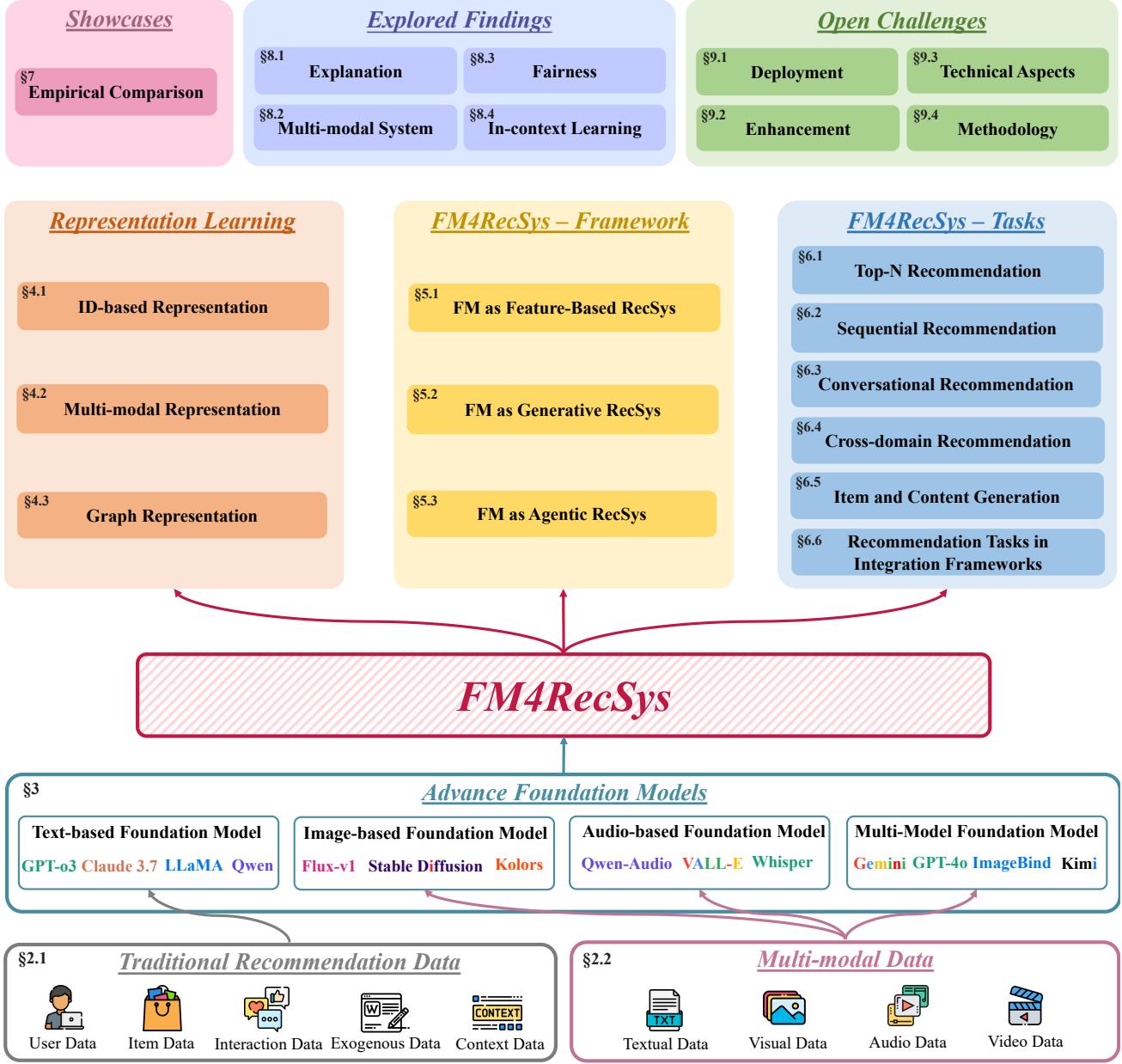


Fig. 2. The taxonomy of FM4RecSys from data characteristics to open problems and opportunities. In contrast to previous surveys, our methodology introduces a unique viewpoint for examining the intersection of FM4RecSys from data characteristics to open problems and opportunities, which are detailed in Section 1.3.

sequential recommendation, and so on, while also identifying task-specific challenges and current solutions. **Section 7** reviews empirical findings related to the opportunities and observed impacts of FM4RecSys, including their potential for enhancing scalability, efficiency, and reducing biases. In **Section 8**, we discuss open challenges and future research directions in FM4RecSys, highlighting unresolved issues in model robustness, explainability, and computational efficiency, as well as proposing directions for advancing this emerging field. Finally, **Section 9** concludes the paper by summarizing the contributions of foundation models to RSs.

## 2 TRADITIONAL RS DATA CHARACTERISTICS

This section explores the diverse types of data used in the RS field. Traditional RS rely on structured sources, such as user demographics, explicit feedback, and behavioral histories, augmented by sequential and network data that capture dynamic interaction patterns [1], [2]. In contrast, recent advances involve integrating multi-modal data, including text, images, audio, and video, which expands the scope for understanding both user preferences and item attributes. Together, these diverse sources provide a robust foundation for addressing various recommendation tasks [25].

### 2.1 Traditional RS Data

**User and item information:** In traditional recommendation systems, user and item information form the foundation

for modeling preferences and generating recommendations. User information typically includes demographic attributes (e.g., age, gender, location), explicit preferences (e.g., ratings, likes, and reviews), and behavioral history (e.g., purchase records and browsing logs) [1]. On the other hand, item information consists of metadata such as category, brand, price, and textual descriptions, as well as multimodal features derived from images, videos, or audio [1], [2]. In addition, user-item interaction matrices play a crucial role in collaborative filtering-based methods. Implicit feedback, such as clicks, dwell time, and purchase actions, is particularly valuable as it reflects natural user engagement patterns without requiring explicit ratings. Side information, including knowledge graphs and attribute embeddings, further enriches user and item representations, improving recommendation quality, particularly in sparse interaction scenarios.

**Sequential information:** Sequential information plays a key role in RS, capturing the temporal sequence of user interactions with items [26]. This data can be particularly informative when user identifiers are available, allowing for the construction of personalized recommendation models that leverage individual historical behaviors. In cases where user identifiers are absent, session-based data can be employed to model interactions within a single session, making it suitable for scenarios involving anonymous users or cold-start problems. Additionally, location-based data, such as check-ins at points of interest (POI) [27], can be integrated to provide context-aware recommendations, which are particularly effective in mobile and location-based services.

**Network data:** Network data refers to the complex relationships between entities, such as users, items, social connections, and citations, which can be exploited by RS to enhance recommendation accuracy [2]. Citation networks, which map the relationships between academic papers through citations, are commonly used in academic recommendation systems [28]. Social networks, capturing the interactions and connections between users, enable RS to recommend friends, content, or groups based on social proximity and shared interests [29].

## 2.2 Multimodal RS Data

**Textual data:** Textual data represents another common source of information in RS. Various forms of textual content, including hashtags, news articles, reviews, and so on, are used to derive rich contextual insights. Hashtags, often employed in social media, facilitate the clustering of similar content and are useful for recommending related items [30]. News articles and headlines can be analyzed to offer users recommendations that align with their reading interests [31]. User-generated reviews provide deep insights into user preferences and item attributes, which are crucial for content-based recommendation approaches, enhancing the personalization of recommendations [32].

**Visual data:** Visual data is increasingly leveraged in RS, particularly in domains where the aesthetic qualities of items are significant. Visual features extracted from images, such as color, shape, and texture, are utilized to recommend items with similar visual characteristics. This approach is particularly valuable in domains like fashion, design, and

art, where visual similarity plays a central role in user preferences [33].

**Audio data:** Audio data, particularly in the form of music, is another key input for RS. Audio features, including genre, tempo, and mood, are analyzed to generate recommendations that align with a user's listening history and preferences. This data source is integral to music streaming services, where personalized playlists and track recommendations are based on complex audio feature analysis [34].

**Video data:** Video data provides a rich source of information for RS, especially in the context of multimedia content recommendation. Features derived from video, such as visual style, genre, and content type, are used to suggest similar videos or to enhance content discovery mechanisms [35]. This data is particularly relevant for video streaming platforms, where user engagement is heavily driven by the timely and relevant recommendation of content.

**Multimodal Fusion:** Drawing from prior multi-source and multimodal data in RSs, multimodal fusion is recognized as a crucial component for FM4RecSys. Several fusion paradigms have been explored in recent literature. *Early fusion* combines modality-specific features at the input or representation level; for instance, embeddings from a product's image and textual description may be concatenated and jointly processed [36], allowing the model to capture inter-modal correlations from the outset. In contrast, *late fusion* processes each modality independently and merges their outputs at a later stage, typically via averaging, weighted summation, or gating mechanisms [37], offering flexibility but potentially missing fine-grained interactions. Hybrid fusion balances these strategies by preserving modality-specific pathways while introducing interaction layers at intermediate or output stages. For example, NOVA [36] adopts a non-invasive hybrid strategy that maintains separate branches for collaborative and content features, with final fusion at the prediction layer to preserve ID-based signals. Attention-based fusion has become a dominant paradigm for adaptive integration, where attention mechanisms dynamically assign weights to modalities or their internal components (e.g., text tokens or image patches) based on contextual relevance. *Co-attention* or cross-modal attention mechanisms are particularly effective for aligning modalities; for instance, CMBF [38] uses cross-attention to align image and text representations, capturing complementary semantics between product visuals and descriptions. These techniques are often integrated into transformer-based architectures. Building on this, *cross-modal Transformers* further enhance multimodal fusion by leveraging self-attention and cross-attention layers to learn joint representations across modalities, and have been successfully applied to both sequential and static recommendation tasks [39], [40]. Additionally, *modality-aware gating* and *graph-based fusion* offer further enhancements. Graph-based approaches model user-item interactions as graphs, where item nodes are enriched by multimodal content, allowing modality-specific message passing and collaborative signal propagation. Models such as MMGCN [37] and MGAT [41] utilize attention or gated mechanisms to adaptively integrate modality signals, while gating networks like those used in MARIO [42] dynamically adjust the influence of each modality based on user preferences or contextual signals,

enabling personalized and context-aware fusion.

#### Key Difference: Traditional and Multimodal RS data

- Traditional RS data primarily relies on fundamental user and item information such as demographics, explicit preferences, and behavioral histories, while multimodal RS data integrates diverse modalities like text, images, audio, and video to significantly broaden the information landscape.
- Traditional RS data is typically structured in clear formats like user-item interaction matrices and defined attributes, whereas multimodal RS data is often unstructured or semi-structured, necessitating specialized feature extraction and projection into a unified latent space.
- Multimodal RS data requires more complex fusion techniques than traditional RS data due to the necessity of integrating heterogeneous modalities.

### 3 RECENT ADVANCES IN FOUNDATION MODELS

In this section, we will give a brief introduction to recent advances in FMs, Multimodal Foundation Models, and Foundation Model Agents.

#### 3.1 Foundation Models

A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT [14], GPT-3 [43], and CLIP [15]. However, the sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible; for example, GPT-3 has 175 billion parameters and can be adapted via natural language prompts to do a passable job on a wide range of tasks despite not being trained explicitly to do many of those tasks. Building upon the success of GPT-3, which is the first model to encompass over 100B parameters, several noteworthy models have been inspired, including GPT-J [44], BLOOM [45], OPT [46], Chinchilla [47], and LLaMA [48]. These models follow the similar Transformer decoder structure as GPT-3 and are trained on various combinations of datasets. Owing to their vast number of parameters, fine-tuning LLMs for specific tasks, such as RS, is often deemed impractical. Consequently, two prevailing methods for applying LLMs have been established: in-context learning (ICL) [49] and parameter-efficient fine-tuning [50]. ICL is one of the emergent abilities of LLMs empowering them to comprehend and furnish answers based on the provided input context, rather than relying merely on their pre-training knowledge. This method requires only the formulation of the task description and demonstrations in natural language, which are then fed as input to the LLM. Notably, parameter tuning is not required for ICL. Additionally, the efficacy of ICL can be further augmented through the adoption of the chain-of-thought prompting, involving multiple demonstrations (describe the chain of thought examples) to guide the model's reasoning process. ICL is the most commonly used method for applying LLMs to information

retrieval. Parameter-efficient fine-tuning aims to reduce the number of trainable parameters while maintaining satisfactory performance. LoRA [50], for example, has been widely applied to open-source LLMs (e.g., LLaMA and BLOOM) for this purpose. Recently, QLoRA [51] has been proposed to further reduce memory usage by leveraging a frozen 4-bit quantized LLM for gradient computation. Despite the exploration of parameter-efficient finetuning for various NLP tasks, its implementation in RS tasks remains relatively limited, representing a potential avenue for future research.

#### 3.2 Multi-modal Foundation Models

Recently, substantial advancements in multimodal Foundation Models (MFMs) have emerged, largely augmenting standard FMs to accommodate multimodal input and output environments through economical training strategies [52]. MM-FMs leverage large language models (LLMs) as central components for semantic understanding and reasoning in multimodal tasks. With marketing models such as GPT-4(Vision) [53] and Gemini [54], [55] demonstrating exceptional multimodal understanding and generation capabilities, there is a surge of interest among researchers in Multimodal Foundation Models. Initial studies primarily revolved around multimodal content comprehension and text generation. This included image-text understanding, as seen in groundbreaking projects such as BLIP-2 [56], BLIP-3 [57], LLaVA [58], MiniGPT4 [59], and OpenFlamingo [60]. Initiatives like VideoChat [61], Video-ChatGPT [62], and LLaMA-VID [63] took this a step further with video-text understanding. The audio-text understanding capabilities of MM-LLMs were also largely explored in projects like Qwen-Audio [64] and Qwen-Audio2 [65]. Subsequent research broadened the abilities of MM-LLMs to include specific modality outputs. Image-to-text output tasks come into the picture with efforts such as Kosmos-G [66], Emu [67], and MiniGPT-5 [68], while projects like SpeechGPT [69] and AudioPaLM [70] herald the advent of speech/audio-text output. In recent times, the focus has been on imitating human-like any-to-any modality conversion, giving a glimpse into the potential path toward artificial general intelligence. Some attempts have been made to combine LLMs with external tools to accomplish comprehensive multimodal comprehension and generation, as showcased by VisualChatGPT [71], HuggingGPT [72], and AudioGPT [73]. In contrast, to minimize the cascading errors in the system, novel initiatives such as NExT-GPT [74], CoDi-2 [75], and Modaverse [76] have been developed. These provide end-to-end MM-LLMs and cover a full spectrum of modalities, signifying a promising step towards effectively modeling multimodal content in the RS field.

#### 3.3 Foundation Model Agents

The rapid evolution of LLM-based AI has spurred significant advancements in Agent AI, fundamentally reshaping how systems interact with complex environments. In recent years, researchers have equipped LLM agents with core components—memory, planning, reasoning, tool utilization, and action execution—that are essential for autonomous decision-making and dynamic interaction [77].

Single-agent systems leverage a unified model that integrates multiple interdependent modules.<sup>23</sup> The memory component acts as a structured repository that stores and retrieves contextually relevant information, such as user preferences and historical interactions [78]. This persistent memory is crucial for maintaining coherent, long-term interactions and forms the foundation for personalization in recommendation settings. The planning module is closely linked with advanced reasoning capabilities. Recent research has identified approaches such as task decomposition, multi-plan selection, external module-aided planning, reflection and refinement, and memory-augmented planning [79]. These techniques enable an agent to break down complex tasks, select and refine strategies based on evolving contexts, and leverage external knowledge sources. Integrated reasoning further enhances decision-making by allowing the system to adapt dynamically to novel scenarios. Frameworks like the ReAct [80] and Reflexion [81] exemplify how interleaving reasoning with concrete actions, such as web-browsing or tool invocation, can significantly improve system robustness and adaptability. Beyond internal cognitive processes, these agents increasingly rely on tool utilization to interface with external data and services. Systems like WebGPT [82] illustrate the effectiveness of using external modules (e.g., web search engines) to retrieve real-time information. Other works, such as Retroformer [83] and AvaTaR [84], further optimize these interactions through policy gradient optimization and contrastive reasoning, respectively, to fine-tune tool usage and enhance performance over time.

In contrast, LLM-based multi-agent systems emphasize collaboration among diverse autonomous agents. These systems are designed to mimic complex human workflows by facilitating inter-agent communication, task specialization, and coordinated decision-making. Frameworks such as CAMEL [85] and AutoGen demonstrate how agents with distinct roles can interact to solve problems more efficiently than a single, monolithic agent. By assigning specialized functions, ranging from ideation and planning to evaluation, these frameworks enable a division of labor that enhances overall system capability and flexibility. Further advancements are seen in approaches like MetaGPT [86] and AgentLite [87], which incorporate meta-programming techniques and lightweight libraries to dynamically allocate roles and coordinate complex workflows. These structured interactions not only improve task efficiency but also offer robustness in dynamic problem-solving environments. Recent developments also include systems such as ChatEval [88] and ChatDev [89], which leverage inter-agent debate and evaluative feedback to produce more nuanced and reliable outputs. This human-like discussion among agents is particularly beneficial in open-ended natural language generation tasks and complex software development processes.

## 4 REPRESENTATION LEARNING

In the pre-foundation model era, RS heavily relied on user and item representations from one-hot encoding for deep

learning models. With the advent of FM4RecSys, there is a shift towards embracing more diverse inputs such as user profiles, item side information, and external knowledge bases like Wikipedia for enhanced recommendation performance. To be specific, numerous works [90], [91] have identified that the key to building FM-based recommenders lies in bridging the gap between FMs' pre-training and recommendation tasks. To narrow the gap, existing work usually represents recommendation data in natural language for fine-tuning on FMs [92]. In this process, each user/item is represented by a unique identifier (e.g., user profile, item title, or numeric ID), and subsequently, the user's historical interactions are converted into a sequence of identifiers. FMs can be fine-tuned on these identifiers to learn their representations to excel at recommendation tasks. Current recommendation data representation methods can be categorized as ID-based representation, multi-modal representation, graph representation, and hybrid representation.

