

Large Language Model Enhanced Recommender Systems: Methods, Applications and Trends

Qidong Liu
Xi'an Jiaotong University &
City University of Hong Kong
Xi'an, China
liuqidong@stu.xjtu.edu.cn

Yejing Wang
City University of Hong Kong
Hong Kong, China
yejing.wang@my.cityu.edu.hk

Xiang Li
Nanyang Technological University
Singapore, Singapore
xiang002@e.ntu.edu.sg

Chong Chen
Tsinghua University
Beijing, China
cstchenc@163.com

Xiangyu Zhao ✉
City University of Hong Kong
Hong Kong, China
xianzhao@cityu.edu.hk

Zijian Zhang
Jilin University
Changchun, China
zhangzijian@jlu.edu.cn

Maolin Wang
City University of Hong Kong
Hong Kong, China
Morin.wang@my.cityu.edu.hk

Wei Huang
Independent Researcher
Beijing, China
hwdzyx@gmail.com

Yuhao Wang
City University of Hong Kong
Hong Kong, China
yhwang25-c@my.cityu.edu.hk

Yuqi Sun
Xi'an Jiaotong University
Xi'an, China
YuqiSun@stu.xjtu.edu.cn

Pengyue Jia
City University of Hong Kong
Hong Kong, China
jia.pengyue@my.cityu.edu.hk

Feng Tian ✉
Xi'an Jiaotong University
Xi'an, China
fengtian@mail.xjtu.edu.cn

Abstract

Due to exceptional reasoning and understanding abilities, the Large Language Model (LLM) has revolutionized the pattern of many fields, including recommender systems (RS). There has been a handful of research that focuses on empowering the RS by LLM. Recently, considering the latency and memory costs in real-world applications, **LLM-Enhanced RS (LLMERS)** is highlighted. This direction pushes the LLM into the online system with a large step by eliminating the utilization of LLM during inference. As a cutting-edge field, there is a clear need for a comprehensive survey to summarize this direction. In this survey, we systematically investigate the most up-to-date works of LLM-enhanced RS to boost this direction. Based on the component of an RS model that the LLM aims to augment, the basic taxonomy includes **Knowledge Enhancement**, **Interaction Enhancement** and **Model Enhancement**. Additionally, we identify several promising research directions. To facilitate access to the surveyed papers, we release a repository ¹.

¹<https://github.com/Applied-Machine-Learning-Lab/Awesome-LLM-Enhanced-Recommender-Systems>

✉ Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1454-2/2025/08
<https://doi.org/10.1145/3711896.3736553>

CCS Concepts

• Information systems → Recommender systems.

Keywords

Recommender Systems; Large Language Model; Enhancement

ACM Reference Format:

Qidong Liu, Xiangyu Zhao ✉, Yuhao Wang, Yejing Wang, Zijian Zhang, Yuqi Sun, Xiang Li, Maolin Wang, Pengyue Jia, Chong Chen, Wei Huang, and Feng Tian ✉. 2025. Large Language Model Enhanced Recommender Systems: Methods, Applications and Trends. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3711896.3736553>

1 Introduction

The large language model has been proven to own emergent capabilities in language understanding and reasoning [3, 15, 16, 108, 110, 113, 115]. In view of the only utilization of collaborative signals by conventional recommender systems [2, 88, 90], supplementing semantic information for RS by the LLM is attractive. Thus, many works have been proposed to fill the gap between natural language and recommendation, leading to a powerful RS.

Despite a certain extent of success in adapting LLM to RS, one significant difference between the dialogue system and the recommender system lies in the **inference latency**. RS often requires low latency for a mass of requests, while the LLM, *e.g.*, LLaMA-7B, can only achieve seconds latency for a response. However, many studies focus on using the LLM for recommendation directly at the early stage [18], which makes it difficult to meet the demands of real-world applications. Recently, more researchers have paid

attention to such an issue and commenced diving into the LLM-enhanced RS for practice. Therefore, to boost such a direction, we write this survey to conclude the most up-to-date works.

In order to clarify the range of this survey, we first give out the definition of LLMERS: *The conventional recommender systems are enhanced by LLM via assistance in training or supplementary for data, while no need for LLM inference during the service.* Though there have been some surveys about LLM for RS, three critical differences exist. **i)** Most current surveys focus on how to use the LLM itself as a better RS, including generative recommendation [35, 38, 93] and discriminative recommendation [4, 6, 26, 41, 76, 115]. By comparison, our survey addresses the LLMERS especially. **ii)** LLM for RS is a cutting-edge direction, which develops rapidly. Some surveys [4, 41, 93, 115] do not include the most up-to-date papers. By comparison, more than 60 works in this survey were published after 2024. **iii)** A few surveys have referred to LLM-enhanced RS [4, 41], but they only focus on the enhancement from the aspect of feature engineering. In contrast, this survey is the first one to conclude LLMERS from a comprehensive view, including feature and model aspects.

1.1 Preliminary

Since LLM-enhanced RS targets to enhance RS, introducing RS’s components and challenges is a prerequisite to understanding where and why the LLM is needed. As shown in Figure 1, the conventional RS often consists of the interaction data (feature and interaction) and the recommendation model.

Interaction Data. The conventional recommender system captures the collaborative signals from user-item records [33], which depends on the interactions in data for training. Besides, many content-based models [56] extract the co-occurrence relationships contained in the features of users and items for recommendation. Therefore, **Feature** and **Interaction** are two necessities of data. However, there exist two challenges of data hindering further advancements for conventional RS.

- **Challenge 1:** For features, they are often converted into numerical or categorical values for utilization while lacking reasoning and understanding from the knowledge aspect.
- **Challenge 2:** For interactions, data sparsity leads to insufficient training for RS models.

Recommendation Model. With the wide application of deep learning techniques, the recommendation model conforms to an “Embedding-Deep Network” pattern. The embedding layer targets transforming the raw features into dense representations [114], and the deep network will then capture the user’s interests [107]. However, they both face a unique challenge:

- **Challenge 3:** For recommendation models, they can only capture the collaborative signals, while ignoring semantic information.

1.2 Taxonomy

LLMERS augments the conventional RS from its basic components, *i.e.*, interaction data and recommendation model, so that only conventional RS models are conducted during serving. According to the challenge LLM will take effect, we categorize the LLM-enhanced RS into three lines, which are illustrated in Figure 1.

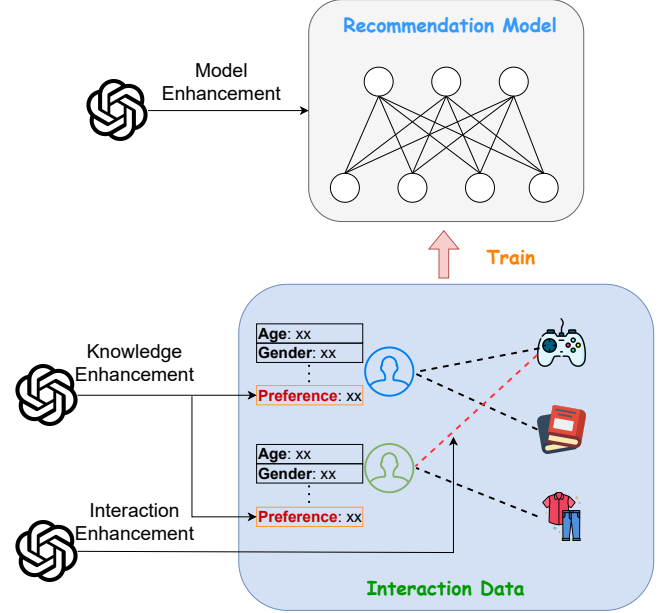


Figure 1: The taxonomy of LLM-enhanced RS.

Knowledge Enhancement. This line of work utilizes reasoning abilities and world knowledge of the LLM to derive textual descriptions for users or items. These descriptions will serve as extra features to supplement reasoning and understanding knowledge, and thus they can face the **Challenge 1**. (Section 2)

Interaction Enhancement. To face the data sparsity issue, *i.e.*, **Challenge 2**, some research studies adopt the LLM to derive new user-item interactions. (Section 3)

Model Enhancement. The LLM can analyze the interactions from a semantic view, so some works have tried to make use of the LLM to assist the conventional recommendation models, addressing the **Challenge 3**. (Section 4)

For clarity, we show all the LLMERS papers according to the taxonomy in Figure 2.

2 Knowledge Enhancement

LLM owns extensive world knowledge and powerful reasoning abilities, which can supplement the RS with external knowledge. In this category, the LLMERS often utilize explicit knowledge to augment the recommender system. For example, as shown in Figure 1, we can prompt the LLM by the user’s textual interactions to output the “preference” of the user as an extra feature. Such a feature, containing the knowledge from the LLM, can enhance the current RS. Since the feature can be derived and saved in advance, the demand for LLM will be avoided during the service, leading to efficient inference. According to the type of the feature, we can further partition this line into unstructured **Summary Text**, structured **Knowledge graph**, and **Combination** of these two. The illustrations of these lines are shown in Figure 3.

2.1 Summary Text

This subcategory refers to utilizing LLM to summarize the characteristics of items or to reason for the preference of users. For

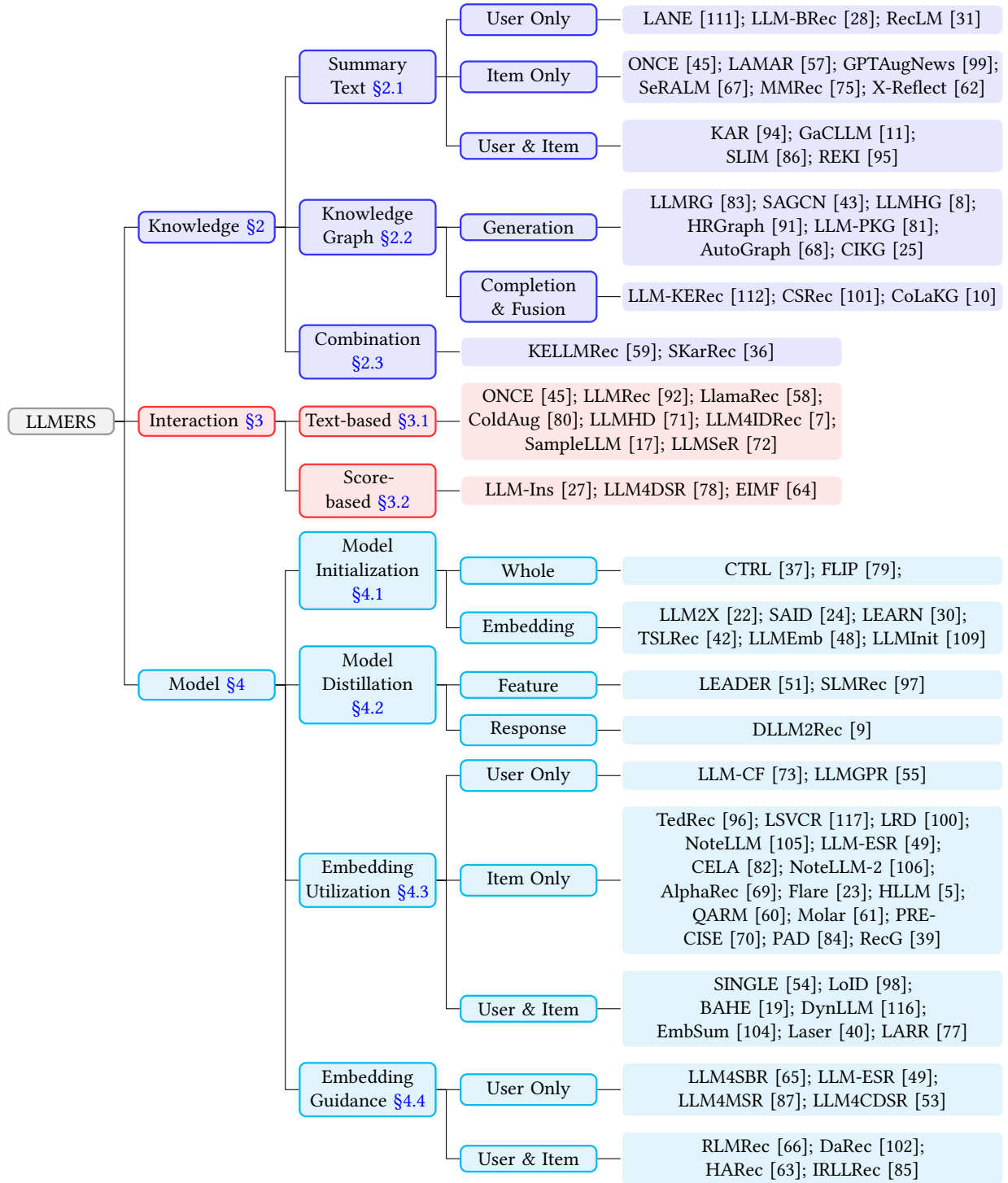


Figure 2: Typology of LLM-Enhanced Recommender Systems.