### 4.1 ID-based Representation

In FM context, recent studies on ID-based representation utilize numeric IDs like “[prefix]+[ID]” (e.g., “user\_123” or “item\_57”) to represent users and items, effectively capturing the uniqueness of items [93], [94]. Nevertheless, numeric IDs lack semantics and fail to leverage the rich knowledge in FMs. Furthermore, FMs require sufficient interactions to fine-tune each ID representation, limiting their generalization ability to large-scale, cold-start, and cross-domain recommendations. In addition, ID indexing necessitates updates to vocabularies to handle out-of-vocabulary (OOV) issues and parameter updates of the FMs that incur extra computational costs, highlighting the need for more informative representations. Meanwhile, sequential ID indexing [94] is utilized to capture the collaborative information in an intuitive way.

### 4.2 Multi-modal Representation

A promising alternative way lies in leveraging multi-modal side information, including utilizing images [95] (such as item visuals), textual content [96]–[101] (encompassing item titles, descriptions, and reviews), multi-modal elements [102]–[105] (like short video clips and music), and external knowledge sources [106]–[108] (such as item relationships detailed in Wikipedia). Yuan *et al.* [109] underscores the advantages of multi-modality-based RS when compared to ID-based counterparts, drawing attention to the performance gains, which emphasizes that richer side information about users and items can enhance performance in cross-domain and cold-start recommendation scenarios.

However, the alignment between pure item side information and user-item interactions may not always be consistent [92], [110]. In other words, two items with similar visual or textual features might not necessarily share similar interaction patterns with users. Thus, it is natural to utilize the hybrid representation that combines ID and multi-modal side information to achieve distinctiveness and semantic richness. For instance, TransRec [111] utilizes multi-faceted identifiers that combine IDs, titles, and attributes to achieve both uniqueness and semantic richness in item representation. CLLM4Rec [92] extends the vocabulary of FMs by

2. <https://github.com/huggingface/smolagents>  
3. <https://www.langchain.com/langgraph>

Representation Learning Approach	Suitable for FM Paradigm	Suitable for Recommendation Tasks	Key Benefits	Challenges
ID-Based Representation	► Primarily Feature-Based	► Top-N Recommendation ► Basic Sequential Recommendation	► High efficiency and scalability ► Easy integration with existing pipelines	► Lack of semantic depth ► Limited adaptability in complex scenarios
Multi-modal Representation	► Complementary to Feature-Based and Generative	► Cross-Domain Recommendation ► Item/Content Generation ► Conversational Recommendation	► Rich semantic context from diverse sources ► Enhanced handling of cold-start problems	► Increased preprocessing complexity ► Alignment issues and potential noise
Graph-Based Representation	► Most suitable for Generative, Agentic Framework	► Sequential Recommendation ► Cross-Domain Recommendation ► Conversational Recommendation	► Captures rich relational and contextual information ► Facilitates dynamic adaptation via interactive modeling	► High computational cost ► Scalability issues in large-scale graphs
Hybrid Representation	► Versatile across Feature-Based, Generative, and Agentic frameworks	► Applicable to a wide range of tasks including Top-N, Sequential, Conversational, Cross-Domain, and Item/Content Generation	► Balances efficiency with rich semantic expressiveness ► Leverages complementary strengths of multiple modalities	► Increased model complexity ► Challenges in effective fusion and alignment

TABLE 1

Comparative analysis of different representation learning approaches. The table shows the suitability of each approach for specific FM paradigms (Section 5) and recommendation tasks (Section 6), along with their key benefits and challenges.

incorporating user/item ID tokens and aligning them with user-item review text information through hard and soft prompting, allowing for accurate modeling of user/item collaborative information and content semantics.

### 4.3 Graph Representation

Moreover, the graphical structure inherent in user-item interactions makes graph representation a natural way to harness structured information from the user-item graph, thereby enhancing the recommendation system's ability to model complex interactions and user preferences [112]. LLM-based graph representation can primarily be categorized into two types. The first type involves encoding entity IDs (including user IDs, item IDs, and attribute IDs) using GNNs/HGNNs(Heterogeneous Graph Neural Networks) to obtain the graph representation. Then, LLMs are used to encode the textual descriptions of the entities themselves to obtain semantic text representations. These two representations are then fused to form a hybrid representation [113]–[116]. Ren *et al.* [117] incorporate auxiliary textual signals obtained through LLMs and align semantic spaces with collaborative relational graph signals. The second type first utilizes LLMs to extract the text representation of the entities. Subsequently, based on the relationships between entities, GNNs/HGNNs are used to encode and obtain the graph representation. Damianou *et al.* [118] train an HGNN on an item-item graph, where items are connected if they have been co-interacted with by the same user. Each node is associated with text embedding node features derived from LLMs applied to the item's description. The final representation combines the item's semantic information with the item-item relational representation. These integrations aim to harness the strengths of LLMs in natural language understanding and GNNs in relational data processing, resulting in a more powerful RS that can understand and recommend items.

In summary, Table 1 presents a comprehensive comparative analysis of various representation learning ap-

proaches as discussed in this section. The table categorizes ID-based, multi-modal, graph-based, and hybrid representations by aligning them with the three FM paradigms, Feature-Based, Generative, and Agentic, and further maps their applicability to specific recommendation tasks such as Top-N, Sequential, Conversational, Cross-Domain, and Item/Content Generation. This overview highlights that while ID-based representations are highly efficient and well-suited for fast, scalable applications like Top-N recommendation, multi-modal methods offer richer semantic context beneficial for cross-domain and content generation tasks. Similarly, graph-based approaches excel in capturing relational and dynamic user-item interactions, which are critical for sequential and conversational recommendations, while hybrid methods aim to combine these strengths to balance efficiency with expressive power. Overall, this table not only underscores the trade-offs associated with each representation strategy but also serves as a foundation for understanding their role within the integrated frameworks outlined in later sections.

## 5 FM4RECSYS: INTEGRATION APPROACHES

As mentioned in Section 1.2, we identified three paradigms of integration in current research: Feature-Based, Generative, and Agentic. Table 2 provides a detailed comparison of these paradigms from different perspectives. Next, we delve deeper into each paradigm, surveying representative methods and discussing their strengths and limitations.

### 5.1 Feature-Based Paradigm: Foundation Models as Feature Enhancers

As shown in Figure 3, existing works along this line can be categorized into two directions: (i) FM Embeddings for RSs, where FMs act as feature encoders to generate high-quality user/item embeddings for conventional recommendation models; and (ii) FM Tokens for RSs, where FMs generate

FM Paradigm	Capabilities	Tasks	Key Benefits	Challenges
Feature-Based RecSys	<ul style="list-style-type: none"> <li>FM embeddings enhance user-item representation.</li> <li>Improves similarity matching &amp; cold-start handling.</li> </ul>	<ul style="list-style-type: none"> <li>Top-N Recommendation</li> <li>Sequential Recommendation</li> </ul>	<ul style="list-style-type: none"> <li>Better generalization to unseen items.</li> <li>Cold-start user/item adaptation.</li> <li>Improved user profiling.</li> </ul>	<ul style="list-style-type: none"> <li>Limited Reasoning Capability</li> <li>High inference cost of feature computation due to FM size</li> </ul>
Generative RecSys	<ul style="list-style-type: none"> <li>Generates personalized recommendations instead of ranking.</li> <li>Explains recommendations in natural language.</li> </ul>	<ul style="list-style-type: none"> <li>Sequential Recommendation</li> <li>IN-context Recommendation</li> <li>Conversational Recommendation</li> <li>Explainable Recommendation</li> </ul>	<ul style="list-style-type: none"> <li>Self-adaptive item generation.</li> <li>Personalized user experience.</li> <li>Multimodal RS enabled (text, image, audio).</li> </ul>	<ul style="list-style-type: none"> <li>Bias Amplification in Generation</li> <li>Explainability &amp; Trust Issues</li> <li>High inference cost due to FM size</li> </ul>
Agentic RecSys	<ul style="list-style-type: none"> <li>Acts as a decision-making agent.</li> <li>Remembers user history &amp; self-improves.</li> <li>Supports multi-turn interactions &amp; goal-driven planning.</li> </ul>	<ul style="list-style-type: none"> <li>Long-term (Sequential) Personalization</li> <li>Multi-turn Adaptive Recommendation</li> <li>Planning &amp; Memory-Driven RS</li> </ul>	<ul style="list-style-type: none"> <li>Context-aware and real-time learning.</li> <li>Continuous user adaptation &amp; memory.</li> <li>Proactive reasoning &amp; negotiation.</li> </ul>	<ul style="list-style-type: none"> <li>Scalability &amp; Real-time Inference Cost</li> <li>Evaluation Challenge of Agent Effectiveness</li> </ul>

TABLE 2

Table 2 provides a detailed comparison of foundation model-powered recommender systems (FM4RecSys) across three major paradigms: Feature-Based, Generative, and Agentic. It summarizes their capabilities, tasks, key benefits, and challenges.

semantic-aware tokens or indices to facilitate token-level generation and retrieval in recommendation.

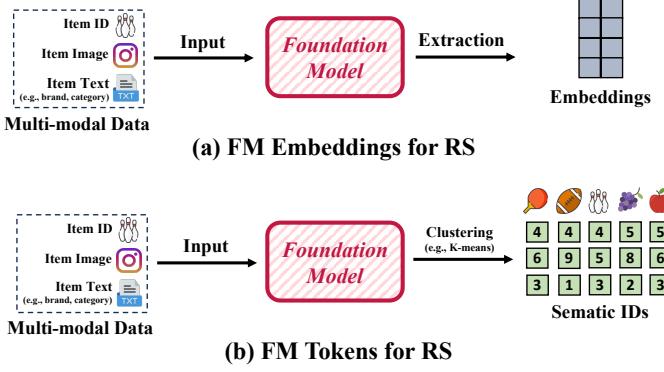


Fig. 3. Examples of FM embeddings and tokens for RSs.

**FM Embeddings for RS:** This modeling paradigm views the language model as a feature extractor, which feeds the features of items and users into LLMs and outputs corresponding embeddings. A traditional RS model can utilize knowledge-aware embeddings for various recommendation tasks. RLMRec [117] treats LLMs as a text encoder to map items or users into a semantic space and aligning their semantic space with collaborative relation modeling for better representation learning. They leverage the advanced text comprehension abilities of LLMs to capture the nuanced semantic aspects of user behaviors and preferences. AlphaRec [119] employs linear mapping to project language representations of item titles into a behavior space for recommendation. Those language representations are transformed into high-quality behavior representations, resulting in outstanding recommendation performance. LLMRec [120] uses LLMs for graph augmentation for RecSys by augmenting user-item interaction edges, item node attributes, and user node profiles. This work addresses the scarcity of implicit feedback signals by enabling LLMs to explicitly reason about user-item interaction patterns. BinLLM [121] encodes collaborative information textually for LLM-based recommendation by converting collaborative

embeddings into binary sequences. Specifically, BinLLM transforms collaborative embeddings from external models into binary sequences, a text format compatible with LLMs, enabling the direct utilization of collaborative information in a text-like format. iDreamRec [122] makes further steps to combine LLMs embeddings with Diffusion Models [123] for recommendation tasks by incorporating text embeddings from LLMs to accurately model item distributions. Specifically, given the metadata about items (e.g., titles), this work prompts GPT to generate detailed textual descriptions, offering richer content than simple item IDs.

**FM Tokens for RS:** This modeling paradigm generates tokens based on the input items' and users' features. The generated tokens capture potential preferences through semantic mining, which can be integrated into the decision-making process of a recommendation system. Recent studies have introduced methods in which items are represented by semantically meaningful tokens, enabling more accurate and context-aware recommendations. For instance, the TIGER framework [16] proposes a generative retrieval approach that autoregressively decodes item identifiers. In this framework, each item is assigned a *Semantic ID*—a tuple of codewords derived from the item's content features—thus allowing the system to predict the next item a user might interact with based on previous interactions. Similarly, the LC-Rec model [124] addresses the semantic gap between large language models and recommender systems by integrating both language and collaborative semantics. It employs a learning-based vector quantization method to assign meaningful item indices, enabling the language model to generate items directly from the entire item set without relying on predefined candidates. Other frameworks extend these ideas further. For example, ColaRec [125] combines content information and collaborative signals within a unified sequence-to-sequence generative framework, while the EAGER framework [126] integrates behavioral and semantic information through a two-stream generative architecture. Moreover, the COBRA framework [127] adopts a cascaded approach that alternates between sparse semantic IDs and dense vectors to capture both semantic insights and collaborative signals from user-item interactions. OneRec [128] ex-

plores related techniques to further enhance recommender systems by leveraging token generation via large language models, thereby highlighting the potential of semantic tokenization in this domain.

## 5.2 Generative Paradigm: Foundation Models as Generators of Recommendations

Different from the feature-based paradigm, the generative paradigm formulates recommendations as a generative task and provides diverse recommendations to meet user preferences. Specifically, it transforms pre-trained generative models into powerful end-to-end recommendation systems. The input of FMs typically includes varying user preference information, such as profile description, behavior prompt, and task instruction, while the output is expected to generate reasonable recommendations. In this section, we mainly focus on how to transfer an FM into a recommender, and thereby classify related research works into three categories: (1) Pre-trained FM for RS, (2) Non-tuning FM4RecSys, and (3) Fine-tuning FM4RecSys as shown in Figure 4.

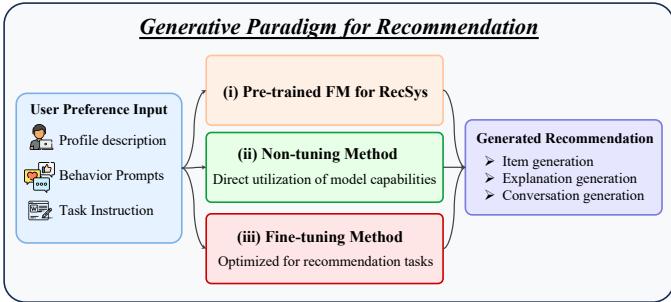


Fig. 4. An illustration of the generative paradigm for recommendation. User preference inputs (e.g., the profile description, behavior prompts, and task instructions) are utilized to guide the pre-trained foundation models (FM) for RS. The model can be leveraged in a non-tuning manner by directly utilizing its capabilities or via fine-tuning for specific recommendation tasks, producing various forms of generated recommendations such as item generation, explanation generation, and conversation generation.

**Pre-trained FM4RecSys:** Few works like M6 Rec [129], PTUM [130] and RecGPT [131] pre-train the whole model on massive recommendation datasets by adopting transformer-based models for next-item prediction and applying different language modeling tasks, such as masked language modeling, permutation language modeling, and so on. This line of work generally requires a large amount of domain data for RS, leading to high training costs.

**Non-tuning FM4RecSys:** Foundation models (FMs) have demonstrated strong zero- and few-shot capabilities across many tasks [43], [132], [133]. Consequently, recent studies assume that FMs inherently possess recommendation capabilities and aim to activate these abilities through the use of tailored prompts where FMs parameters remain unchanged. Due to the rich textual side information, prompting with LLMs has demonstrated powerful recommendation ability. Non-tuning FM4RecSys focuses on designing appropriate prompts to stimulate the recommendation abilities of LLMs. Liu *et al.* [134] propose a prompt construction framework to evaluate the ability of ChatGPT on five common recommendation tasks, providing zero-shot and few-shot versions for each type of prompt. He

*et al.* [8] not only use prompts to evaluate the ability of LLMs on sequential recommendation but also introduce recency-focused prompting and in-context learning strategies to alleviate order perception and position bias issues of LLMs. More recently, some works [17] have focused on designing novel structures in prompting for FM4RecSys. Yao *et al.* [135] include heuristic prompts including item attributes in natural language, along with collaborative filtering information presented through text templates and knowledge graph reasoning paths. Similarly, Rahdari *et al.* [136] craft hierarchical prompt structures that encapsulate information about recommended items and top-k similar item information in the user interaction history. Hou *et al.* [8] leverage the zero-shot and in-context learning capabilities of LLMs to construct prompts for sequential movie recommendations. However, using LLMs as recommender systems without pre-training or fine-tuning, only depending on in-context learning, still lags behind traditional supervised methods like SASRec [137] in sequential tasks [8]. The primary reason is that LLMs have limited ability to perceive the order of historical interaction sequences. As the length of these sequences increases, the performance of LLM recommendations tends to decline. Therefore, optimizing the long-sequence modeling capability of LLMs presents a potential opportunity to enhance their effectiveness in recommendation scenarios.

**Fine-tuning FM4RecSys:** This line of work generally requires a large amount of domain-specific data for retraining the FMs, leading to extra computational costs. Although FMs generally exhibit strong zero- and few-shot capabilities, it is unsurprising that they may fall short of outperforming recommendation models specifically trained on task-related data for a given task. Therefore, one straightforward approach is to fine-tune powerful FMs, which contain world knowledge, using task-specific data for downstream recommendation tasks. Recently, there has been extensive exploration of fine-tuning large language models (LLMs) for recommendation tasks. TCF [138] adopts and fine-tunes LLMs to create a universal item representation for recommendation tasks, contrasting this approach with the increasingly popular prompt-based approach using ChatGPT. Unfortunately, despite utilizing an item encoder with tens of billions of parameters, it still requires re-adaptation for new data to achieve optimal recommendations. Furthermore, this type of model has not demonstrated the strong transferability that was expected, suggesting that constructing large-scale foundational recommender models may be more challenging than in the fields of CV and NLP. InstructRec [139] designs abundant instructions for tuning, including 39 manually designed templates with preference, intention, task form, and context of a user. After instruction tuning, LLMs can understand and follow different instructions for recommendations. TallRec [140] uses LoRA [141], a parameter-efficient tuning method, to handle the two-stage tuning for LLMs. It is first fine-tuned on the general data of Alpaca [142], and then further fine-tuned with the historical information of users. It utilizes item titles as input and shows effectiveness for cold-start recommendations. BIGRec [143] emphasizes that LLMs struggle to integrate statistical data such as popularity and collaborative filtering because of their inherent semantic biases. To address this, BIGRec fine-

tunes LLMs through instruction tuning to produce tokens that symbolize items. However, aligning LLM outputs with real-world items is challenging due to their inventive nature. BIGRec subsequently aligns these generated tokens with real items in the recommendation database by incorporating statistical data like item popularity. DEALRec [144] has introduced a data pruning method that employs two scores: the influence score, which estimates the impact of sample removal on performance using a small surrogate model, and the effort score, which prioritizes challenging samples for LLMs. This approach enables efficient fine-tuning of LLMs, thereby enhancing both efficiency and accuracy.

Meanwhile, Diffusion Models (DMs) [123], [145] are another generative FMs that have recently started being applied to customize the best visual recommendation content to different users. It is non-trivial to consider both the user's potential preferences, based on historical and contextual information, and the visual coherence and correlation of content. The emergence of visual FMs, particularly Stable Diffusion [145], offers a promising direction for automating and even personalizing item display content generation. Di-Fashion [146] can not only generate complementary fashion items but also create personalized outfit images from scratch based on user preferences. The model fine-tunes the latest Stable Diffusion model to ensure high fidelity, compatibility, and personalization in the generated fashion images. Ad-Booster [147] introduces the generative creative optimization task and leverages user interest signals to personalize advertisement creative generation using the inpainting technique of Stable Diffusion. The subsequent work, CG4CTR [148], further refine the solution by introducing a new automated creative generation for the click-through rate (CTR) optimization pipeline. Specifically, it employs the inpainting mode of Stable Diffusion to generate background images while preserving the main product details. More recently, DynaPIG [149] leverages diffusion models to generate visually appealing personalized product images, enhancing user engagement with recommendations. However, since diffusion-based FMs are primarily designed for vision tasks, they are not inherently suited for item recommendation. Many recent diffusion-based works [150]–[153] tend to train generative recommender from scratch for domain-specific tasks. Despite these advancements, a significant gap remains in effectively integrating the strengths of diffusion-based FMs with broader frameworks for recommendation. To bridge this gap, researchers have begun exploring how FMs can act not only as generative models but also as central components in autonomous systems, enabling more dynamic and interactive recommendation experiences.

### 5.3 Agentic Paradigm: Foundation Models as Interactive Recommender Agents

In the FM-powered autonomous agent system, FMs function as the agent's brain, complemented by key components like planning, memory, and tool use [154]. There are many inspiring works, such as AutoGPT and BabyAGI proving the FM-based agents' potential. These agents can store their past experiences and make better decisions for future behaviors. In RS scenarios, agents are typically represented as either User Simulators or the Recommender System itself, as shown in Figure 5.

**Agent as User Simulator:** This paradigm uses agents to simulate user behaviors for real-world recommendations. Gathering sufficient and high-quality user behavior data is expensive and ethically complex. Furthermore, real-world user interaction data is often very sparse, such as in the case of cold-start users. Besides, traditional methods [155], [156] often face challenges in simulating complex user behaviors due to models' capabilities, whereas FMs have demonstrated potential in this area [157]. Consequently, employing personalized agents powered by FMs for RSs emerges as a logical and effective strategy. Wang *et al.* [157] treat each user as an FM-based autonomous agent within a virtual simulator named RecAgent. This simulator allows for the free interaction, behavior, and evolution of different agents, taking into account not only actions within the RS, like item browsing and clicking, but also external factors like social interactions. Zhang *et al.* [158] further investigate the extent to which FM-empowered generative agents can accurately simulate real human behavior for movie recommendation. They design Agent4Rec, a recommender system simulator with 1,000 LLM-empowered generative agents interacting with personalized movie recommendations in a page-by-page manner with various actions. Since previous FM methods perform poorly in long-term recommendations, Shi *et al.* [159] propose a Bi-level Learnable LLM Planner (BiLLP) framework to enhance these systems. BiLLP leverages LLMs to balance short-term and long-term user satisfaction by combining macro-learning for high-level planning and micro-learning for personalized actions. This framework shows significant potential in addressing the challenges associated with long-term recommendation planning. After that, [160] treats both users and items as agents and enables a collaborative learning process that optimizes the interactions between agents. Recently, Zhang *et al.* [161] propose the USimAgent that can simulate users' querying, clicking, and stopping behaviors during search sessions through LLMs, and thus, is capable of generating complete search sessions for specific search tasks. BASES [162] utilizes LLM-based agents for large-scale user simulation in web search, generating diverse user profiles and search behaviors. It has demonstrated effectiveness through evaluations on both Chinese and English benchmarks. Meanwhile, Huang *et al.* [163] propose the LLM Interaction Simulator (LLM-InS), which simulates user behavior patterns based on content features. This approach transforms cold items into warm items, addressing the challenge of recommending cold items due to the lack of historical user interactions. Specifically, the LLM-InS simulator essentially functions as a CTR (Click-Through Rate) model. It takes information about a specific user and a specific item as input and predicts whether the user will click on the item. For a cold-start item, a subset of users is "recalled," and the simulator predicts whether these users will click on the cold-start item, thereby generating interaction data. This simulated interaction data is then used to update the item embeddings.

**Agent as RecSys:** This paradigm leverages the robust capability of FMs, including reasoning, reflection, and tool usage for recommendation. RAH [164] improves alignment with user personalities and mitigate biases by incorporating FM-based agents and a Learn-Act-Critic loop. Then, Wang *et al.* [165] first introduce a Self-Inspiring planning algorithm

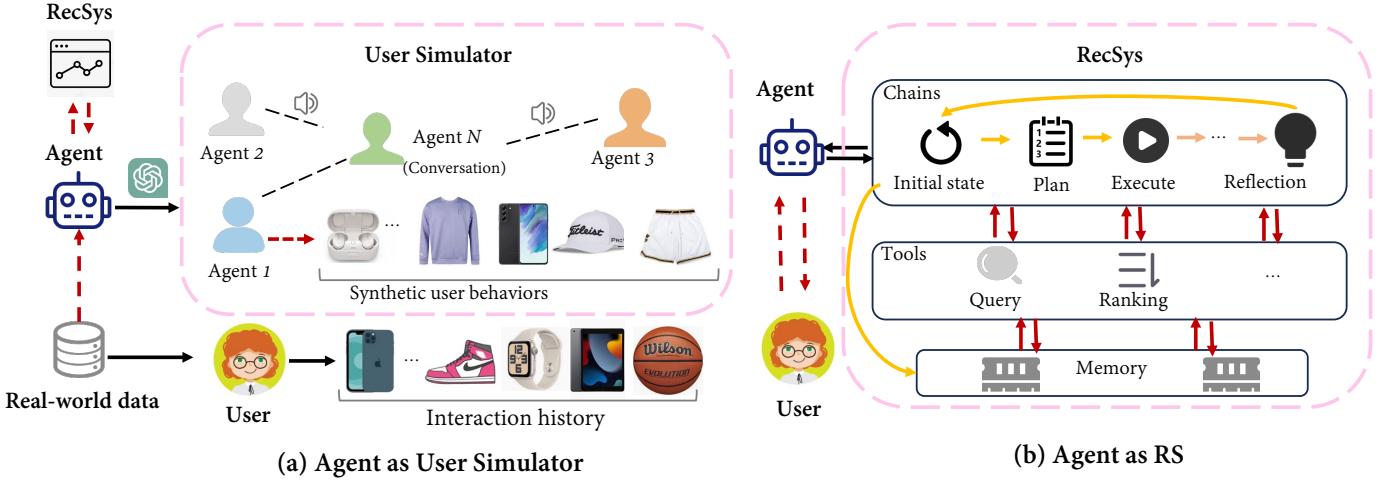


Fig. 5. Two types of personalized agents in FM4RecSys: (a) Agent as User Simulator and (b) Agent as Recommender System.

that keeps track of all past steps of the agent to help generate new states. At each step, the agent looks back at all the paths it has taken before to figure out what to do next. This approach aids in employing databases, search engines, and summarization tools, combined with user data, for producing tailored recommendations. InteRecAgent [166] models the FMs as the brain and recommendation models as tools providing domain-specific knowledge, enabling FMs to parse user intent and generate responses. They specify a core set of tools essential for RS tasks, Information Query, Item Retrieval, and Item Ranking, and introduce a candidate memory bus, allowing previous tools to access and modify the pool of item candidates.

**Agent both for Simulator and RecSys:** More recently, to address the gap in multi-agent collaboration within recommendation systems, Wang *et al.* [167] introduce MACRec, a framework that enhances recommendation tasks through the collaboration of specialized agents like Manager, User/Item Analyst, Reflector, Searcher, and Task Interpreter. MACRec can be applied to various tasks, including rating prediction, sequential recommendation, conversational recommendation, and explanation generation, tackling recommendation tasks through the collaboration of various agents. Cai *et al.* [168] propose the PUMA framework, which uses a memory system to retrieve relevant past user interactions, enhancing the agent's ability to align actions with user preferences.

To summarize, as shown in Table 3, the simulation-oriented work focuses on using agents to simulate user behaviors and item characteristics in RSs. This line of research seeks to enhance the understanding of user preferences but lacks integration into RSs. The goal of recommender-oriented studies is to build a “recommender agent” with planning and memory components to tackle recommendation tasks.

#### 5.4 Critical Analysis

**Feature-based RecSys:** In the feature-based framework, foundation models are mainly used as high-quality feature extractors to generate improved embeddings for users and

items. This approach benefits from leveraging vast pre-trained knowledge, which enhances representation quality and facilitates better ranking in tasks like Top-N recommendation and sequential recommendation. However, a critical drawback is that these models function in an auxiliary capacity and remain largely decoupled from the core recommendation decision process. This separation limits the system’s ability to dynamically adapt to contextual shifts or user-specific feedback. Additionally, the computational overhead associated with employing large-scale foundation models for embedding extraction can be a concern in real-world scenarios. Although the modularity and ease of integration make the feature-based approach attractive for enhancing existing systems, its limited reasoning and interactive capabilities constrain its application in more complex, dynamic environments.

**Generative RecSys:** The generative framework redefines recommendation by transforming it into an end-to-end natural language generation problem. This approach takes advantage of foundation models’ inherent ability to generate personalized recommendations, natural language explanations, and even novel items, which proves beneficial in zero- or few-shot settings. Despite its promise, the generative paradigm faces significant challenges in terms of output controllability and alignment with user intent. The models’ emphasis on fluency can sometimes lead to recommendations that are less precise or relevant (e.g., the OOV problem: generated items are out of vocabulary), and the qualitative nature of generated content makes it difficult to evaluate performance using standard ranking metrics. Moreover, the training and inference processes in generative methods tend to be resource-intensive, thereby raising concerns about latency and scalability in practical applications. Thus, while the generative approach offers innovative avenues for personalization and explanation, its challenges in output quality control and computational cost must be addressed to ensure its viability in real-world recommendation systems.

**Agentic RecSys:** Agentic frameworks treat the recommender system as an autonomous agent capable of interactive decision making, memory retention, and real-time plan-

Model	Objectives	Single-type Agents	Multi-type Agents	Diverse Rec. Scenarios	Open-source
RecAgent [157]	User Simulation	✓			✓
Agent4Rec [158]	User Simulation	✓			✓
LLM-Ins [163]	User Simulation	✓			
PMG [169]	User Simulation	✓			✓
BiLLP [159]	User Simulation	✓			✓
BASES [162]	User Simulation	✓			
USimAgent [161]	User Simulation	✓			
AgentCF [160]	U-I Inter Simulation		✓		✓
WebAgent [168]	U-I Inter Simulation		✓	✓	✓
RAH [164]	Recommender		✓		
RecMind [165]	Recommender	✓		✓	
InteRecAgent [166]	Recommender	✓			
MACRec [167]	Recommender	✓	✓	✓	✓

TABLE 3

Comparison among Foundation Model Agents for RecSys. Note that *Single-type Agents* indicate all agents serve the same role (e.g., users), while *Multi-type Agents* refer to agents having multiple roles and capabilities (e.g., managers, reflectors).

ning. This paradigm promises significant advancements by engaging in multi-turn dialogues, incorporating feedback, and even utilizing external tools to refine recommendations. However, its complexity poses substantial challenges. The integration of memory and planning components, while theoretically enabling dynamic adaptation, introduces issues regarding scalability and real-time performance. The unpredictable nature of autonomous decision-making can lead to inconsistencies in recommendation quality, and managing long-term user satisfaction while maintaining immediate responsiveness is a non-trivial problem. Additionally, the higher computational burden and the difficulties in designing robust evaluation methods for interactive systems are notable obstacles. Despite these limitations, the agentic approach is particularly promising for creating personalized experiences that more closely mimic human-like reasoning and decision-making.

### Comparative Discussion

- **Feature-Based Framework:** Excels in simplicity and producing high-quality semantic embeddings, but lacks dynamic adaptability and interactive reasoning.
- **Generative Framework:** Provides personalized and context-rich outputs, yet struggles with output controllability and entails high computational costs.
- **Agentic Framework:** Offers advanced interactive capabilities and real-time adaptation, though it is hindered by complexity and scalability concerns.

## 6 FM4RECSYS – TASKS

In this section, we first revisit the general formulation of the main recommendation tasks, including Top-N recommendation, sequential recommendation, conversational recommendation, and cross-domain recommendation, followed by the recent progress of FM4RecSys.

### 6.1 Top-N Recommendation

The Top-N recommendation task aims to generate a ranked list of the most relevant items for a user, typically repre-

senting the top-N items deemed most suitable based on their preferences [217]. However, if user information (including meta-information and item interaction history) is overly lengthy, it may exceed the input length capacity of foundation models. To address this, one way is to leverage a feature-based paradigm, using foundation models' embeddings to replace traditional user and item embeddings to execute the Top-N recommendation [4], [207]. However, the limitation of this approach is that it often ignores the rich contextual semantics and lacks the generative capabilities needed to generalize to unseen users or items. Another way is that foundation models use a prompt that only includes user information, asking the foundation models to directly generate recommendations for those users [93], [212]. In the case of multimodal and generative representation methods, the generated recommendation items can undergo similarity calculations with the multi-modal representation of ranking candidates [215]. Additionally, some approaches [213], [214] follow practices from the NLP field. They select K-item negative samples or hard examples, feed them along with user prompts to FMs, and obtain the final rankings. However, these approaches target the idealized experimental scenario and may not be practical for real-world recommendation systems with millions of items. More recently, the Llama4Rec [216] framework synergistically integrates both ID and text representation through data and prompt augmentation strategies and an adaptive aggregation module, resulting in significant improvements in recommendation performance.

### 6.2 Sequential Recommendation

Various FM-based approaches have been proposed to exploit their capability in the realm of context-aware recommendation. Not only can the extensive world knowledge stored in FMs serve as a rich source of background information for items [207], but also can the reasoning capability of FMs augment the next item prediction [107], [208]. Jesse *et al.* [207] first explores three different methods of utilizing foundation models' knowledge for context-aware recommendations, based on FM semantic similarity,

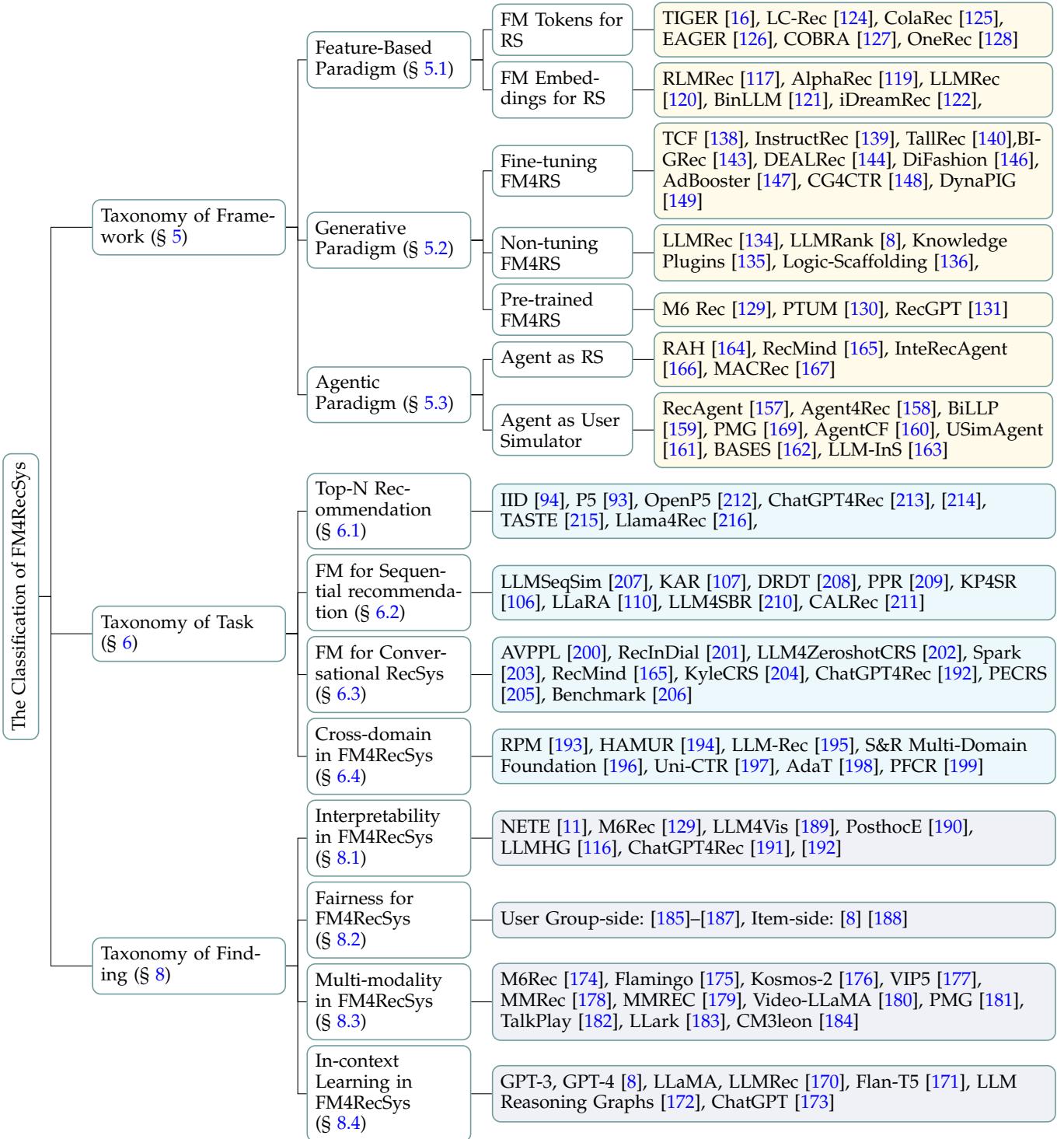


Fig. 6. The taxonomy of Foundation Model (FM) for Recommendation Systems (FM4RecSys). Representative works are shown under each sub-category for reference.