example, the prompt for summarization can be “Given a user who has viewed *<Browsing History>*, please explain what he or she is interested in: ”. The summary texts are then often encoded by a general language model, such as Bert [32]. Some works enhance the RS for either the user or item side, while others address both.

2.1.1 User Only and Item Only. For the user-only literature, LANE [111] and LLM-BRec [28] fabricate prompts to capture multi-preferences of users. For better utilization, LANE additionally designs an attention mechanism to align the preference representations of RS and LLM. On the other hand, as basic units in the RS,

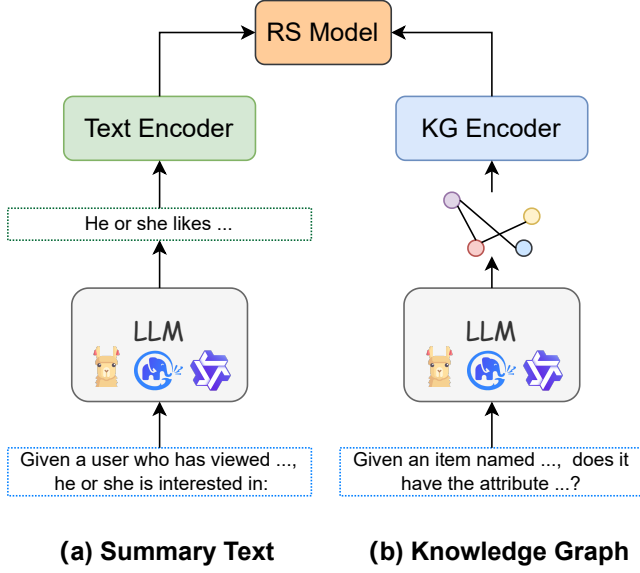


Figure 3: The illustration of Knowledge Enhancement.

more works focus on the item-only aspect. As one of the pioneers in this field, ONCE [45] utilizes LLM as a content summarizer for news recommendation. Besides, it proposes to extend the news content by LLM before encoding, while Yada *et al.* [99] uses the LLM to generate category descriptions for augmenting the news content. Similarly, LAMAR [57] adopts LLM to output factual properties of movies. As for SeRALM [67], it additionally devises an alignment training pattern for better integration of LLM’s knowledge. Recently, multimodal RS [46] has attracted much attention. MMRec [75] and X-Reflect [62] propose to use multimodal LLM to summarize the visual feature into textual description for use.

2.1.2 User & Item. Some studies generate summary texts for both user and item to enhance RS, especially content-based RS [29]. As one of the pioneers in this line, KAR [94] gets the reasoning knowledge of users and factual knowledge of items in textual form. The knowledge is then encoded by the designed multi-expert text encoders and integrated into the traditional CTR models, such as DeepFM [21] and DIN [118]. REKI [95], a successor of KAR, further refines the knowledge extraction module for various scales of recommendation scenarios. Different from these two works, GaCLLM [11] enhances the graph-based RS models by the knowledge of users and items, considering the LLM as an aggregator to extract structural information and thereby augment the RS.

2.2 Knowledge Graph

Summary text is a type of unstructured knowledge for enhancement, while the structured Knowledge Graph (KG) may drive better integration. Thus, many current works have explored how to apply the LLM to generate a KG (**Generation**) or augment an existing KG (**Completion & Fusion**) for enhancing traditional RS.

2.2.1 Generation. In this subcategory, the KG is generated initially and then used as supplementary features for enhancing RS.

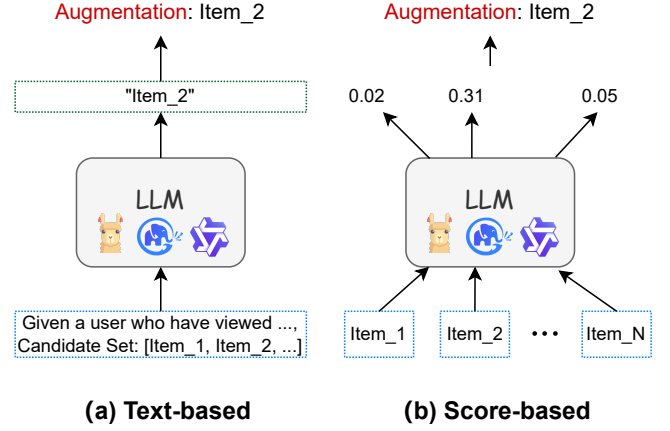


Figure 4: The illustration of Interaction Enhancement.

LLMRG [83] is the first to explore the KG generation for RS. It fabricates the reasoning prompt to derive possible interaction sequences from LLM and the verification prompt to filter out illogical sequences. The generated sequence forms an interaction graph, which is then used to enhance the user representation. Similarly, LLMHG [8] also prompts LLM to generate a KG based on the user’s interaction records, but instead of an interaction graph, it derives user-interest hyper-graphs. Besides, SAGCN [43] imposes the LLM to distinguish the user-item interactions from various semantic aspects, generating multiple interaction graphs corresponding to these aspects. All of these works tend to generate KG at the interaction level. By comparison, HRGraph [91] utilizes LLM to conduct job entity extraction from job descriptions and curriculum vitae, producing the specified graph for job recommendation.

2.2.2 Completion & Fusion. Different from generation, this line of work often imposes the LLM on existing KG. For example, LLM-KERec [112] adopts the LLM to identify complementary relationships within an item KG. Then, the KG-based RS can derive better recommendations by enriched entity relationships. Besides, Yang *et al.* [101] find that the direct fusion of existing KG and LLM-generated entity relations is sub-optimal. To address such an issue, they propose an improved fusion method to better integrate the two sources.

2.3 Combination

In consideration of the effectiveness of these two types, some works also resort to combining both of them. KELLMRec [59] is one representative, which integrates the KG for constructing prompts to avoid the hallucination problem when getting the summary texts from LLM. Besides, SKarRec [36] also links the entity in KG to the textual prompts, allowing the LLM to generate KG-aware knowledge that augments the concept recommendations.

3 Interaction Enhancement

To address the sparsity problem in traditional RS, data augmentation is a straightforward way [34, 52], generating new interactions. The augmented interactions are only supplemented for training traditional RS models, so no extra inference burden will be brought.

Inspired by the semantic view of LLM, some research studies have explored the interaction generation by LLM. We categorize these approaches into two types based on the output of the LLM, *i.e.*, **Text-based** and **Score-based**, shown in Figure 4. The former gives out the names of pseudo-interacted items as the augmentation. By comparison, score-based methods derive the logits of the possible interactions and generate the augmented items by ranking.

3.1 Text-based

ONCE [45] is the first to investigate the data augmentation for RS using LLM. It prompts the LLM to recommend news based on a user’s historical records. The generated news is added to the dataset as an augmented interaction for this user. However, the direct choice from a large item set is difficult for LLM. Thus, later works propose narrowing the selection to a smaller set of items for higher augmentation accuracy. For example, LLMRec [92] first adopts a well-trained traditional RS to filter out a list of items and then prompts LLM to choose from them. Similarly, LlamaRec [58] imposes the LLM to generate interactions from a random-sampled item set. Furthermore, Wang *et al.* [80] proposes to utilize the LLM to judge which item a user prefers, generating pair-wise interactions for cold-start RS. In addition to generating new data, distinguishing noisy data is also a vital topic in data augmentation. LLMHD [71] addresses this issue by applying the LLM to distinguish hard or noisy samples in sequential recommendation.

3.2 Score-based

As mentioned in the last subsection, it is difficult for LLM to select from a large number of items. Besides, the textual output often faces the out-of-corpus problem [1], *i.e.*, the generated item name cannot correspond to anyone in the item set. To avoid these two problems, some works explore the score-based method. One typical work of this line is LLM-InS [27]. It firstly imposes the LLM to derive the semantic embeddings of users and items, and then generates augmentations by calculating the similarity between them. By comparison, LLMDSR [78] considers the probability of word tokens as the confidence in deciding a noisy item and generates the new item by the similarity of embeddings.

4 Model Enhancement

Beyond data-level enhancement, the powerful abilities and semantics of LLM can also be injected into traditional recommendation models directly. RS models are generally composed of two main parts, *i.e.*, embedding layer and hidden layers. The former often captures relationships between items, while the latter extracts more complex user preferences. According to the various parts, we categorize this line into **Model Initialization**, **Model Distillation**, **Embedding Utilization** and **Embedding Guidance**, as shown in Figure 5. The first two categories involve applying the LLM to the entire RS model, while the embedding utilization and guidance only impose the LLM on the embedding layer. It is worth noting that Model Initialization, Model Distillation, and Embedding Guidance only need the LLM to assist in training RS models, so LLM usage is eliminated during serving. Besides, Embedding Utilization can cache the LLM embeddings before inference. Thus, they all can meet the latency requirements of real-time service.

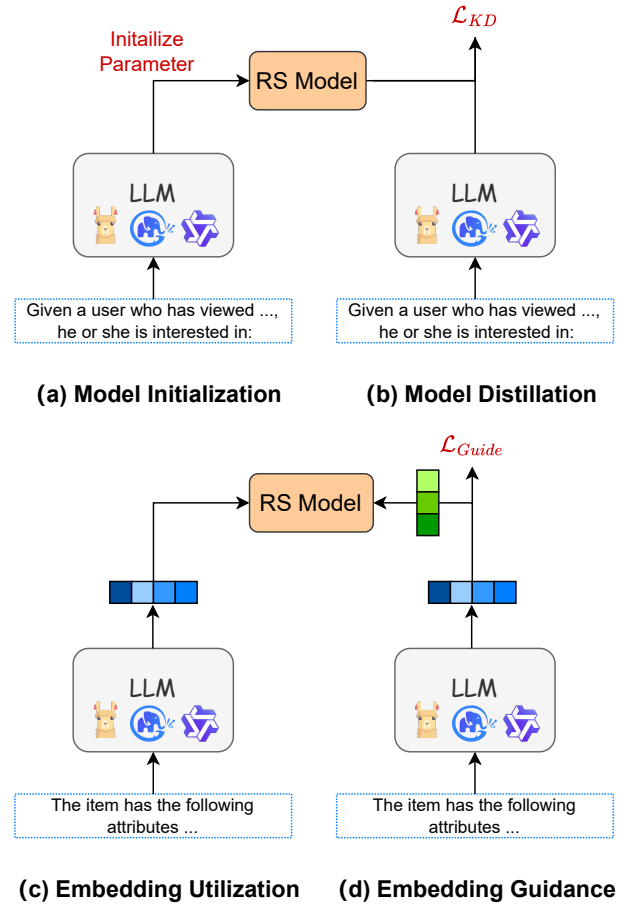


Figure 5: The illustration of Model Enhancement.

4.1 Model Initialization

Pretrain, as one of the initialization techniques, has shown potential in the recommendation field [44, 47, 89]. On the one hand, pretrained weights can help model convergence. On the other hand, the information learned from the pretrain stage will be preserved and affect the downstream recommendation tasks. Thus, recent works have explored saving the semantics of LLM to the weights of RS models before training. Some approaches apply pretrained weights to the whole RS model, denoted as **Whole**, while the others are for the embedding layer, referred to as **Embedding**.