FM prompt fine-tuning, and BERT4Rec initialized by FM semantic embedding. After that, Artun *et al.* [218] propose three orthogonal methods: LLMSeqSim, LLMSeqPrompt, and LLM2Sequential, for leveraging FMs in sequential recommendation, along with hybrid strategies that combine their strengths based on item popularity and session context. Extensive experiments across multiple datasets demonstrate that LLM-enhanced models significantly improve

accuracy, diversity, and coverage, with fine-tuned GPT-3.5 notably outperforming PaLM 2 in next-item prediction tasks. Wu *et al.* [209] generate personalized soft prompts using user profile knowledge and employ prompt-oriented contrastive learning for effective training. After that, Zhai *et al.* [106] introduce knowledge prompt-tuning for sequential recommendations, which effectively integrates external knowledge bases with FMs, transforming structured

knowledge into prompts to refine recommendations by bridging semantic gaps and reducing noise. Then, Liao *et al.* [110] employ a hybrid approach for item representation in input prompts for FMs, combining ID-based item embeddings from traditional recommenders with textual item features, bridging the modality gap between traditional recommender systems and FMs through an adapter, and facilitating the transfer of user behavioral knowledge to the FM's input space. The LLM4SBR [210] framework transforms session data into a bimodal form of text and behavior, leveraging large language models (LLMs) for enhanced inference and alignment. Meanwhile, Wang *et al.* [208] utilize the reasoning capability of Foundation Models (FMs) and introduce a collaborative in-context demonstration retrieval method, abstracting high-level user preferences and reducing noise to improve the recommendation process without the need for FM fine-tuning. More recently, Li *et al.* [211] proposed CALRec, a contrastive-aligned generative framework for adapting LLMs to sequential recommendation. It uses two-stage fine-tuning: first, a two-tower setup with contrastive and language modeling losses on multi-domain data; then, fine-tuning on the target domain. However, CALRec struggles in cold-start scenarios, often generating item descriptions memorized from training.

### 6.3 Conversational Recommendation

The goal of conversational recommendation is to not only to suggest items to users over multiple rounds of interactions, but also to provide human-like responses for multiple purposes such as preference refinement, knowledgeable discussion, or recommendation justification [219], [220]. The emergence of FMs has undoubtedly impacted conversational RSs, especially CRS-related research. He *et al.* [202] presents empirical evidence that FMs, even without fine-tuning, can surpass existing conversational recommendation models in a zero-shot setting. After that, a series of works [165], [192], [203], [204] adopt the role-playing prompt to guide ChatGPT/GPT-4 in simulating user interaction with conversational recommendation agents. These works augment FMs' capability through techniques such as RAG and Chain-of-Thought (CoT). Meanwhile, several studies are built based on prior work [221] in knowledge graph-based conversational recommendation. For instance, Wang *et al.* [201] introduce a framework that integrates pre-trained language models like DialoGPT with a knowledge graph to generate dialogues and recommend items, showcasing how FM's generative capability can be utilized for conversational recommendation. Zhang *et al.* [200] explores a user-centric approach, emphasizing the adaptation of FMs to users' evolving preferences through graph-based reasoning and reinforcement learning. However, most methods rely on external knowledge graphs, require additional data labeling, and may suffer from training inefficiencies and semantic misalignment issues. In contrast, Mathieu *et al.* [205] propose PECRS, a unified and parameter-efficient conversational recommender system (CRS). PECRS formulates CRS as a natural language processing task, directly leveraging one pre-trained FM to encode items, understand user intent, perform item recommendations, and generate dialogues. Recently, Wang *et al.* [206] critique the current evaluation

protocols for conversational RSs and introduce an FM-based user simulator approach, iEvalM, which significantly enhances evaluation accuracy and explainability. However, FMs for conversational recommendation are still hindered by a tendency towards popularity bias and sensitivity to geographical regions [222]. Meanwhile, in the context of multi-turn conversational recommendation, determining the appropriate timing for state transitions during human-RS interactions is a significant challenge. For example, deciding whether to continue the conversation with the user or make a recommendation at a given moment is crucial. A key issue that needs addressing is how to effectively model a state checker using an FM to handle these decisions.

### 6.4 Cross-domain Recommendation

In real-world scenarios, data sparsity is a pervasive issue for Collaborative Filtering (CF) recommender systems, as users rarely rate or review a broad range of items, particularly new ones. Cross-domain recommendation (CDR) tackles this by harnessing abundant data from a well-informed source domain to enhance recommendations in a data-scarce target domain. Multi-domain recommendation (MDR) extends this concept by utilizing auxiliary information across multiple domains to recommend items within those domains to specific users [223]. However, domain conflicts remain a significant hurdle, potentially limiting the effectiveness of recommendations. The advent of foundation models that are pre-trained on extensive data across various domains and possess the cross-domain analogical reasoning ability [193] presents a promising solution to these challenges.

HAMUR [194] designs a domain-specific adapter to be integrated into existing models and a domain-shared hyper-network that dynamically generates adapter parameters to tackle the mutual interference and the lack of adaptability in previous models. Tang *et al.* [195] discuss the application of FMs in multi-domain recommendation systems by mixing the user behavior across different domains, concatenating the title information which items into a sentence, and modeling the user behavior with a pre-trained language model, which demonstrates the effectiveness across diverse datasets. The S&R (Search and Recommendation) Multi-Domain FM [196] employs FMs to refine text features from queries and items, improving CTR predictions in new user or item scenarios. KAR [107] further leverages the power of FMs for open-world reasoning and factual knowledge extraction, and adaptation. It introduces a comprehensive three-stage process encompassing knowledge, reasoning and generation, adaptation, and subsequent utilization. Based on the S&R Multi-Domain FM, Uni-CTR [197] employs a unique prompting strategy to convert features into a prompt sequence that FMs can use to generate semantic representations, capturing commonalities between domains while also learning domain-specific characteristics through domain-specific networks. More recently, Fu *et al.* [198] investigate the efficacy of adapter-based learning for CDR, which is designed to leverage raw item modality features, like texts and images, for making recommendations. They conduct empirical studies to benchmark existing adapters and examine key factors affecting their performance. How-

ever, it is worth mentioning that CDR faces challenges of domain privacy leakage and ineffective knowledge transfer in existing FM4RecSys methods. To address these issues, Guo *et al.* [199] proposed the PFCR framework, which introduces a privacy-preserving federated learning schema using local client interactions and gradient encryption. This framework models items in a universal feature space through description texts and leverages federated content representations with prompt fine-tuning strategies.

## 6.5 Item and Content Generation for Recommendation

A forward-looking application of FMs in RS is item generation – creating new content that can be recommended to the user. This goes beyond the classical remit of RS, which typically selects from existing items. However, with generative models, the line between recommending an existing item and generating a new item tailored to the user can blur. The main tasks of item and content generation as follows:

**Bundle Generation:** In E-commerce, bundle recommendation is crucial for increasing average order value. BundleGen [224] introduces a diffusion-based framework to generate item bundles conditioned on a seed item and user preference. Instead of retrieving co-purchased items, it learns to generate a set of compatible, stylistically coherent items using a denoising process in the item embedding space. The model iteratively refines random noise into a meaningful bundle, optimizing for both intra-bundle compatibility and personalization. Compared to autoregressive models, diffusion models offer better control over diversity and global structure, especially in high-dimensional item spaces.

**Playlist Generation:** In music recommendation, playlist generation is evolving from heuristic-based sequencing to fully generative paradigms. Recent work, such as MusicGen [225], fine-tunes pre-trained language models like GPT-2 to autoregressively generate sequences of song IDs, conditioned on a user’s history or intent prompt (e.g., “relaxing jazz for evening”). These models treat playlist generation as a language modeling task over item IDs, allowing flexible incorporation of user intent, mood, or temporal context. By training on curated playlists, such models capture high-order dependencies and smooth transitions between tracks, outperforming retrieval-based baselines on diversity and coherence metrics.

**Text Content Recommendations:** In news recommendation, generative models are used to go beyond list ranking by enabling LLMs to generate summaries or narratives that align with a user’s interest profile. GNR (Generative News Recommendation) [226] adopts a two-stage pipeline: first retrieving relevant news articles, then using GPT-based models to synthesize a cohesive summary that connects the stories under a unifying theme. This approach transforms passive article lists into engaging narratives. Similarly, Prompt4NewsRec [227] uses prompt-based generation to model user interest evolution and recommend news in natural language, leveraging personalization at the prompt level.

**Creative Content Personalization:** Many recent content-aware works leverage visual FMs, such as Stable Diffusion [145], to generate personalized recommendation content. DiFashion [146] creates personalized outfit images from

scratch based on user preferences. AdBooster [147] leverages user interest signals to personalize ad creative generation based on the Stable Diffusion. CG4CTR [148] further employs Stable Diffusion to generate background images while preserving the main product details as diverse ads for different users. DynaPIG [149] leverages diffusion-based FMs to create visually appealing personalized product images.

**New Item Cold-Start as Generation:** Cold-start items lack behavioral data, but generative models can hallucinate metadata or simulate interactions. Acharya *et al.* [228] use few-shot prompting with LLMs to generate item descriptions (e.g., genre, plot, key attributes) for new movies or books based on minimal metadata (like title or category), improving recall when fed into traditional recommenders. ColdLLM [229] proposes a two-stage framework: a lightweight filtering model selects promising user candidates, and then GPT-4 simulates interactions (clicks, ratings) between these users and cold-start items. These synthetic logs are fed into downstream models, yielding strong gains in offline metrics and GMV in real-world A/B tests. These approaches demonstrate the generative capability of LLMs in bootstrapping cold-start items without any real interactions.

## 6.6 Discussion

In this section, we analyze the suitability of the three integration frameworks for various recommendation tasks. By comparing the inherent capabilities of Feature-Based, Generative, and Agentic paradigms, we can identify which framework is best suited for each type of task and where potential hybrid approaches might be beneficial.

For the Top-N Recommendation task, which primarily involves ranking and matching based on user preferences, the Feature-Based framework has a clear advantage. This approach focuses on extracting high-quality embeddings that capture semantic similarities, allowing for efficient and accurate ranking. While Generative methods can also operate in zero- or few-shot settings, they tend to suffer from issues of output controllability, and Agentic systems often introduce unnecessary complexity when the task does not require dynamic interaction or multi-turn feedback.

For Sequential Recommendation, the goal is to model user behavior over time and capture the evolving nature of preferences. Feature-based methods, with their emphasis on learning latent patterns from sequential data, remain effective in this area. However, Agentic frameworks show promise by integrating memory and planning capabilities to better handle long-term dependencies and adapt to changes in user behavior. Generative approaches might contribute through techniques such as chain-of-thought prompting for sequence modeling, but they usually face challenges with preserving the natural order of interactions. For Conversational Recommendation, which inherently relies on multi-turn dialogue and context-aware interactions, both Generative and Agentic paradigms excel. The natural language generation capability of Generative models enables them to produce personalized and contextually rich responses, while Agentic systems add value by dynamically managing interactions, incorporating feedback, and adapting recommendations over the course of a conversation. This task

benefits most from a combination of the two, where conversational fluency and interactive reasoning are paramount. Cross-domain recommendation tasks require the system to generalize across diverse data sources and domains. Here, the Feature-Based framework is particularly effective when it leverages multi-modal embeddings to capture semantic similarities between different domains. Generative methods can help in bridging data sparsity by generating contextual connections, although they might lack the consistency needed for reliable domain transfer. Agentic frameworks could further refine cross-domain recommendations by dynamically integrating external data, but their added complexity may not be necessary unless adaptive, real-time adjustments are required. Finally, for Item/Content Generation, where new content such as item descriptions, images, or creative outputs must be generated, the Generative framework is inherently best suited. Its strength lies in the ability to produce novel, high-quality content directly. Agentic approaches might complement this task through iterative planning and feedback incorporation, yet the core generation task is most effectively handled by methods that specialize in flexible, creative output.

### Comparative Discussion

- **Top-N Recommendation:** Best served by Feature-Based methods due to robust embedding extraction and efficient ranking; Generative and Agentic frameworks add complexity that may be unnecessary for static ranking tasks.
- **Sequential/Conversational Recommendation:** Sequential tasks can leverage both Feature-Based and Agentic approaches for capturing temporal dependencies, while conversational tasks benefit significantly from the natural language generation of Generative models, with Agentic systems enhancing interactive adaptability.
- **Cross-Domain and Item Generation:** Feature-Based approaches are advantageous for cross-domain tasks by providing semantic alignment across domains, whereas Item/Content Generation is most aligned with Generative models, possibly enhanced by Agentic feedback loops for real-time refinement.

## 7 FM4RecSys SHOWCASE: EMPIRICAL COMPARISON OF FM FOR SEQUENTIAL RECOMMENDATION

As shown in Table 4, we conduct a comprehensive empirical study to benchmark representative FM4RecSys frameworks across various sequential recommendation datasets. Overall, feature-based models maintain strong performance on standard benchmarks, benefiting from explicit feature engineering and effective modeling of historical interactions. For instance, models like HyperGraph-LLM and ReAT consistently achieve competitive results on the Beauty, Toys, and Sports datasets. Generative-based models further demonstrate superior performance, especially on complex datasets with sparse user behavior signals or rich semantic information (e.g., Yelp, MIND, GoodRead). Models such as POD

and GenRec leverage powerful language models (LMs) as behavior generators, achieving state-of-the-art NDCG@5 on multiple datasets, and highlighting their ability to capture complex user preference patterns. Since most of the agentic methods take a training-free strategy, indicating their potential for handling interactive and personalized recommendation scenarios.

This empirical comparison reveals several trends: (i) feature-based models remain strong baselines for structured recommendation tasks; (ii) generative models offer superior flexibility and generalization in diverse recommendation environments, and (iii) agentic models open up new research directions for interactive and autonomous recommender systems. These findings provide practical guidelines for the development of future FM4RecSys models and motivate further exploration of hybrid paradigms that unify the advantages of different FM families.

## 8 FM4RecSys: EXPLORED OPPORTUNITIES AND FINDINGS

As discussed above, the emergence of FMs has opened up unprecedented opportunities for advancing RSs, fundamentally reshaping how user preferences and behaviors are modeled, predicted, and interacted with. However, integrating these powerful models into RS scenarios also presents a host of new challenges that require innovative solutions across multiple dimensions. This section outlines key research directions and recent advancements aimed at enhancing RS capabilities in the FM era. We highlight developments in dynamic temporal extrapolation, multi-modal agent intelligence, retrieval-augmented generation (RAG), explainability, life-long personalization, and system-level efficiency and scalability. Together, these emerging areas offer a comprehensive roadmap for building RSs that are more adaptive, reliable, and capable of meeting evolving user needs in complex, real-world environments.

### 8.1 Explanations and Interpretation

A common task in enhancing the interpretability of recommendation systems is the generation of natural language explanations [243]. This involves directing the recommender or external model to produce, in a sentence or a paragraph, the reasons behind recommending a specific item to a particular user. For instance, given a user  $u$  and an item  $i$ , the model is tasked to generate a coherent and understandable explanation in natural language that elucidates why item  $i$  is recommended to user  $u$ . A series of works use ID-based representation and leverage prompts like “explain to the user  $u$  why item  $i$  is recommended” [11]. However, the use of IDs alone in prompts may lead to vague explanations, lacking clarity on specific aspects of the recommendation. To address this, Cui *et al.* [129] propose to integrate item features as hint words in the prompt, aiming to guide the model more effectively in its explanatory process. Meanwhile, Wang *et al.* [189] introduced LLM4Vis, a ChatGPT-based method for visualization recommendations that uses in-context learning to generate visualizations and human-like explanations, avoiding the need for a large dataset of examples like traditional methods. Ngoc *et al.* [190] combines

Reference	Framework	Backbone	NS	NDCG@5 by Dataset								
				Beauty	Toys	Sports	Games	CDs	Office	Yelp	MIND	GoodRead
SARSRec [137]	-	Transformer	0	0.0249	0.0306	0.0154	0.0365	-	-	0.0100	-	-
BertRec [230]	-	Bert	0	0.0124	0.0071	0.0075	0.0311	-	-	0.0033	-	-
P5 [93]	Feature-based	T5	0	0.0107	0.0050	0.0041	-	-	-	-	-	-
TIGER [16]	Feature-based/Generative	T5	0	0.0321	0.0371	0.0181	-	-	-	-	-	-
LMIndexer [231]	Feature-based	T5	0	0.0262	0.0268	0.0142	-	-	-	-	-	-
HyperGraph-LLM [232]	Feature-based	GPT-4	0	0.0376	0.0379	-	-	-	-	-	-	-
ReAT [233]	Feature-based	T5	99	0.0382	0.0390	0.0188	-	-	-	-	-	-
LC-Rec [124]	Feature-based/Generative	GPT-3.5	Random	-	-	-	0.0560	-	-	-	-	-
ONCE [234]	Feature-based	GPT-3.5		4	-	-	-	-	-	-	0.3872	0.7196
EmbSum [235]	Feature-based	T5	4	-	-	-	-	-	-	-	0.3675	0.5486
LLMHD [236]	Feature-based	GPT-3.5	-	-	-	-	-	-	-	0.0542	-	-
KDA [237]	Feature-based	GPT-3	99	-	-	-	-	-	0.3403	-	-	-
POD [238]	Generative	T5	99	0.0395	0.0599	0.0396	-	-	-	-	-	-
GenRec [239]	Generative	T5	0	0.0397	-	0.0332	-	-	-	0.0475	-	-
BIGRec [240]	Generative	Llama	-	-	-	-	0.0189	-	-	-	-	-
RDRec [241]	Generative	T5	99	0.0461	0.0593	0.0408	-	-	-	-	-	-
RecGPT [242]	Generative	GPT-1	0	0.0143	0.0355	0.0408	-	-	-	0.0107	-	-
RecMind [165]	Agentic	GPT-3.5	99	0.0289	-	-	-	-	-	0.0342	-	-
AgentCF [160]	Agentic	GPT-3.5	9	-	-	-	-	0.4373	0.3589	-	-	-

TABLE 4

Performance comparison of different FM4RecSys frameworks on sequential recommendation tasks over multiple datasets. N@5 refers to NDCG@5, and NS denotes the number of negative samples. Missing values are denoted as '-'.

embedding-based and semantic-based models to generate post-hoc explanations, leveraging ontology-based knowledge graphs to enhance interpretability and user trust. Then, Chu *et al.* [116] combined the reasoning capabilities of FMs with the structural advantages of HNNs (Hypergraph Neural Networks) to better capture and interpret individual user interests. Recently, Liu *et al.* [191] leverage continuous prompt vectors instead of discrete prompt templates. Remarkably, it is found that ChatGPT, operating under in-context learning without fine-tuning, outperforms several traditional supervised methods [192].