4.1.1 Whole. CTRL [37] and FLIP [79] are the representatives in this line. CTRL first considers the semantic representation extracted from LLM and collaborative representation derived from the traditional RS model as two distinct modalities [12–14]. It then aligns such two “modalities” by contrastive tasks. The parameters updated during the alignment process serve as the initialization of the whole traditional RS model. Building upon CTRL, FLIP [79] fabricates fine-grained masked tabular and language modeling alignment tasks for better model initialization.

4.1.2 Embedding. Embedding plays a key role in the traditional RS models [114], preliminarily capturing the relationships between

items and features. The LLMERS in this sub-category all aim to extract semantic relationships between items by the derived embeddings from LLM (*i.e.*, LLM embeddings). Harte *et al.* [22] initiates this line, who input the item title to a large text embedding model. Then, the derived LLM embeddings are used to initialize the embedding layer of a sequential recommendation model. Later, LEARN [30] proposes to adopt the hidden states of a frozen open-sourced LLM as the item embedding. However, the general-purpose LLM without fine-tuning may encounter incompatible challenges for the recommendation task. To address this, SAID [24] and TSLRec [42] specialize the item attribute word generation and information reconstruction tasks, respectively, to better adapt the LLM to recommendation. More recently, LLMEmb [48] integrates an attribute-level contrastive task to further extract fine-grained semantic relationships between items.

4.2 Model Distillation

Knowledge distillation (KD) is one of the compression techniques [20], which is also promising to bring the powerful abilities of LLM to small RS models. However, the gap between discriminative RS models and generative LLM may hinder the success of KD. Thus, existing studies design **Feature-based** and **Response-based** distillation ways to address such an issue.

4.2.1 Feature-based. Feature-based distillation transfers knowledge directly between the hidden states of models, regardless of the output type. LEADER [51] is the pioneer of this line, which distills the RS model using the hidden state from the final layer of the LLM with a trainable adapter. Notably, LEADER also fine-tunes the LLM to transform the LLM into a discriminative model. In contrast, SLMRec [97] maintains the basic structure of LLM, *i.e.*, stack of several transformer layers, but reduces the number of layers to create a smaller-scale RS model. It then imposes knowledge distillation on several transformer layers between the LLM and the RS model.

4.2.2 Response-based. Following the advancement in KD for recommendation, recent works attempt to use LLM to generate a ranking list from a semantic view, followed by the ranking distillation loss [74] for the RS model. Specifically, DLLM2Rec [9] proposes to fine-tune the LLM to adapt the recommendation task and then distill the small RS model by the ranking list derived from the LLM.

4.3 Embedding Utilization

In general, the LLM generates natural language to complete various tasks. However, the textual outputs are often difficult and inefficient to be integrated into RS models directly. To avoid the deficiency and better utilize the semantics of LLM, many recent works utilize the embeddings derived from LLM, *e.g.*, the hidden states of the last transformer layer of the LLM, to enhance traditional RS as a semantic supplement. From the aspect that LLM embeddings address in the traditional RS, this line of work can be categorized into **User Only**, **Item Only** and **User & Item**.

4.3.1 User Only. In this cluster, the LLM is used to generate the representation of the user's preference directly. The LLM embeddings will then be fed into the traditional RS model for augmentation. A notable example in this line is LLM-CF [73]. It first designs a

data mixture method to fine-tune the LLM for recommendation capability. Next, the well-trained LLM can generate Chain-of-Thought (CoT) reasoning to enhance user preferences.

4.3.2 Item Only. In some types of recommender systems, *e.g.*, sequential recommendation, the item embeddings play a fundamental role. Thus, most existing embedding utilization focuses on the item-only enhancement. Many early studies in this sub-category directly adopt the frozen LLM without fine-tuning. For example, TedRec [96] applies LLaMA to encode the item texts and then proposes a frequency-based fusion to combine the derived semantic embeddings and ID embeddings. LRD [100] adopts LLM embeddings akin to TedRec but introduces a novel self-supervised relation discovery task to optimize the utilization. However, these works may lose the informative semantics of LLM embeddings during training. To address such an issue, LLM-ESR [49] and AlphaRec [69] both devise a new pattern of frozen LLM embeddings with a trainable adapter to maintain the original semantic information.

Despite successful utilization, the general-purpose LLM is often not well-suited for recommendation tasks. Thus, several works propose to fine-tune the LLM [50] before utilization. LEARN [30] proposes to align the user preferences derived from LLM with those of RS models, which refines the item LLM embeddings. NoteLLM [105] and NoteLLM-2 [106] are both devised for item-item recommendation, where the LLM derives the item embeddings for retrieval. NoteLLM only targets combining the textual semantics and collaborative signals, while NoteLLM-2 gives out a novel path to fine-tune the LLM for understanding the multimodal features of items. Recently, HLLM [5] proposes a novel hierarchical framework for enhancement. The first level involves an item LLM, which encodes the textual features of the item. The item LLM embeddings will then be fed into a user LLM for final recommendation.

4.3.3 User & Item. In this line, the current papers can be clustered into two categories. One is to utilize the user and item LLM embeddings for recommendation by a similarity function directly. For example, LoID [98] first aligns the LLM embeddings with collaborative ID information, then concatenates the user and item representations derived from the LLM for final recommendation. However, the user LLM embeddings are often sub-optimal because of the lengthy textual prompts of the user's interaction histories. To face this challenge, BAHE [19] proposes to partition an LLM into lower layers and higher layers. It uses the lower layers of the LLM to encode the item attributes to get the comic item representation. Then, the corresponding comic vectors of the user's interacted items are fed into the higher layer. By comparison, EmbSum [104] partitions the user's interaction records into several sessions and then uses the LLM to encode them separately. Except for the lengthy prompt issue, the general user LLM embedding also faces a lack of dynamics. To enhance dynamic graph recommendation, DynLLM [116] generates multi-faceted profiles for users.

The other line of work supplements the user and item LLM embeddings as extra features to augment the content-based RS models. For instance, Laser [40] utilizes the derived user and item LLM embeddings with an adapter for feature enhancement. LARR [77] adopts the LLM embeddings as extra features but further highlights the importance of fine-tuning the LLM specialized for RS.

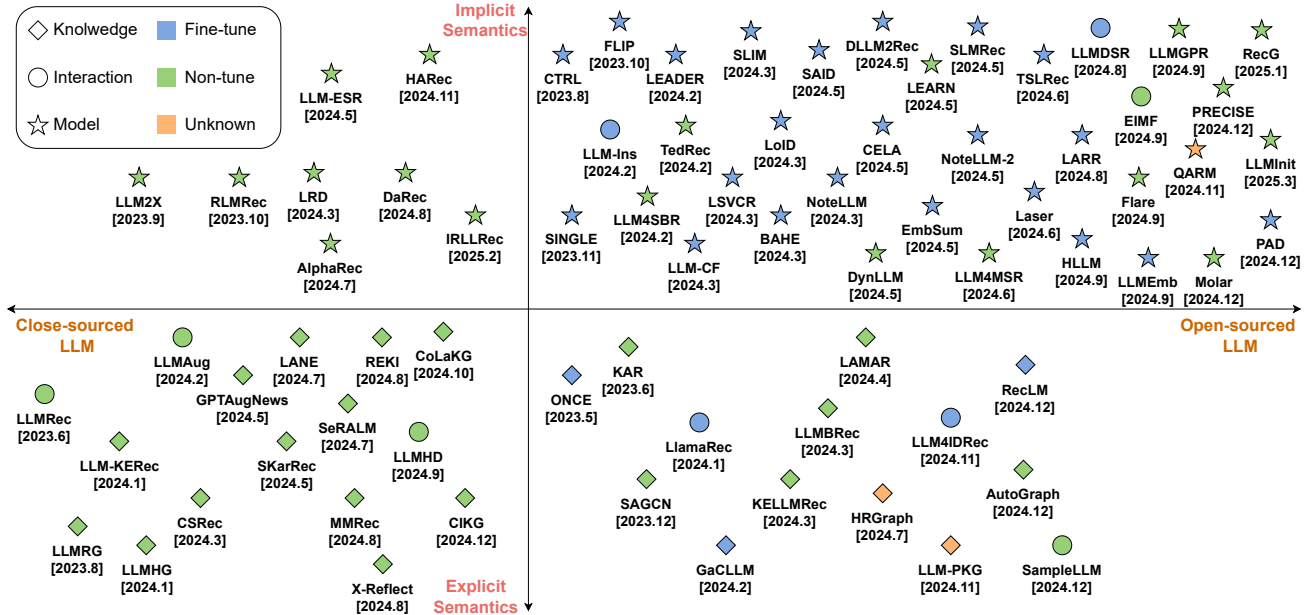


Figure 6: Trends of LLM-enhanced recommender systems. The horizontal and vertical axes represent open-sourced LLM or close-sourced LLM and explicit or implicit semantics, respectively. The dots in different colors and shapes are the categories of the works and whether the LLM is fine-tuned. In each quadrant, the works are ordered by time from left to right.

4.4 Embedding Guidance

Different from the direct embedding utilization, this line of papers only uses the LLM embeddings as the guidance for training or parameter synthesis. According to which aspect the LLM embeddings aim to enhance, we further categorize the current works into **User Only** and **User & Item**.

Due to the lengthy prompt of user interactions, user LLM embeddings are often noisy. Thus, considering the LLM embeddings as guidance is a feasible way for **user-only** enhancement. LLM4SBR [65] enhances the short-term and long-term preferences for session-based recommendation models by identifying the representative items via LLM embeddings. By comparison, LLM-ESR [49] retrieves similar users by LLM embeddings to enhance the long-tail users during the training. These two works both aim to search for relevant users or items from a semantic view and enhance the training process. Differently, LLM4MSR [87] generates the user-specified parameters for multi-scenario recommendation models by LLM embeddings accompanied by meta networks. As for **User & Item** line, the research studies guide both the user and item representation learning. RLMRec [66] kicks off this line, which adds an extra loss to align the collaborative user and item embeddings with corresponding LLM embeddings during training. Based on RLMRec’s insights, DaRec [102] proposes to disentangle the shared and specific parts for both LLM and collaborative embeddings. Only imposing the alignment on shared parts will lead to better performance.

5 Trends

To investigate trends of the LLMERS, we visualize the current works based on the type of LLM and semantics in Figure 6. As mentioned before, the key to LLM enhancement lies in the semantics. At the early stage, many works [94, 112] prompt the LLM to derive natural

language, which contains knowledge and semantics for augmentation. Since they are readable and understandable, we denote them as explicit semantics. By comparison, the cluster of implicit semantics means that the mediation from LLM for enhancement is non-verbal, e.g., the hidden states of LLM. Though humans may favor explicit knowledge due to its explainability, we find that recent works prefer the implicit one. The reason lies in the better performance of implicit knowledge, while explicit knowledge suffers from information loss during the extra encoding process.

As for the type of LLM, the figure reveals that more LLMERS works resort to the open-sourced LLM. One reason is that it can save money spent on calling the API. More importantly, the open-sourced LLM can be fine-tuned to suit the recommendation task [48]. Furthermore, observing the quadrants I and IV, the fine-tuned LLM has become prevalent recently, which indicates the significance of aligning the LLM with recommendation tasks. Overall, we can sum up the trends of current LLMERS as follows:

- The current LLMERS studies move from explicit semantics to **implicit semantics**.
- For a better adaptation to the recommendation task, **fine-tuned open-sourced LLM** becomes more prevalent.
- **Model enhancement** has gained more attention recently, because it can combine implicit semantics and fine-tuned LLM.

6 Applications and Efficiency

Applications. Traditional recommender systems have been widely adopted in various applications, such as news recommendations. Though LLMERS is also based on traditional RS, it often requires abundant side information, especially informative texts, to make full use of the reasoning and understanding abilities of LLM. Thus, current LLMERS works often focus on applications that have one of

Table 1: The applications and corresponding datasets of LLMERS.