## 8.2 Fairness for FM4RecSys

The imperative of fairness in RS stems from its widespread use in decision-making and meeting user demands. Nonetheless, there currently exists a deficiency in comprehending the degree of fairness manifested by FMs in RSs, as well as in identifying suitable methodologies for impartially addressing the needs of diverse user and item groups within these models [185], [186]. For the user group side, Hua *et al.* [185] propose the Unbiased Foundation Model for Fairness-aware Recommendation (UP5) based on Counterfactually-Fair-Prompting (CFP) techniques. After that, Zhang *et al.* [186] crafted metrics and a dataset that accounts for different sensitive attributes in two recommendation scenarios: music and movies, and evaluate ChatGPT's fairness in RS concerning various sensitive attributes of the user side. Recently, Deldjoo *et al.* [187] proposed CFaiRLLM, a comprehensive evaluation framework designed to assess and mitigate biases in LLM-based RS by examining how recommendations vary with the inclusion of sensitive attributes such as gender and age, underscoring the impact of different user profile sampling strategies on fairness outcomes. However, they only focus on age and sex, suggesting the need for future research to consider a broader array of sensitive attributes and to validate the framework across diverse domains.

For the item side, Hou *et al.* [8] guide FMs with prompts to formalize the recommendation task as a conditional

ranking task to improve item-side fairness. Besides, Jiang *et al.* [188] investigate the impact of historical user interactions and inherent semantic biases in FMs and introduce a framework called IFairLRS, which adapts traditional fairness methods to enhance item-side fairness in FM4RecSys without compromising recommendation accuracy. Research on non-discrimination and fairness in FM4RecSys is still in its early stages, indicating a need for further investigation.

## 8.3 Multimodal Recommender Systems

Recent years have seen the rapid advancement of MFMs that can jointly process and understand multiple modalities within a unified architecture. These foundation models have opened new frontiers in recommendation by enabling both enhanced content understanding and generative capabilities. A key aspect of multimodal recommendation is *modality fusion*, combining signals from text (e.g., item descriptions), images (e.g., product photos), audio (e.g., music or speech), and video to enrich user and item representations. Models like M6-Rec [174] unify recommendation tasks (retrieval, ranking, explanation, content creation) by converting multimodal inputs into a text-based format and applying large Transformer models for reasoning and generation. Similarly, Flamingo [175] and Kosmos-2 [176] adopt cross-attention mechanisms that allow language models to attend to visual embeddings, enabling few-shot reasoning grounded in visual context. In domains such as fashion or home decor, incorporating image features has been shown to significantly improve item matching and ranking quality, especially in cold-start scenarios [177]. Frameworks such as MMRec [178] and MMREC [179] support the integration of pre-trained visual and textual encoders into recommendation pipelines, providing flexible architectures for early fusion, co-training, and hybrid models. In the video domain, models like Video-LLaMA [180] represent recent breakthroughs in integrating visual frames, audio tracks, and text transcriptions into large language models to enable content-based video recommendation and captioning. These advances highlight the trend of leveraging raw multimodal content, rather than relying

solely on sparse interaction data or metadata. Beyond understanding existing content, multimodal foundation models enable **personalized content generation**, effectively blurring the boundary between recommendation and creation. In e-commerce, systems like M6-Rec [174] and PMG [181] can generate personalized product images or fashion styles from scratch, conditioned on user preferences inferred from interaction histories. In the music domain, models such as TalkPlay [182] and LLark [183] generate textual explanations or music tags directly from audio input, enabling conversational recommendation interfaces and playlist generation with better contextual understanding. Such generative capabilities are supported by models like CM3leon [184], which integrate text-to-image and image-to-text generation within a decoder-only transformer, and by retrieval-augmented models that can dynamically fetch relevant content snippets to condition output generation. These architectures open up use cases such as personalized ads, playlist artwork, video summaries, or even AI-created recommendation items tailored to each user.

Further progress in this area will involve improving the efficiency and scalability of multi-modal foundation models for real-time recommendation, developing more robust evaluation protocols for generative recommendation quality, and enhancing personalization strategies to better align multimodal generation with individual user intent. As these systems continue to mature, they are expected to support more intelligent, adaptive, and engaging recommendation experiences that unify content understanding and creation.

#### 8.4 In-context Learning Capability

In-context learning, a hallmark of modern FMs, refers to the ability to perform tasks without parameter updates by conditioning on prompts that contain task descriptions and examples. Within RS, this capability manifests through zero-shot, few-shot, and reasoning-based approaches that allow LLMs to generate relevant item suggestions, often without explicit supervised training for the recommendation task. Recent work has shown that this paradigm opens new avenues for flexible, domain-agnostic, and user-friendly recommendations.

**Zero-shot and Few-shot Recommendation:** LLMs such as GPT-3, GPT-4, and LLaMA have demonstrated notable performance in zero-shot and few-shot recommendation scenarios. Researchers have formulated next-item prediction and top-N recommendation as language modeling tasks, prompting LLMs with user histories, candidate items, and instruction-based queries. Hou *et al.* [8] show that GPT-4, when prompted appropriately, can act as a zero-shot ranker, exhibiting competitive ranking accuracy on benchmark datasets. Similarly, Wang *et al.* [244] demonstrates that a structured 3-step prompt can enable GPT-3 to outperform some traditional sequential models on MovieLens. Few-shot strategies further enhance performance by conditioning the model on a handful of annotated recommendation examples. Benchmarks like LLMRec [170] reveal that while LLMs underperform on raw recommendation metrics compared to specialized models, they excel in explanation generation and preference summarization—showcasing their capacity to integrate semantics and simulate commonsense reasoning.

**Reasoning and Instruction-following:** Instruction-following LLMs can generate recommendations based on free-form user descriptions or complex goals. For instance, Zhang *et al.* [171] reformulate recommendation as an instruction-following task, where user profiles and intents are described in natural language. Their fine-tuned Flan-T5 model not only generates relevant recommendations but also adapts flexibly across domains. Similarly, LLM Reasoning Graphs [172] leverage LLMs to generate logical chains linking user interests to item features, which are then used to guide downstream recommenders. In conversational recommendation, in-context reasoning enables LLMs to process feedback iteratively. Spurlock *et al.* [173] show that ChatGPT can refine its recommendations through user feedback via reprompting, thus aligning better with user preferences. This showcases the dual strengths of LLMs in understanding nuanced natural language and adapting recommendations through multi-turn reasoning.

In-context learning enables LLMs to serve as adaptable, explainable, and domain-transferable recommenders. While their performance in traditional metrics can lag behind specialized models, their strengths in semantic reasoning, instruction adherence, and human-centric interaction make them powerful tools for next-generation recommendation systems. Ongoing research continues to explore how to better align, prompt, and integrate LLMs into broader recommendation architectures.

#### Explored Opportunities And Findings

- FMs have paved the way for more intuitive explanations in RSs. By integrating rich item features and leveraging advanced prompting techniques, these models now produce natural language explanations that not only clarify recommendation rationale but also boost user trust and system transparency.
- Emerging frameworks focused on fairness indicate promising directions for mitigating bias on both user and item sides. Although initial efforts have shown encouraging results in handling sensitive attributes, there remains a significant opportunity to broaden these techniques and ensure equitable performance across diverse application domains.
- The incorporation of multi-modal data and in-context learning has markedly enhanced the personalization capabilities of recommendation systems. This progress has enabled richer, more dynamic representations; however, it also highlights pressing challenges in scalability, evaluation, and computational efficiency.

## 9 FM4REC SYS: OPEN CHALLENGES AND OPPORTUNITIES

While significant progress has been made, the field of foundation model-powered recommender systems is still in its early days. Numerous challenges remain open, and they coincide with exciting opportunities for future research. In this section, we outline some of the key open issues and potential directions, grouped into three broad categories:

online deployment, enhancing recommender system capabilities, technical scalability and efficiency, and methodological improvements.

### 9.1 Online Deployment

Foundation models, particularly large language models (LLMs), have shown promising capabilities in enhancing recommendation systems through improved generalization, natural language understanding, and multi-task unification. However, deploying such models in real-time, large-scale industrial settings introduces substantial engineering and system design challenges. Several industry leaders have taken early steps toward integration. For instance, Alibaba's M6-Rec [174] has been deployed as a unified foundation model across retrieval, ranking, and content generation in their e-commerce platform. It employs parameter-efficient tuning (e.g., prefix tuning) and caching strategies to meet latency constraints. Amazon leverages LLMs for cold-start enhancement by generating synthetic metadata [228], while Spotify's LLark [183] explores music understanding via LLMs for personalized playlist generation and natural language tagging. Despite these advancements, real-world deployment remains constrained by several bottlenecks. First, latency and throughput pose a major hurdle, as LLMs are computationally intensive and often incompatible with the sub-100ms inference requirements of online systems [245], [246]. Techniques such as hybrid two-stage architectures (retrieval + reranking) [247], early exit transformers [248], and query caching [249] have been proposed to mitigate this. Second, memory and cost efficiency become critical at scale, with methods like LoRA, adapter tuning [177], and distillation [250] increasingly adopted to reduce the serving footprint. Third, personalization vs. generalization presents a modeling challenge, while LLMs offer strong general knowledge, encoding user-specific nuances without overfitting or memory overhead is non-trivial. Lightweight user-specific prompts or adapters offer partial solutions. Additionally, integrating FMs into existing infrastructure involves compatibility issues, as output formats and ranking scores from generative models may not align with classical CTR-based pipelines. Production deployment also requires robust evaluation protocols, since traditional metrics like NDCG may not capture the full effect of generative personalization, necessitating A/B testing and user satisfaction tracking. Lastly, concerns around explainability, trust, and model staleness remain—hallucinations from LLMs can mislead users, and large models are harder to update regularly compared to standard recommender embeddings. Despite these challenges, foundation models continue to gain traction in recommendation deployment, with hybrid, latency-aware architectures emerging as a practical compromise to unlock their potential at scale.

## 9.2 Enhancing Recommender System Capabilities

### 9.2.1 Temporal Extrapolation

Recent studies [251] have demonstrated that FMs can extrapolate time series data in a zero-shot manner, achieving performance on par with or superior to specialized models trained on specific tasks. This success is largely due to FM's

ability to capture multi-modal distributions and their inclination towards simplicity and repetition, which resonates with the repetitive and seasonal trends commonly observed in time series data. Time series modeling, distinct from other sequence modeling due to its variable scales, sampling rates, and occasional data gaps, has not fully benefited from large-scale pretraining. To address this, LLMTIME2 [252] leverages LLMs for continuous time series prediction by encoding time series as numerical strings and treating forecasting as a next-token prediction task. This method, which transforms token distributions into continuous densities, facilitates an easy application of LLMs to time series forecasting without the need for specialized knowledge or high computational costs, making it particularly beneficial for resource-constrained scenarios. Moreover, by viewing user preference data as a time series sequence, these models might adeptly adapt to long-term shifts in preferences and enhance personalization and predictive accuracy over time, especially with the zero-shot capability of approaches like LLMTIME2, which enables a rapid adaptation to user preference change without the need for extensive retraining.

### 9.2.2 Multi-modal Agent AI for RecSys

Multimodal Agent AI [77] is an emerging field that explores AI systems capable of perceiving and acting across various domains through a unified understanding of multimodal data. These systems leverage generative models and diverse data sources for reality-agnostic training and can operate in both physical and virtual environments. In RS field, such agents can infer user preferences, adapt to real-time feedback, and provide personalized recommendations. Notably, they have the potential to act as simulators—not only for systems but also for user behavior—enabling offline data collection and training that could reduce real-world A/B testing costs. While early works have demonstrated proof-of-concept applications in settings like route planning and medical recommendation, many capabilities remain aspirational. Key open challenges include grounding multi-modal perception in recommendation-specific tasks, ensuring alignment with real user preferences, and scaling interactive agents in practical deployment scenarios.

### 9.2.3 RAG meets RecSys

Retrieval-augmented generation (RAG) is a technique used in FMs to enhance their generative capability by integrating external data retrieval into the generative process [253]. This approach improves the accuracy, credibility, and relevance of FM outputs, particularly in knowledge-intensive tasks like information retrieval and RS. RAG aims to address outdated knowledge, the generation of incorrect information (hallucinations), and limited domain expertise by combining the FM's internal knowledge with dynamic external knowledge bases. RAG is suitable for enhancing the FM4RecSys, especially in modeling lifelong user behavior sequences in real-world RS environments [254]. It could potentially ensure that the RSs remain up-to-date with continuous shifts in user preferences and trends, which is critical for precise identification and documentation of long-term behavioral patterns. For instance, considering the input token length restriction of FMs, RAG may be utilized to selectively extract pertinent portions of a user's interaction history and

associated external knowledge, thereby conforming to the model's input constraint. Additionally, RAG may lessen the likelihood of producing irrelevant recommendations or non-existent items (hallucinations), thereby enhancing the reliability of FM4RecSys.

#### 9.2.4 Explainability and Trustworthiness

Enhancing explainability and trustworthiness in RS is always a significant challenge, especially in the FM era. The complexity and size of FMs introduce new hurdles in explaining FM4RecSys. There are two primary approaches to explainability in RS: one involves generating natural language explanations for recommendations, and the other dives into the model's internal workings. The former approach has seen considerable exploration [255] in the pre-FM era, whereas the latter is less developed. There are also some works [13], [136] that align FMs, such as their prompts, with explicit knowledge bases like knowledge graphs. This alignment can make the model's decision-making process traceable as specific paths in the knowledge graph, offering a clearer explanation. However, these approaches are still in their preliminary phase and might be further enhanced by techniques such as Chain/Tree of Thoughts.

#### 9.2.5 Personalization

LLMs with their advanced multi-modal understanding and generation capabilities, can process and analyze vast amounts of multi-modal data, capturing intricate user preferences and behaviors that traditional models might overlook. This deep understanding allows LLM-based recommender systems to provide highly tailored recommendations, enhancing the relevance and usefulness of the suggestions. By leveraging LLMs, recommender systems can adapt to individual user needs dynamically, considering context, past interactions, and subtle nuances in user input, which leads to more engaging and personalized user interactions. The ability of LLMs to generate human-like text also plays a critical role in creating more relatable and persuasive recommendations, ultimately driving user engagement.

Despite the remarkable advancements by LLMs, there are significant challenges in achieving effective personalization in recommender systems. Ensuring fairness and mitigating biases in recommendations remains a critical issue, as LLMs can inadvertently propagate and even amplify existing biases present in the training data [256], [257], leading to unfair or skewed recommendations. Moreover, there are serious concerns regarding data privacy and security, as LLMs may inadvertently leak personal information embedded in prompts or pre-trained data [258]–[260]. Another major challenge is the static nature of existing LLMs. Given the same input, LLMs typically produce similar outputs, whereas recommender systems need to be highly dynamic. In the same user session, LLM behavior needs to adapt drastically based on a few user interactions. Across sessions, LLM behavior should evolve over time to reflect changes in user preferences and interactions [261]–[263]. Addressing these challenges requires ongoing research and the development of innovative techniques to optimize LLMs for efficiency, fairness, robustness, and data privacy in personalization tasks.

### 9.3 Technical Challenges: Efficiency and Scalability

#### 9.3.1 Long-sequences in FM4RecSys

FM4RecSys faces challenges with long input sequences due to its fixed context window limitations, impacting its efficacy in tasks requiring extensive context [264], [265], like in sequential recommendations. Sequential RSs, relying on a user's comprehensive interaction history and extensive item ranking lists, often exceed the FMs' context capacity, leading to less effective recommendations. Adaptations from NLP techniques are being explored, including segmenting and summarizing inputs to fit within the context window and employing strategies like attention mechanisms and memory augmentation to enhance the focus on pertinent segments of the input. The RoPE technique [266], with its innovative rotary position embedding, shows promise in managing long inputs and offers a potential solution for maintaining the performance of an RS despite the context window constraint of FMs.

#### 9.3.2 System Performance Analysis: APIs cost, Training & Inference Efficiency

In the development of FM-based RSs, a critical aspect is the cost assessment, which varies depending on the data and model selections throughout the training and inference phases [5]. The training phase incurs costs due to the recommendation model's pretraining, fine-tuning, and algorithmic development, where the complexity and the need for specialized engineering can drive up expenses. During the recommendation inference stage, costs persist in the form of system upkeep, updates, and computational demands of API-driven service provision. For instance, systems like OpenAI's GPT-3/4 [133], [267] have costs associated with API usage and token interactions, escalating with more intricate or extensive usage. Furthermore, the incorporation of RAG tools can further elevate expenses by extending prompt lengths and, consequently, the number of tokens processed, leading to higher API fees. Additionally, customization through fine-tuning also adds to the overall expenses.