Characteristic	Applications	Dataset	Link
Features	E-commerce	Amazon	https://bit.ly/4ibCNY1
		Alibaba	https://bit.ly/4itdGjq
		Online Retail	-
	Movie	MovieLens	https://bit.ly/3B6pV4J
		Netflix	https://bit.ly/3ZrfWYt
	POI	Yelp	https://bit.ly/4gfgvfd
		Delivery Hero	https://bit.ly/3Zug150
		Foursquare	https://bit.ly/4io706d
	Video	KuaiSAR	https://bit.ly/3Vv7TzY
	News	MIND	https://bit.ly/3ZzU9Xq
Texts	Book	GoodReads	https://bit.ly/4f3Vjio
		BookCrossing	-
		WeChat-ArticleRec	-
	Job	Personalized	-

two following characteristics: (i) it has a number of normal **features** that LLM can understand; (ii) it has a volume of **texts** that LLM can summarize. The E-commerce application belongs to the former one, where users have many profile features and commodities have attributes [49, 83]. Since the LLM owns open-world knowledge, these features can be understood by it and derive semantics for enhancement. By comparison, the news recommendation is the typical application that utilizes the LLM to summarize the abundant news texts [45]. To show it more clearly, we list the typical applications that LLMERS has targeted in Table 1.

Efficiency. Compared with LLM as RS, the most important distinction and merit of LLMERS lies in high efficiency. Thus, we discuss how each category of LLMERS is efficient in this paragraph. i) **Knowledge Enhancement** minimizes runtime LLM dependency by pre-generating static intermediaries (text summaries or knowledge graphs), but introduces persistent encoding overhead. Summary-based methods append cached summaries as supplementary features via text encoders [37], while KG follows dual paths: feature-based KG integrates with content-based systems through embeddings [27], and interaction-based KG enhances collaborative filtering via graph processing [83]. Hybrid methods combine KG-refined summaries with text encoding, trading compounded costs for semantic enrichment. ii) **Interaction Enhancement** exclusively amplifies training data through LLM-synthesized user interactions [101], preserving inference efficiency and achieving universal scalability across recommendation architectures. iii) **Model Enhancement** employs three strategies with varying architectural dependencies: initialization of components (the whole model [37] or embedding layers [24]) is enhanced by LLM-generated knowledge, and only small RS models are used for service. Distillation prioritizes traditional RS models [9], eliminating LLM during inference. Embedding utilization and guidance decouple LLM from inference by caching embeddings as semantic features (content-based RS [54]) or cross-modal alignment anchors (collaborative filtering [30]), incurring only storage and view-encoding costs.

7 Future Directions

LLMERS is a cutting-edge direction, where many problems have not been addressed. Therefore, we give out several directions to inspire later research.

- **Exploration for More Recommendation Tasks.** Existing LLMERS works have focused on revolutionizing several basic recommendation tasks, such as collaborative filtering [73] and sequential recommendation [49]. They have proven the effectiveness of bringing the semantics of LLM to traditional RS. Thus, it is also promising to replicate their success in some other recommendation tasks.
- **Multimodal RS.** The multimodal RS [46] has become prevalent recently due to the emergence of more multimedia services. However, the existing multimodal RS often faces the challenge of the extraction and fusion of features in different modalities. The adoption of multimodal LLM [103] may be a feasible way to augment existing multimodal RS and even eliminate the extraction and fusion procedures.
- **User-side Enhancement.** There have been many efforts of LLMERS imposed on the item side, but few on the user side. The reason lies in that the user’s historical interactions are often organized into texts orderly. However, such a prompt often faces the overlength problem [19]. Besides, the textual interactions are often difficult to understand for LLM. This problem thus hinders the user-side enhancement severely and needs to be addressed.
- **Explainability.** Explainability is an important aspect of constructing a trustworthy RS. The traditional RS is often short in this aspect since it only adopts meaningless identities. The LLM enhances the RS by understanding users and items from a semantic view, which is promising for deriving an explanation.
- **Benchmark.** Since LLMERS is a newborn direction, there is no benchmark in this field. Therefore, developing a comprehensive and usable benchmark is an urgent need, because it can facilitate fresh researchers to contribute to this field and accelerate the advancement of this direction.

8 Conclusion

Large language model-enhanced recommender systems (LLMERS) have attracted much attention due to their effectiveness and practicability. In this paper, we conclude the most recent efforts in this field. According to the component of traditional RS that LLMERS targets, we categorize the papers into three lines, *i.e.*, knowledge, interaction and model enhancement. Furthermore, to facilitate and inspire the researchers, we summarize developing trends and give out several future directions.

Acknowledgments

This research was partially supported by Research Impact Fund (No.R1015-23), Collaborative Research Fund (No.C1043-24GF), Huawei (Huawei Innovation Research Program, Huawei Fellowship), National Natural Science Foundation of China (No.62192781, No.62177038, No.62293551, No.62277042, No.62137002, No.61721002, No.61937001, No.62377038), Project of China Knowledge Centre for Engineering Science and Technology.