Addressing efficiency in FM4RecSys is a practical challenge with direct implications for system performance and resource utilization. Referencing Table 5, we outline targeted solutions:

**Training Cost Reduction:** For pre-training or fine-tuning Foundation Models within recommender systems, it is necessary to carefully select the most informative and diverse data so that the model can efficiently capture essential user-item interaction patterns and features and accelerate the learning process [144], [268], [269]. Meanwhile, model distillation [281] offers an alternative approach to reducing fine-tuning costs. Wang *et al.* [272] propose the Step-by-step Knowledge Distillation Framework for Recommendation (SLIM), which leverages the reasoning capabilities of FMs in a resource-efficient manner. By using Chain-of-Thought (CoT) prompting and rationale distillation, SLIM enables smaller models to perform effective and meaningful sequential recommendations at lower costs. In addition, employing techniques [198] like LoRA [141] and LoftQ [282] for fine-tuning can help in managing memory usage and reducing training time.

Cost and Efficiency	Possible Methods	References
<i>Training Cost</i>	Data Selection for Pre-training Data Selection for Fine-tuning Parameter-efficient Fine-tuning Model Distillation	[268], [269] [144], [270], [271] [198] [272]
<i>Inference Latency</i>	Embedding Caching Lightweight FMs	[207], [273] [274], [275], [276]
<i>API Cost</i>	Data Selection Prompt Selection Adaptive RAG	[277] [278] [279], [280]

TABLE 5  
An organization of representative methods for reducing the cost and improving the efficiency for FM4RecSys.

**Inference Latency Reduction:** The computational demand for FM inference is notable. Strategies such as employing pre-computed embedding caches [207], [273] (e.g., VQ-Rec or LLM4Seq) can offer some relief by speeding up inference. Similarly, efforts to compact the model size through distillation [275], pruning [274], and quantization [276] can lead to improvements in memory cost and inference speed. Recently, Kaur *et al.* [283] has proposed a hybrid task allocation strategy that leverages the strengths of both LLMs and traditional RSs to minimize inference latency. This strategy involves two criteria for categorizing users into two groups: strong users, who continue to receive recommendations from traditional RSs, and weak users, who are shifted to LLM recommendations if LLMs demonstrate superior performance.

**API Cost Reduction:** In FM-based API recommender systems, efficient data selection can enhance the fine-tuning efficiency by using a selected set of data points [277]. Additionally, refining prompt engineering with methods like prompt generation or compression [278] may lead to more efficient processing of FM inputs by making prompts more concise or better tailored, though the gains should be considered within realistic expectations. Besides, utilizing RAG to enhance API-based RSs can result in an additional context length, especially when retrieving longer item descriptions as prompt inputs. Therefore, adopting adaptive RAG [279] is also an effective method to reduce API costs in that case.

## 9.4 Methodological Improvements: Evaluation and Development

### 9.4.1 Benchmarks and Evaluation Metrics

Liu *et al.* [134] benchmark four state-of-the-art LLMs on five recommendation system tasks, using both quantitative and qualitative methods. However, they only focus on specific LLMs like ChatGPT and ChatGLM, limiting experiments to the Amazon Beauty dataset due to the high computational cost. Thus, because of the domain-specific nature of recommendation systems, there is a need for more datasets, recommendation tasks, and evaluation metrics to create a more unified benchmark. Moreover, for multi-modal and personalized agents FMs, devising new benchmarks and evaluation metrics specifically for recommendation scenarios is essential. Finally, data pollution presents a significant challenge. Ensuring that the evaluation data has not been

inadvertently used during FM training is difficult, making it a crucial issue to address in the current setting.

In summary, to comprehensively evaluate and enhance the performance of FM-based recommendation systems, a holistic and diversified benchmark is essential. Such a benchmark should include a variety of datasets, diverse recommendation tasks, and metrics that are adaptable to different models.

### 9.4.2 Causality in FM4RecSys

The rise of FMs has notably expanded the potential for exploring causality in recommender systems. Equipped with vast knowledge bases and sophisticated architectures, FMs offer unique opportunities for causal discovery and analysis, making them a major focus in current RS research efforts. Although FMs are talented at predicting user preferences, interpreting the causal relationships behind these preferences is increasingly crucial. The pursuit of causal inference in recommender systems seeks to provide recommendations that are not only transparent and reliable but also more interpretable. A key open question is how to effectively apply causal methods to address pressing concerns regarding bias and fairness within these systems. These areas are critically important and represent key directions for future research, demanding extensive exploration to uncover and address the underlying complexities of causality in recommendation systems.

### 9.4.3 Safety and Trust in FM4RecSys

We draw on a concise discussion of how the impressive understanding and generation power of FMs act as a dual-edged weapon in the context of FM4RecSys.

From a perspective of safety, FMs are vulnerable to red teaming attacks, where malicious actors craft poison prompts to manipulate the models into producing undesirable content. This content can range from fraudulent or racist material to misinformation or content inappropriate for younger audiences, potentially causing significant societal harm and putting users at risk [284], [285]. Thus, in the context of FM4RecSys especially when employing conversational interfaces, aligning FMs with human values becomes crucial. This alignment involves gathering relevant negative data and employing supervised fine-tuning techniques such as online and offline human preference training [286], [287]. These methods can help in refining the models to adhere more closely to human instructions

and expectations, ensuring the generative contents made by FM4RecSys are safe, reliable, and ethically sound.

*From a perspective of privacy*, if FMs are trained directly on a large amount of sensitive user interaction data, it might be possible for third parties to use methods like prompt injection to access specific user interaction histories, thereby constructing user profiles. In that sense, the incorporation of approaches such as federated learning [288] and machine unlearning [289], [290] into FM4RecSys represents a promising direction for the future.

## 10 CONCLUSION

In this paper, we have furnished a thorough review of FM4RecSys, providing detailed comparisons and highlighting future research paths. We proposed a classification scheme for organizing and clustering existing publications and discussed the advantages and disadvantages of using foundation models for recommendation tasks. In addition, we detailed some of the most pressing open problems and promising future extensions. We hope this survey provides an overview of the challenges, recent progress, open questions, and opportunities in foundation models to the recsys research community.

## REFERENCES

- [1] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in *Recommender Systems Handbook*. Springer, 2015, pp. 1–34.
- [2] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 5:1–5:38, 2019.
- [3] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, J. Gao *et al.*, "Multimodal foundation models: From specialists to general-purpose assistants," *Foundations and Trends® in Computer Graphics and Vision*, vol. 16, no. 1-2, pp. 1–214, 2024.
- [4] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, X. Li, C. Zhu, H. Guo, Y. Yu, R. Tang, and W. Zhang, "How can recommender systems benefit from large language models: A survey," *CoRR*, vol. abs/2306.05817, 2023.
- [5] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunsell, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajah, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, and et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.
- [6] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.
- [7] H. Ding, Y. Ma, A. Deoras, Y. Wang, and H. Wang, "Zero-shot recommender systems," *CoRR*, vol. abs/2105.08318, 2021.
- [8] Y. Hou and et al., "Large language models are zero-shot rankers for recommender systems," *arXiv preprint arXiv:2402.04521*, 2024.
- [9] C. Gao, W. Lei, X. He, M. de Rijke, and T. Chua, "Advances and challenges in conversational recommender systems: A survey," *AI Open*, vol. 2, pp. 100–126, 2021.
- [10] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M. Kan, and T. Chua, "Estimation-action-reflection: Towards deep interaction between conversational and recommender systems," in *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA*. ACM, 2020, pp. 304–312.
- [11] L. Li, Y. Zhang, and L. Chen, "Generate neural template explanations for recommendation," in *The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland*. ACM, 2020, pp. 755–764.
- [12] J. Sun, C. Zheng, and et al., "A survey of reasoning with foundation models," *CoRR*, vol. abs/2312.11562, 2023.
- [13] Y. Wang, Z. Chu, X. Ouyang, S. Wang, H. Hao, Y. Shen, J. Gu, S. Xue, J. Y. Zhang, Q. Cui, L. Li, J. Zhou, and S. Li, "Enhancing recommender systems with large language model reasoning graphs," *CoRR*, vol. abs/2308.10835, 2023.
- [14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.
- [16] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost *et al.*, "Recommender systems with generative retrieval," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10 299–10 315, 2023.
- [17] L. Xu, J. Zhang, B. Li, J. Wang, M. Cai, W. X. Zhao, and J. Wen, "Prompting large language models for recommender systems: A comprehensive framework and empirical analysis," *CoRR*, vol. abs/2401.04997, 2024.
- [18] C. Huang, J. Wu, Y. Xia, Z. Yu, R. Wang, T. Yu, R. Zhang, R. A. Rossi, B. Kveton, D. Zhou *et al.*, "Towards agentic recommender systems in the era of multimodal large language models," *arXiv preprint arXiv:2503.16734*, 2025.
- [19] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback, 2022," *arXiv preprint arXiv:2212.08073*, vol. 8, no. 3, 2022.
- [20] M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu *et al.*, "Human-level play in the game of diplomacy by combining language models with strategic reasoning," *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.
- [21] P. Liu, L. Zhang, and J. A. Gulla, "Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems," *CoRR*, vol. abs/2302.03735, 2023.
- [22] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, and E. Chen, "A survey on large language models for recommendation," *CoRR*, vol. abs/2305.19860, 2023.
- [23] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, and Q. Li, "Recommender systems in the era of large language models (llms)," *CoRR*, vol. abs/2307.02046, 2023.
- [24] L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," *CoRR*, vol. abs/2309.01157, 2023.
- [25] S. Zhang, Y. Tay, L. Yao, A. Sun, and C. Zhang, "Deep learning for recommender systems," in *Recommender Systems Handbook*. Springer US, 2022, pp. 173–210.
- [26] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. A. Orgun, "Sequential recommender systems: Challenges, progress and prospects," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 2019, pp. 6332–6338.
- [27] M. A. Islam, M. M. Mohammad, S. S. S. Das, and M. E. Ali, "A survey on deep learning based point-of-interest (POI) recommendations," *Neurocomputing*, vol. 472, pp. 306–325, 2022.
- [28] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *Int. J. Digit. Libr.*, vol. 17, no. 4, pp. 305–338, 2016.
- [29] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Comput. Commun.*, vol. 41, pp. 1–10, 2014.
- [30] S. Balineni and W. Andreopoulos, "Graph deep learning hashtag recommender for reels," in *IEEE Ninth International Conference on Big Data Computing Service and Applications, BigDataService 2023, Athens, Greece, July 17-20, 2023*. IEEE, 2023, pp. 119–126.

- [31] M. Karimi, D. Jannach, and M. Jugovac, "News recommender systems - survey and roads ahead," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 1203–1227, 2018.
- [32] E. Hasan, M. Rahman, C. Ding, J. X. Huang, and S. Raza, "Review-based recommender systems: A survey of approaches, challenges and future perspectives," *CoRR*, vol. abs/2405.05562, 2024.
- [33] R. He and J. J. McAuley, "VBPR: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. AAAI Press, 2016, pp. 144–150.
- [34] P. Knees and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 10, no. 1, pp. 2:1–2:21, 2013.
- [35] M. Ge and F. Persia, "A survey of multimedia recommender systems: Challenges and opportunities," *Int. J. Semantic Comput.*, vol. 11, no. 3, p. 411, 2017.
- [36] Y. Liu, H. Yin, B. Cui, Y. Chen, K. Wang, and Z. Huang, "Nova: Non-invasive and fusing variational autoencoder for recommendation with side information," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 404–414.
- [37] X. Wei, H. Zhang, L. Cao, L. Nie, Y. Yang, and T.-S. Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.
- [38] L. Chen, L. Cao, X. Wei, and Y. Yang, "Cmbf: Co-attentive multi-modal feature fusion for recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 549–558.
- [39] W. Chen, J. Zhang, H. Liu, X. Liu, J. Wu, Y. Yang, and J. Tang, "Mkgformer: Multi-modal knowledge graph enhanced transformer for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- [40] M. Jeong, K.-W. Kim, M. Song, and H. Park, "Camrec: Co-attentive multimodal recommendation with pre-trained encoders," in *Proceedings of the Web Conference 2024*, 2024.
- [41] C. Tao, R. Wang, Y. Wang, F. Wu, T. Tan, and L. Wang, "Mgat: Multi-modal graph attention network for recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 169–178.
- [42] J. Kim, K. Lee, S. Kim, H. Lim, B. Kim, and H. Kim, "Modality-aware recommendation with interaction-level fusion," in *Proceedings of the 45th International ACM SIGIR Conference*, 2022, pp. 149–158.
- [43] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022*, 2022.
- [44] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800gb dataset of diverse text for language modeling," *CoRR*, vol. abs/2101.00027, 2021.
- [45] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Lucioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekerman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Ajji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, and et al., "BLOOM: A 176b-parameter open-access multilingual language model," *CoRR*, vol. abs/2211.05100, 2022.
- [46] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: open pre-trained transformer language models," *CoRR*, vol. abs/2205.01068, 2022.
- [47] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," *CoRR*, vol. abs/2203.15556, 2022.
- [48] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023.
- [49] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu et al., "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.
- [50] Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, and Q. Tian, "Qa-lora: Quantization-aware low-rank adaptation of large language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- [51] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [52] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 12401–12430.
- [53] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.
- [54] R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schriftwieser, A. Glaese, J. Chen, E. Pitler, T. P. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, and et al., "Gemini: A family of highly capable multimodal models," *CoRR*, vol. abs/2312.11805, 2023.
- [55] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schriftwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakera, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, and et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *CoRR*, vol. abs/2403.05530, 2024.
- [56] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 19730–19742.
- [57] L. Xue, M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo, S. Kendre, J. Zhang, C. Qin, S. Zhang, C. Chen, N. Yu, J. Tan, T. M. Awalgaonkar, S. Heinecke, H. Wang, Y. Choi, L. Schmidt, Z. Chen, S. Savarese, J. C. Niebles, C. Xiong, and R. Xu, "Xgen-mm (BLIP-3): A family of open large multimodal models," *CoRR*, vol. abs/2408.08872, 2024.
- [58] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [59] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *CoRR*, vol. abs/2305.06355, 2023.
- [60] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Y. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *CoRR*, vol. abs/2308.01390, 2023.

- [61] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *CoRR*, vol. abs/2305.06355, 2023.
- [62] M. Maaz, H. A. Rasheed, S. Khan, and F. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 12 585–12 602.
- [63] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, ser. Lecture Notes in Computer Science, vol. 15104. Springer, 2024, pp. 323–340.
- [64] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *CoRR*, vol. abs/2311.07919, 2023.
- [65] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *CoRR*, vol. abs/2407.10759, 2024.
- [66] X. Pan, L. Dong, S. Huang, Z. Peng, W. Chen, and F. Wei, "Kosmos-g: Generating images in context with multimodal large language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- [67] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang, "Emu: Generative pretraining in multimodality," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [68] K. Zheng, X. He, and X. E. Wang, "Minigpt-5: Interleaved vision-and-language generation via generative vokens," *CoRR*, vol. abs/2310.02239, 2023.
- [69] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023, pp. 15 757–15 773.
- [70] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quiriy, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rosenberg, T. N. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirovic, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. H. Frank, "Audiopalms: A large language model that can speak and listen," *CoRR*, vol. abs/2306.12925, 2023.
- [71] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *CoRR*, vol. abs/2303.04671, 2023.
- [72] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [73] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Y. Zou, Z. Zhao, and S. Watanabe, "Audiogpt: Understanding and generating speech, music, sound, and talking head," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*. AAAI Press, 2024, pp. 23 802–23 804.
- [74] S. Wu, H. Fei, L. Qu, W. Ji, and T. Chua, "Next-gpt: Any-to-any multimodal LLM," *CoRR*, vol. abs/2309.05519, 2023.
- [75] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal, "Codi-2: In-context, interleaved, and interactive any-to-any generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 27 415–27 424.
- [76] X. Wang, B. Zhuang, and Q. Wu, "Modaverse: Efficiently transforming modalities with llms," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 26 596–26 606.
- [77] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi *et al.*, "Agent ai: Surveying the horizons of multimodal interaction," *arXiv preprint arXiv:2401.03568*, 2024.
- [78] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A survey on the memory mechanism of large language model based agents," *arXiv preprint arXiv:2404.13501*, 2024.
- [79] X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang, and E. Chen, "Understanding the planning of llm agents: A survey," *arXiv preprint arXiv:2402.02716*, 2024.
- [80] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.
- [81] N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning (2023)," *arXiv preprint cs.AI/2303.11366*, 2023.
- [82] J. Hilton, R. Nakano, S. Balaji, and J. Schulman, "Webgpt: Improving the factual accuracy of language models through web browsing," *OpenAI Blog, December*, vol. 16, 2021.
- [83] W. Yao, S. Heinecke, J. C. Niebles, Z. Liu, Y. Feng, L. Xue, R. Murthy, Z. Chen, J. Zhang, D. Arpit *et al.*, "Retroformer: Retrospective large language agents with policy gradient optimization," *arXiv preprint arXiv:2308.02151*, 2023.
- [84] S. Wu, S. Zhao, Q. Huang, K. Huang, M. Yasunaga, K. Cao, V. N. Ioannidis, K. Subbian, J. Leskovec, and J. Zou, "Avatar: Optimizing llm agents for tool-assisted knowledge retrieval," *arXiv preprint arXiv:2406.11200*, 2024.
- [85] G. Li, H. Hammoud, H. Itani, D. Khizbulin, and B. Ghanem, "Camel: Communicative agents for "mind" exploration of large language model society," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 991–52 008, 2023.
- [86] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou *et al.*, "Metagpt: Meta programming for multi-agent collaborative framework," *arXiv preprint arXiv:2308.00352*, 2023.
- [87] Z. Liu, W. Yao, J. Zhang, L. Yang, Z. Liu, J. Tan, P. K. Choubey, T. Lan, J. Wu, H. Wang *et al.*, "Agentlite: A lightweight library for building and advancing task-oriented llm agent system," *arXiv preprint arXiv:2402.15538*, 2024.
- [88] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better llm-based evaluators through multi-agent debate," *arXiv preprint arXiv:2308.07201*, 2023.
- [89] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, and M. Sun, "Communicative agents for software development," *arXiv preprint arXiv:2307.07924*, 2023.
- [90] K. Bao, J. Zhang, Y. Zhang, W. Wenjie, F. Feng, and X. He, "Large language models for recommendation: Progresses and future directions," in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Beijing, China*. ACM, 2023, pp. 306–309.
- [91] W. Hua, L. Li, S. Xu, L. Chen, and Y. Zhang, "Tutorial on large language models for recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 1281–1283.
- [92] Y. Zhu, L. Wu, Q. Guo, L. Hong, and J. Li, "Collaborative large language model for recommender systems," *CoRR*, vol. abs/2311.01343, 2023.
- [93] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 299–315.
- [94] W. Hua, S. Xu, Y. Ge, and Y. Zhang, "How to index item ids for recommendation foundation models," *arXiv preprint arXiv:2305.06569*, 2023.
- [95] R. Sarkar, N. Bodla, M. I. Vasileva, Y. Lin, A. Beniwal, A. Lu, and G. Medioni, "Outfittransformer: Learning outfit representations for fashion recommendation," in *IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA*. IEEE, 2023, pp. 3590–3598.
- [96] J. Li, M. Wang, J. Li, J. Fu, X. Shen, J. Shang, and J. J. McAuley, "Text is all you need: Learning language representations for sequential recommendation," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA*. ACM, 2023, pp. 1258–1267.
- [97] Z. Zhang and B. Wang, "Prompt learning for news recommendation," in *Proceedings of the 46th International ACM SIGIR Conference*

- on Research and Development in Information Retrieval, Taipei, Taiwan.* ACM, 2023, pp. 227–237.
- [98] S. Doddapaneni, K. Sayana, A. Jash, S. S. Sodhi, and D. Kuzmin, “User embedding model for personalized language prompting,” *CoRR*, vol. abs/2401.04858, 2024.
- [99] J. Zhou, Y. Dai, and T. Joachims, “Language-based user profiles for recommendation,” *CoRR*, vol. abs/2402.15623, 2024.
- [100] S. Gao, J. Fang, Q. Tu, Z. Yao, Z. Chen, P. Ren, and Z. Ren, “Generative news recommendation,” in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*. ACM, 2024, pp. 3444–3453.
- [101] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O’Banion, and J. Xie, “User-llm: Efficient LLM contextualization with user embeddings,” *CoRR*, vol. abs/2402.13598, 2024.
- [102] Y. Shen, L. Zhang, K. Xu, and X. Jin, “Autotransition: Learning to recommend video transition effects,” in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel*, ser. Lecture Notes in Computer Science, vol. 13698. Springer, 2022, pp. 285–300.
- [103] K. Youwang, J. Kim, and T. Oh, “Clip-actor: Text-driven recommendation and stylization for animating human meshes,” in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel*, ser. Lecture Notes in Computer Science, vol. 13663. Springer, 2022, pp. 173–191.
- [104] C. Huang, S. Wang, X. Wang, and L. Yao, “Dual contrastive transformer for hierarchical preference modeling in sequential recommendation,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, 2023, pp. 99–109.
- [105] ——, “Modeling temporal positive and negative excitation for sequential recommendation,” in *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*. ACM, 2023, pp. 1252–1263.
- [106] J. Zhai, X. Zheng, C. Wang, H. Li, and Y. Tian, “Knowledge prompt-tuning for sequential recommendation,” in *Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada*. ACM, 2023, pp. 6451–6461.
- [107] Y. Xi, W. Liu, J. Lin, J. Zhu, B. Chen, R. Tang, W. Zhang, R. Zhang, and Y. Yu, “Towards open-world recommendation with knowledge augmentation from large language models,” *CoRR*, vol. abs/2306.10933, 2023.
- [108] W. Luo, C. Song, L. Yi, and G. Cheng, “Kellmrec: Knowledge-enhanced large language models for recommendation,” *CoRR*, vol. abs/2403.06642, 2024.
- [109] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni, “Where to go next for recommender systems? ID- vs. modality-based recommender models revisited,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan*. ACM, 2023, pp. 2639–2649.
- [110] J. Liao, S. Li, Z. Yang, J. Wu, Y. Yuan, and X. Wang, “Llara: Aligning large language models with sequential recommenders,” *CoRR*, vol. abs/2312.02445, 2023.
- [111] X. Lin, W. Wang, Y. Li, F. Feng, S. Ng, and T. Chua, “A multi-facet paradigm to bridge large language model and recommendation,” *CoRR*, vol. abs/2310.06491, 2023.
- [112] X. Wang, X. He, M. Wang, F. Feng, and T. Chua, “Neural graph collaborative filtering,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France*. ACM, 2019, pp. 165–174.
- [113] N. Guo, H. Cheng, Q. Liang, L. Chen, and B. Han, “Integrating large language models with graphical session-based recommendation,” *CoRR*, vol. abs/2402.16539, 2024.
- [114] N. Choudhary, E. W. Huang, K. Subbian, and C. K. Reddy, “An interpretable ensemble of graph and language models for improving search relevance in e-commerce,” in *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore*. ACM, 2024, pp. 206–215.
- [115] Q. Zhao, H. Qian, Z. Liu, G. Zhang, and L. Gu, “Breaking the barrier: Utilizing large language models for industrial recommendation systems through an inferential knowledge graph,” *CoRR*, vol. abs/2402.13750, 2024.
- [116] Z. Chu, Y. Wang, Q. Cui, L. Li, W. Chen, S. Li, Z. Qin, and K. Ren, “Llm-guided multi-view hypergraph learning for human-centric explainable recommendation,” *CoRR*, vol. abs/2401.08217, 2024.
- [117] X. Ren, W. Wei, L. Xia, L. Su, S. Cheng, J. Wang, D. Yin, and C. Huang, “Representation learning with large language models for recommendation,” in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*. ACM, 2024, pp. 3464–3475.
- [118] A. Damianou, F. Fabbri, P. Gligoli, M. D. Nadai, A. Wang, E. Palumbo, and M. Lalmas, “Towards graph foundation models for personalization,” in *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore*. ACM, 2024, pp. 1798–1802.
- [119] L. Sheng, A. Zhang, Y. Zhang, Y. Chen, X. Wang, and T.-S. Chua, “Language models encode collaborative signals in recommendation,” *arXiv preprint arXiv:2407.05441*, 2024.
- [120] W. Wei, X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, J. Wang, D. Yin, and C. Huang, “Llmrec: Large language models with graph augmentation for recommendation,” in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 806–815.
- [121] Y. Zhang, K. Bao, M. Yan, W. Wang, F. Feng, and X. He, “Text-like encoding of collaborative information in large language models for recommendation,” *arXiv preprint arXiv:2406.03210*, 2024.
- [122] G. Hu, Z. Yang, Z. Cai, A. Zhang, and X. Wang, “Generate and instantiate what you prefer: Text-guided diffusion for sequential recommendation,” *arXiv preprint arXiv:2410.13428*, 2024.
- [123] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [124] B. Zheng, Y. Hou, H. Lu, Y. Chen, W. X. Zhao, M. Chen, and J.-R. Wen, “Adapting large language models by integrating collaborative semantics for recommendation,” in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 1435–1448.
- [125] Y. Wang, Z. Ren, W. Sun, J. Yang, Z. Liang, X. Chen, R. Xie, S. Yan, X. Zhang, P. Ren et al., “Content-based collaborative generation for recommender systems,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2420–2430.
- [126] Y. Wang, J. Xun, M. Hong, J. Zhu, T. Jin, W. Lin, H. Li, L. Li, Y. Xia, Z. Zhao et al., “Eager: Two-stream generative recommender with behavior-semantic collaboration,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3245–3254.
- [127] Y. Yang, Z. Ji, Z. Li, Y. Li, Z. Mo, Y. Ding, K. Chen, Z. Zhang, J. Li, S. Li et al., “Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations,” *arXiv preprint arXiv:2503.02453*, 2025.
- [128] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou, “Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment,” *arXiv preprint arXiv:2502.18965*, 2025.
- [129] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, “M6-rec: Generative pretrained language models are open-ended recommender systems,” *arXiv preprint arXiv:2205.08084*, 2022.
- [130] C. Wu, F. Wu, T. Qi, J. Lian, Y. Huang, and X. Xie, “Ptum: Pre-training user model from unlabeled user behaviors via self-supervision,” *arXiv preprint arXiv:2010.01494*, 2020.
- [131] H. Ngo and D. Q. Nguyen, “Recpt: Generative pre-training for text-based recommendation,” *arXiv preprint arXiv:2405.12715*, 2024.
- [132] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, “Your diffusion model is secretly a zero-shot classifier,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.
- [133] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [134] J. Liu, C. Liu, P. Zhou, Q. Ye, D. Chong, K. Zhou, Y. Xie, Y. Cao, S. Wang, C. You, and P. S. Yu, “Llmrec: Benchmarking large language models on recommendation task,” *CoRR*, vol. abs/2308.12241, 2023.
- [135] J. Yao, W. Xu, J. Lian, X. Wang, X. Yi, and X. Xie, “Knowledge plugins: Enhancing large language models for domain-specific recommendations,” *CoRR*, vol. abs/2311.10779, 2023.
- [136] B. Rahdari, H. Ding, Z. Fan, Y. Ma, Z. Chen, A. Deoras, and B. Kveton, “Logic-scaffolding: Personalized aspect-instructed rec-

- ommendation explanation generation using llms," *CoRR*, vol. abs/2312.14345, 2023.
- [137] W. Kang and J. J. McAuley, "Self-attentive sequential recommendation," in *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. IEEE Computer Society, 2018, pp. 197–206.
- [138] R. Li, W. Deng, Y. Cheng, Z. Yuan, J. Zhang, and F. Yuan, "Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights," *CoRR*, vol. abs/2305.11700, 2023.
- [139] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," *CoRR*, vol. abs/2305.07001, 2023.
- [140] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "Tallrec: An effective and efficient tuning framework to align large language model with recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 1007–1014.
- [141] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR*, 2022.
- [142] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," 2023.
- [143] K. Bao, J. Zhang, W. Wang, Y. Zhang, Z. Yang, Y. Luo, F. Feng, X. He, and Q. Tian, "A bi-step grounding paradigm for large language models in recommendation systems," *CoRR*, vol. abs/2308.08434, 2023.
- [144] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, Y. Wei, and T. Chua, "Data-efficient fine-tuning for llm-based recommendation," *CoRR*, vol. abs/2401.17197, 2024.
- [145] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [146] Y. Xu, W. Wang, F. Feng, Y. Ma, J. Zhang, and X. He, "Diffusion models for generative outfit recommendation," in *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 2024, pp. 1350–1359.
- [147] V. Shilova, L. D. Santos, F. Vasile, G. Racic, and U. Tanielian, "Adbooster: Personalized ad creative generation using stable diffusion outpainting," in *Workshop on Recommender Systems in Fashion and Retail*. Springer, 2023, pp. 73–93.
- [148] H. Yang, J. Yuan, S. Yang, L. Xu, S. Yuan, and Y. Zeng, "A new creative generation pipeline for click-through rate with stable diffusion model," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 180–189.
- [149] Á. T. Czapp, M. Jani, B. Domíán, and B. Hidasi, "Dynamic product image generation and recommendation at scale for personalized e-commerce," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 768–770.
- [150] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T.-S. Chua, "Diffusion recommender model," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 832–841.
- [151] Z. Yang, J. Wu, Z. Wang, X. Wang, Y. Yuan, and X. He, "Generate what you prefer: Reshaping sequential recommendation via guided diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 247–24 261, 2023.
- [152] Z. Li, A. Sun, and C. Li, "Diffurec: A diffusion model for sequential recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 3, pp. 1–28, 2023.
- [153] H. Huang, C. Huang, T. Yu, X. Chang, W. Hu, J. McAuley, and L. Yao, "Dual conditional diffusion models for sequential recommendation," *arXiv preprint arXiv:2410.21967*, 2024.
- [154] L. Weng, "Llm-powered autonomous agents," *lilian-weng.github.io*, Jun 2023. [Online]. Available: <https://lilianweng.github.io/posts/2023-06-23-agent/>
- [155] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, and D. Cai, "What to do next: Modeling user behaviors by time-lstm," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne*, 2017, pp. 3602–3608.
- [156] E. Ie, C. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier, "Recsim: A configurable simulation platform for recommender systems," *CoRR*, vol. abs/1909.04847, 2019.
- [157] L. Wang, J. Zhang, X. Chen, Y. Lin, R. Song, W. X. Zhao, and J.-R. Wen, "Recagent: A novel simulation paradigm for recommender systems," 2023.
- [158] A. Zhang, L. Sheng, Y. Chen, H. Li, Y. Deng, X. Wang, and T. Chua, "On generative agents in recommendation," *CoRR*, vol. abs/2310.10108, 2023.
- [159] W. Shi, X. He, Y. Zhang, C. Gao, X. Li, J. Zhang, Q. Wang, and F. Feng, "Enhancing long-term recommendation with bi-level learnable large language model planning," *CoRR*, vol. abs/2403.00843, 2024.
- [160] J. Zhang, Y. Hou, R. Xie, W. Sun, J. J. McAuley, W. X. Zhao, L. Lin, and J. Wen, "Agentcf: Collaborative learning with autonomous language agents for recommender systems," *CoRR*, vol. abs/2310.09233, 2023.
- [161] E. Zhang, X. Wang, P. Gong, Y. Lin, and J. Mao, "Usimagent: Large language models for simulating search users," *CoRR*, vol. abs/2403.09142, 2024.
- [162] R. Ren, P. Qiu, Y. Qu, J. Liu, W. X. Zhao, H. Wu, J. Wen, and H. Wang, "BASES: large-scale web search user simulation with large language model based agents," *CoRR*, vol. abs/2402.17505, 2024.
- [163] F. Huang, Z. Yang, J. Jiang, Y. Bei, Y. Zhang, and H. Chen, "Large language model interaction simulator for cold-start item recommendation," *CoRR*, vol. abs/2402.09176, 2024.
- [164] Y. Shu, H. Gu, P. Zhang, H. Zhang, T. Lu, D. Li, and N. Gu, "Rah! recsys-assistant-human: A human-central recommendation framework with large language models," *CoRR*, vol. abs/2308.09904, 2023.
- [165] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, and Y. Yang, "Recmind: Large language model powered agent for recommendation," *arXiv preprint arXiv:2308.14296*, 2023.
- [166] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie, "Recommender AI agent: Integrating large language models for interactive recommendations," *CoRR*, vol. abs/2308.16505, 2023.
- [167] Z. Wang, Y. Yu, W. Zheng, W. Ma, and M. Zhang, "Multi-agent collaboration framework for recommender systems," *CoRR*, vol. abs/2402.15235, 2024.
- [168] H. Cai, Y. Li, W. Wang, F. Zhu, X. Shen, W. Li, and T.-S. Chua, "Large language models empowered personalized web agents," *arXiv preprint arXiv:2410.17236*, 2024.
- [169] X. Shen, R. Zhang, X. Zhao, J. Zhu, and X. Xiao, "PMG : Personalized multimodal generation with large language models," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13–17, 2024*. ACM, 2024, pp. 3833–3843.
- [170] J. Liu and et al., "Llmrec: Benchmarking large language models on recommendation tasks," *arXiv preprint arXiv:2305.17105*, 2023.
- [171] J. Zhang and et al., "Recommendation as instruction following: A large language model empowered recommendation approach," *arXiv preprint arXiv:2305.17897*, 2023.
- [172] Y. Wang and et al., "Enhancing recommender systems with large language model reasoning graphs," *arXiv preprint arXiv:2310.13476*, 2023.
- [173] C. Spurlock and et al., "Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback," *arXiv preprint arXiv:2401.10001*, 2024.
- [174] L. Cui, T. Huang, L. Sun et al., "M6-rec: A multi-task, multimodal, multilingual and multidomain recommender system," in *Proceedings of KDD*, 2022.
- [175] J.-B. Alayrac, J. Donahue, and et al., "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [176] B. Peng, X. Lin, Y. Wang et al., "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023.
- [177] S. Geng, Z. Hu, and et al., "Vip5: Visual instruction prompting for recommendation," in *Proceedings of the 46th International ACM SIGIR Conference*, 2023.
- [178] T. Zhou and et al., "Mmrec: An open-source toolkit for multimodal recommendation," *arXiv preprint arXiv:2303.02977*, 2023.
- [179] C. Tian and et al., "Mmrec: Multimodal recommendation with unified multimodal language model," *arXiv preprint arXiv:2403.03412*, 2024.
- [180] Y. Zhu, C. Wang, and et al., "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

- [181] H. Shen, Y. Zhang, and et al., "Pmg: Personalized multimodal generation for recommendation," *arXiv preprint arXiv:2402.00302*, 2024.
- [182] L. He and et al., "Talkplay: Dialogue-aware multimodal language models for music recommendation," *arXiv preprint arXiv:2402.06752*, 2024.
- [183] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multimodal instruction-following language model for music," *arXiv preprint arXiv:2310.07160*, 2023.
- [184] R. Beaumont and et al., "Cm3leon: Open foundation models for multimodal understanding and generation," *arXiv preprint arXiv:2306.06535*, 2023.
- [185] W. Hua, Y. Ge, S. Xu, J. Ji, and Y. Zhang, "UP5: unbiased foundation model for fairness-aware recommendation," *CoRR*, vol. abs/2305.12090, 2023.
- [186] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 993–999.
- [187] Y. Deldjoo and T. D. Noia, "Cfairllm: Consumer fairness evaluation in large-language model recommender system," *CoRR*, vol. abs/2403.05668, 2024.
- [188] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, and X. He, "Item-side fairness of large language model-based recommendation system," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*. ACM, 2024, pp. 4717–4726.
- [189] L. Wang, S. Zhang, Y. Wang, E. Lim, and Y. Wang, "Llm4vis: Explainable visualization recommendation using chatgpt," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023, pp. 675–692.
- [190] N. L. Lê, M. Abel, and P. Gouspillou, "Combining embedding-based and semantic-based models for post-hoc explanations in recommender systems," in *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2023, Honolulu, Oahu, HI, USA, October 1-4, 2023*. IEEE, 2023, pp. 4619–4624.
- [191] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu, "A first look at llm-powered generative news recommendation," *arXiv preprint arXiv:2305.06566*, 2023.
- [192] J. Liu, C. Liu, R. Lv, K. Zhou, and Y. Zhang, "Is chatgpt a good recommender? a preliminary study," *arXiv preprint arXiv:2304.10149*, 2023.
- [193] X. Hu, S. Storks, R. L. Lewis, and J. Chai, "In-context analogical reasoning with pre-trained language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 2023, pp. 1953–1969.
- [194] X. Li, F. Yan, X. Zhao, Y. Wang, B. Chen, H. Guo, and R. Tang, "HAMUR: hyper adapter for multi-domain recommendation," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom*. ACM, 2023, pp. 1268–1277.
- [195] Z. Tang, Z. Huan, Z. Li, X. Zhang, J. Hu, C. Fu, J. Zhou, and C. Li, "One model for all: Large language models are domain-agnostic recommendation systems," *CoRR*, vol. abs/2310.14304, 2023.
- [196] Y. Gong, X. Ding, Y. Su, K. Shen, Z. Liu, and G. Zhang, "An unified search and recommendation foundation model for cold-start scenario," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom*. ACM, 2023, pp. 4595–4601.
- [197] Z. Fu, X. Li, C. Wu, Y. Wang, K. Dong, X. Zhao, M. Zhao, H. Guo, and R. Tang, "A unified framework for multi-domain CTR prediction via large language models," *CoRR*, vol. abs/2312.10743, 2023.
- [198] J. Fu, F. Yuan, Y. Song, Z. Yuan, M. Cheng, S. Cheng, J. Zhang, J. Wang, and Y. Pan, "Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights," *CoRR*, vol. abs/2305.15036, 2023.
- [199] L. Guo, Z. Lu, J. Yu, Q. V. H. Nguyen, and H. Yin, "Prompt-enhanced federated content representation learning for cross-domain recommendation," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*. ACM, 2024, pp. 3139–3149.
- [200] G. Zhang, "User-centric conversational recommendation: Adapting the need of user with large language models," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 1349–1354.
- [201] L. Wang, H. Hu, L. Sha, C. Xu, K. Wong, and D. Jiang, "Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph," *CoRR*, vol. abs/2110.07477, 2021.
- [202] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. J. McAuley, "Large language models as zero-shot conversational recommenders," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom*. ACM, 2023, pp. 720–730.
- [203] G. Lin and Y. Zhang, "Sparks of artificial general recommender (AGR): early experiments with chatgpt," *CoRR*, vol. abs/2305.04518, 2023.
- [204] K. D. Spurlock, C. Acun, E. Saka, and O. Nasraoui, "Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback," *CoRR*, vol. abs/2401.03605, 2024.
- [205] M. Ravaut, H. Zhang, L. Xu, A. Sun, and Y. Liu, "Parameter-efficient conversational recommender system as a language processing task," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*. Association for Computational Linguistics, 2024, pp. 152–165.
- [206] X. Wang, X. Tang, W. X. Zhao, J. Wang, and J.-R. Wen, "Rethinking the evaluation for conversational recommendation in the era of large language models," *arXiv preprint arXiv:2305.13112*, 2023.
- [207] J. Harte, W. Zorgdrager, P. Louridas, A. Katsifodimos, D. Jannach, and M. Fragkoulis, "Leveraging large language models for sequential recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 1096–1102.
- [208] Y. Wang, Z. Liu, J. Zhang, W. Yao, S. Heinecke, and P. S. Yu, "DRDT: dynamic reflection with divergent thinking for llm-based sequential recommendation," *CoRR*, vol. abs/2312.11336, 2023.
- [209] Y. Wu, R. Xie, Y. Zhu, F. Zhuang, X. Zhang, L. Lin, and Q. He, "Personalized prompts for sequential recommendation," *CoRR*, vol. abs/2205.09666, 2022.
- [210] S. Qiao, C. Gao, J. Wen, W. Zhou, Q. Luo, P. Chen, and Y. Li, "LLM4SBR: A lightweight and effective framework for integrating large language models in session-based recommendation," *CoRR*, vol. abs/2402.13840, 2024.
- [211] Y. Li, X. Zhai, M. Alzantot, K. Yu, I. Vulic, A. Korhonen, and M. Hammad, "Calrec: Contrastive alignment of generative llms for sequential recommendation," *CoRR*, vol. abs/2405.02429, 2024.
- [212] S. Xu, W. Hua, and Y. Zhang, "Openp5: Benchmarking foundation models for recommendation," *CoRR*, vol. abs/2306.11134, 2023.
- [213] X. Li, Y. Zhang, and E. C. Malthouse, "A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, and fake news," in *Proceedings of the International Workshop on News Recommendation and Analytics co-located with the 2023 ACM Conference on Recommender Systems, Singapore*, ser. CEUR Workshop Proceedings, vol. 3561. CEUR-WS.org, 2023.
- [214] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, Z. Sun, X. Zhang, and J. Xu, "Uncovering chatgpt's capabilities in recommender systems," in *Proceedings of the 17th ACM Conference on Recommender Systems, Singapore*. ACM, 2023, pp. 1126–1132.
- [215] Z. Liu, S. Mei, C. Xiong, X. Li, S. Yu, Z. Liu, Y. Gu, and G. Yu, "Text matching improves sequential recommendation by reducing popularity biases," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom*. ACM, 2023, pp. 1534–1544.
- [216] S. Luo, Y. Yao, B. He, Y. Huang, A. Zhou, X. Zhang, Y. Xiao, M. Zhan, and L. Song, "Integrating large language models into recommendation via mutual augmentation and adaptive aggregation," *CoRR*, vol. abs/2401.13870, 2024.
- [217] V. W. Anelli, A. Bellogin, T. Di Noia, D. Jannach, and C. Pomo, "Top-n recommendation algorithms: A quest for the state-of-the-art," in *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 2022, pp. 121–131.
- [218] A. Boz, W. Zorgdrager, Z. Kotti, J. Harte, P. Louridas, V. Karakoidas, D. Jannach, and M. Fragkoulis, "Improving sequential recommendations with llms," *ACM Transactions on Recommender Systems*, 2024.

- [219] D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A survey on conversational recommender systems," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 105:1–105:36, 2022.
- [220] Y. Sun and Y. Zhang, "Conversational recommender system," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA*. ACM, 2018, pp. 235–244.
- [221] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J. Wen, and J. Yu, "Improving conversational recommender systems via knowledge graph based semantic fusion," in *The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, CA, USA*. ACM, 2020, pp. 1006–1014.
- [222] J. M. Lichtenberg, A. Buchholz, and P. Schwöbel, "Large language models as recommender systems: A study of popularity bias," *arXiv preprint arXiv:2406.01285*, 2024.
- [223] F. Zhu, Y. Wang, C. Chen, J. Zhou, L. Li, and G. Liu, "Cross-domain recommendation: Challenges, progress, and prospects," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Virtual Event / Montreal, Canada*. ijcai.org, 2021, pp. 4721–4728.
- [224] K. Zhou, J. Zhang, and C. Li, "Bundlegen: Generative bundle recommendation via diffusion models," in *Proceedings of the Web Conference*, 2023.
- [225] Y. Wang, M. Zhang, and J. Liu, "Musicgen: Personalized music playlist generation with pre-trained language models," in *Proceedings of the 46th International ACM SIGIR Conference*, 2023.
- [226] Y. Yao, C. Xu, and J. Wang, "Generative news recommendation with user-interest-aware document synthesis," in *EMNLP Findings*, 2022.
- [227] X. Li, L. Wu, and F. Liu, "Prompt4newsrec: Prompt-based generative news recommendation with user summarization," *arXiv preprint arXiv:2305.14520*, 2023.
- [228] A. Acharya and T. Smith, "Leveraging large language models for generating item metadata in cold-start recommendation," in *RecSys*, 2023.
- [229] Y. Huang and Y. Zhang, "Coldilm: Generating user-item interactions via llms for cold-start recommendation," in *WSDM*, 2024.
- [230] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.
- [231] B. Jin, H. Zeng, G. Wang, X. Chen, T. Wei, R. Li, Z. Wang, Z. Li, Y. Li, H. Lu *et al.*, "Language models as semantic indexers," *arXiv preprint arXiv:2310.07815*, 2023.
- [232] Z. Chu, Y. Wang, Q. Cui, L. Li, W. Chen, Z. Qin, and K. Ren, "Llm-guided multi-view hypergraph learning for human-centric explainable recommendation," *arXiv preprint arXiv:2401.08217*, 2024.
- [233] Y. Cao, N. Mehta, X. Yi, R. Keshavan, L. Heldt, L. Hong, E. H. Chi, and M. Sathiamoorthy, "Aligning large language models with recommendation knowledge," *arXiv preprint arXiv:2404.00245*, 2024.
- [234] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu, "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*, 2024, pp. 452–461.
- [235] C. Zhang, Y. Sun, M. Wu, J. Chen, J. Lei, M. Abdul-Mageed, R. Jin, A. Liu, J. Zhu, S. Park *et al.*, "Embsum: Leveraging the summarization capabilities of large language models for content-based recommendations," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 1010–1015.
- [236] T. Song, W. Chao, and H. Liu, "Large language model enhanced hard sample identification for denoising recommendation," *arXiv preprint arXiv:2409.10343*, 2024.
- [237] S. Yang, W. Ma, P. Sun, Q. Ai, Y. Liu, M. Cai, and M. Zhang, "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*, 2024, pp. 335–344.
- [238] L. Li, Y. Zhang, and L. Chen, "Prompt distillation for efficient llm-based recommendation," in *CIKM*, 2023.
- [239] P. Cao and P. Liò, "Genre: Generative sequential recommendation with large language models," *arXiv preprint arXiv:2407.21191*, 2024.
- [240] K. Bao, J. Zhang, W. Wang, Y. Zhang, Z. Yang, Y. Luo, C. Chen, F. Feng, and Q. Tian, "A bi-step grounding paradigm for large language models in recommendation systems," *ACM Transactions on Recommender Systems*, 2023.
- [241] X. Wang, J. Cui, Y. Suzuki, and F. Fukumoto, "Rdrec: Rationale distillation for llm-based recommendation," *arXiv preprint arXiv:2405.10587*, 2024.
- [242] Y. Zhang, W. Yu, E. Zhang, X. Chen, L. Hu, P. Jiang, and K. Gai, "Recgpt: Generative personalized prompts for sequential recommendation via chatgpt training paradigm," *arXiv preprint arXiv:2404.08675*, 2024.
- [243] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *Found. Trends Inf. Retr.*, vol. 14, no. 1, pp. 1–101, 2020.
- [244] L. Wang and E.-P. Lim, "Zero-shot next-item recommendation using large pretrained language models," *arXiv preprint arXiv:2306.06078*, 2023.
- [245] A. Agrawal, N. Kedia, A. Panwar, J. Mohan, N. Kwatra, B. Gulavani, A. Tumanov, and R. Ramjee, "Taming {Throughput-Latency} tradeoff in {LLM} inference with {Sarathi-Serve}, " in *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024, pp. 117–134.
- [246] H. Ghosh, "Enabling efficient serverless inference serving for llm (large language model) in the cloud," *arXiv preprint arXiv:2411.15664*, 2024.
- [247] J. Lu, K. Hall, J. Ma, and J. Ni, "Hyrr: Hybrid infused reranking for passage retrieval," *arXiv preprint arXiv:2212.10528*, 2022.
- [248] W. Shan, L. Meng, T. Zheng, Y. Luo, B. Li, T. Xiao, J. Zhu *et al.*, "Early exit is a natural capability in transformer-based models: An empirical study on early exit without joint optimization," *arXiv preprint arXiv:2412.01455*, 2024.
- [249] Y. An, Y. Cheng, S. J. Park, and J. Jiang, "Hyperrag: Enhancing quality-efficiency tradeoffs in retrieval-augmented generation with reranker kv-cache reuse," *arXiv preprint arXiv:2504.02921*, 2025.
- [250] M. Li *et al.*, "Recommendation models meet language models: A survey," *arXiv preprint arXiv:2402.00072*, 2024.
- [251] M. Jin, Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, X. Li, S. Pan, V. S. Tseng, Y. Zheng, L. Chen, and H. Xiong, "Large models for time series and spatio-temporal data: A survey and outlook," *CoRR*, vol. abs/2310.10196, 2023.
- [252] N. Gruver, M. A. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [253] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *CoRR*, vol. abs/2312.10997, 2023.
- [254] J. Lin, R. Shan, C. Zhu, K. Du, B. Chen, S. Quan, R. Tang, Y. Yu, and W. Zhang, "Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation," *CoRR*, vol. abs/2308.11131, 2023.
- [255] Y. Zhang, X. Chen *et al.*, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.
- [256] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *CoRR*, vol. abs/2309.00770, 2023.
- [257] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, T. Yu, H. Deilamsalehy, R. Zhang, S. Kim, and F. Dernoncourt, "Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes," *CoRR*, vol. abs/2402.01981, 2024.
- [258] Y. Yao, J. Duan, K. Xu, Y. Cai, E. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *CoRR*, vol. abs/2312.02003, 2023.
- [259] Z. Wan, A. Cheng, Y. Wang, and L. Wang, "Information leakage from embedding in large language models," *CoRR*, vol. abs/2405.11916, 2024.
- [260] C. Song and A. Raghunathan, "Information leakage in embedding models," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9–13, 2020*. ACM, 2020, pp. 377–390.
- [261] S. Wozniak, B. Koptyra, A. Janz, P. Kazienko, and J. Kocon, "Personalized large language models," *CoRR*, vol. abs/2402.09269, 2024.
- [262] K. Zhang, L. Qing, Y. Kang, and X. Liu, "Personalized LLM response generation with parameterized memory injection," *CoRR*, vol. abs/2404.03565, 2024.

- [263] C. Wang, M. Zhang, W. Ma, Y. Liu, and S. Ma, "Make it a chorus: Knowledge- and time-aware item modeling for sequential recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020.* ACM, 2020, pp. 109–118.
- [264] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *International Conference on Learning Representations*, 2019.
- [265] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [266] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.
- [267] OpenAI, "Gpt-4 technical report," *OpenAI*, 2023.
- [268] M. R. Glass, A. Gliozzo, R. Chakravarti, A. Ferritto, L. Pan, G. P. S. Bhargav, D. Garg, and A. Sil, "Span selection pre-training for question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online*. Association for Computational Linguistics, 2020, pp. 2773–2782.
- [269] S. M. Xie, S. Santurkar, T. Ma, and P. Liang, "Data selection for language models via importance resampling," *CoRR*, vol. abs/2302.03169, 2023.
- [270] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "LIMA: less is more for alignment," *CoRR*, vol. abs/2305.11206, 2023.
- [271] Y. Cao, Y. Kang, and L. Sun, "Instruction mining: High-quality instruction data selection for large language models," *CoRR*, vol. abs/2307.06290, 2023.
- [272] Y. Wang, C. Tian, B. Hu, Y. Yu, Z. Liu, Z. Zhang, J. Zhou, L. Pang, and X. Wang, "Can small language models be good reasoners for sequential recommendation?" in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024.* ACM, 2024, pp. 3876–3887.
- [273] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," in *Proceedings of the ACM Web Conference*, 2023, pp. 1162–1171.
- [274] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, "The lottery ticket hypothesis for pre-trained BERT networks," in *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, virtual*, 2020.
- [275] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, ser. Findings of ACL. Association for Computational Linguistics, 2020, pp. 4163–4174.
- [276] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han, "AWQ: activation-aware weight quantization for LLM compression and acceleration," *CoRR*, vol. abs/2306.00978, 2023.
- [277] H. Chen, Y. Zhang, Q. Zhang, H. Yang, X. Hu, X. Ma, Y. Yang-gong, and J. Zhao, "Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning," *CoRR*, vol. abs/2305.09246, 2023.
- [278] J. Mu, X. L. Li, and N. D. Goodman, "Learning to compress prompts with gist tokens," *CoRR*, vol. abs/2304.08467, 2023.
- [279] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada. Association for Computational Linguistics, 2023, pp. 9802–9822.
- [280] C. Huang, R. Wang, K. Xie, T. Yu, and L. Yao, "Learn when (not) to trust language models: A privacy-centric adaptive model-aware approach," *arXiv preprint arXiv:2404.03514*, 2024.
- [281] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.
- [282] Y. Li, Y. Yu, C. Liang, P. He, N. Karampatziakis, W. Chen, and T. Zhao, "Loftq: Lora-fine-tuning-aware quantization for large language models," *CoRR*, vol. abs/2310.08659, 2023.
- [283] K. Kaur and C. Shah, "Efficient and responsible adaptation of large language models for robust top-k recommendations," *arXiv preprint arXiv:2405.00824*, 2024.
- [284] B. Deng, W. Wang, F. Feng, Y. Deng, Q. Wang, and X. He, "Attack prompt generation for red teaming and defending large language models," in *Findings of the Association for Computational Linguistics, Singapore*. Association for Computational Linguistics, 2023, pp. 2176–2189.
- [285] J. Zhang, Y. Liu, Q. Liu, S. Wu, G. Guo, and L. Wang, "Stealthy attack on large language model based recommendation," *CoRR*, vol. abs/2402.14836, 2024.
- [286] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, "Aligning large language models with human: A survey," *CoRR*, vol. abs/2307.12966, 2023.
- [287] G. Xu, J. Liu, M. Yan, H. Xu, J. Si, Z. Zhou, P. Yi, X. Gao, J. Sang, R. Zhang, J. Zhang, C. Peng, F. Huang, and J. Zhou, "Cvalues: Measuring the values of chinese large language models from safety to responsibility," *CoRR*, vol. abs/2307.09705, 2023.
- [288] Y. Yu, Q. Liu, L. Wu, R. Yu, S. L. Yu, and Z. Zhang, "Untargeted attack against federated recommendation systems via poisonous item embeddings and the defense," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, Washington, DC, USA.* AAAI Press, 2023, pp. 4854–4863.
- [289] C. Chen, F. Sun, M. Zhang, and B. Ding, "Recommendation unlearning," in *The ACM Web Conference, Virtual Event, Lyon, France*. ACM, 2022, pp. 2768–2777.
- [290] H. Wang, J. Lin, B. Chen, Y. Yang, R. Tang, W. Zhang, and Y. Yu, "Towards efficient and effective unlearning of large language models for recommendation," *CoRR*, vol. abs/2403.03536, 2024.