References

- [1] Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *arXiv preprint arXiv:2308.08434* (2023).
- [2] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.
- [3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.
- [4] Bo Chen, Xinyi Dai, Huifeng Guo, Wei Guo, Weiwen Liu, Yong Liu, Jiarui Qin, Ruiming Tang, Yichao Wang, Chuhan Wu, et al. 2024. All roads lead to rome: Unveiling the trajectory of recommender systems across the llm era. *arXiv preprint arXiv:2407.10081* (2024).
- [5] Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. 2024. HLLM: Enhancing Sequential Recommendations via Hierarchical Large Language Models for Item and User Modeling. *arXiv preprint arXiv:2409.12740* (2024).
- [6] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.
- [7] Lei Chen, Chen Gao, Xiaoyi Du, Hengliang Luo, Depeng Jin, Yong Li, and Meng Wang. 2024. Enhancing ID-based Recommendation with Large Language Models. *arXiv preprint arXiv:2411.02041* (2024).
- [8] Zhixuan Chu, Yan Wang, Qing Cui, Longfei Li, Wenqing Chen, Sheng Li, Zhan Qin, and Kui Ren. 2024. Llm-guided multi-view hypergraph learning for human-centric explainable recommendation. *arXiv preprint arXiv:2401.08217* (2024).
- [9] Yu Cui, Feng Liu, Pengbo Wang, Bohao Wang, Heng Tang, Yi Wan, Jun Wang, and Jiawei Chen. 2024. Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Models. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 507–517.
- [10] Ziqiang Cui, Yunpeng Weng, Xing Tang, Fuyuan Lyu, Dugang Liu, Xiuqiang He, and Chen Ma. 2024. Comprehending Knowledge Graphs with Large Language Models for Recommender Systems. *arXiv preprint arXiv:2410.12229* (2024).
- [11] Yingpeng Du, Ziyang Wang, Zhu Sun, Haoyan Chua, Hongzhi Liu, Zhonghai Wu, Yining Ma, Jie Zhang, and Youchen Sun. 2024. Large Language Model with Graph Convolution for Recommendation. *arXiv preprint arXiv:2402.08859* (2024).
- [12] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. 2023. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.
- [13] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Zichuan Xu, Wenzheng Xu, Junyang Chen, and Renfu Li. 2024. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *AAAI*.
- [14] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE TMM* (2022).
- [15] Zichuan Fu, Xiangyang Li, Chuhan Wu, Yichao Wang, Kuicai Dong, Xiangyu Zhao, Mengchen Zhao, Huifeng Guo, and Ruiming Tang. 2023. A unified framework for multi-domain ctr prediction via large language models. *ACM Transactions on Information Systems* (2023).
- [16] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2025. LLM4Rerank: LLM-based Auto-Reranking Framework for Recommendations. In *Proceedings of the ACM on Web Conference 2025*. 228–239.
- [17] Jingtong Gao, Zhaocheng Du, Xiaopeng Li, Xiangyu Zhao, Yichao Wang, Xiangyang Li, Huifeng Guo, and Ruiming Tang. 2025. SampleLLM: Optimizing Tabular Data Synthesis in Recommendations. *arXiv preprint arXiv:2501.16125* (2025).
- [18] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524* (2023).
- [19] Binzong Geng, Zhaoxin Huan, Xiaolu Zhang, Yong He, Liang Zhang, Fajie Yuan, Jun Zhou, and Linjian Mo. 2024. Breaking the length barrier: Llm-enhanced CTR prediction in long textual user behaviors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2311–2315.
- [20] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.
- [21] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1725–1731.
- [22] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1096–1102.
- [23] Liam Hebert, Marialena Kyriakidi, Hubert Pham, Krishna Sayana, James Pine, Sukhdeep Sodhi, and Ambarish Jash. 2024. FLARE: Fusing Language Models and Collaborative Architectures for Recommender Enhancement. *arXiv preprint arXiv:2409.11699* (2024).
- [24] Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM on Web Conference 2024*. 103–111.
- [25] Zheng Hu, Zhe Li, Ziyun Jiao, Satoshi Nakagawa, Jiawen Deng, Shimin Cai, Tao Zhou, and Fuji Ren. 2024. Bridging the User-side Knowledge Gap in Knowledge-aware Recommendations with Large Language Models. *arXiv preprint arXiv:2412.13544* (2024).
- [26] Chengkai Huang, Tong Yu, Kaige Xie, Shuai Zhang, Lina Yao, and Julian McAuley. 2024. Foundation models for recommender systems: A survey and new perspectives. *arXiv preprint arXiv:2402.11143* (2024).
- [27] Feiran Huang, Zhenghang Yang, Junyi Jiang, Yuanchen Bei, Yijie Zhang, and Hao Chen. 2024. Large Language Model Interaction Simulator for Cold-Start Item Recommendation. *arXiv preprint arXiv:2402.09176* (2024).
- [28] Raksha Jalan, Tushar Prakash, and Niranjana Pedanekar. 2024. LLM-BRec: Personalizing Session-based Social Recommendation with LLM-BERT Fusion Framework. In *The Second Workshop on Generative Information Retrieval*.
- [29] Umair Javed, Kamran Shaukat, Ibrahim A Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhui Luo. 2021. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)* 16, 3 (2021), 274–306.
- [30] Jian Jia, Yipei Wang, Yan Li, Honggang Chen, Xuehan Bai, Zhaocheng Liu, Jian Liang, Quan Chen, Han Li, Peng Jiang, et al. 2024. Knowledge Adaptation from Large Language Model to Recommendation for Practical Industrial Application. *arXiv preprint arXiv:2405.03988* (2024).
- [31] Yangqin Jiang, Yuhao Yang, Lianghao Xia, Da Luo, Kangyi Lin, and Chao Huang. 2024. ReCLM: Recommendation Instruction Tuning. *arXiv preprint arXiv:2412.19302* (2024).
- [32] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-HLT*, Vol. 1. Minneapolis, Minnesota, 2.
- [33] Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook* (2021), 91–142.
- [34] Riwei Lai, Li Chen, Rui Chen, and Chi Zhang. 2024. A Survey on Data-Centric Recommender Systems. *arXiv preprint arXiv:2401.17878* (2024).
- [35] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2024. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 10146–10159.
- [36] Qingyao Li, Wei Xia, Kounianhua Du, Qiji Zhang, Weinan Zhang, Ruiming Tang, and Yong Yu. 2024. Learning Structure and Knowledge Aware Representation with Large Language Models for Concept Recommendation. *arXiv preprint arXiv:2405.12442* (2024).
- [37] Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. 2023. CTRL: Connect Collaborative and Language Model for CTR Prediction. *arXiv preprint arXiv:2306.02841* (2023).
- [38] Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024. A survey of generative search and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924* (2024).
- [39] Yunzhe Li, Junting Wang, Hari Sundaram, and Zhining Liu. 2025. A Zero-Shot Generalization Framework for LLM-Driven Cross-Domain Sequential Recommendation. *arXiv preprint arXiv:2501.19232* (2025).
- [40] Jianghao Lin, Xinyi Dai, Rong Shan, Bo Chen, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Large Language Models Make Sample-Efficient Recommender Systems. *arXiv preprint arXiv:2406.02368* (2024).
- [41] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. 2023. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817* (2023).
- [42] Dugang Liu, Shenxian Xian, Xiaolin Lin, Xiaolian Zhang, Hong Zhu, Yuan Fang, Zhen Chen, and Zhong Ming. 2024. A Practice-Friendly Two-Stage LLM-Enhanced Paradigm in Sequential Recommendation. *arXiv preprint arXiv:2406.00333* (2024).
- [43] Fan Liu, Yaqi Liu, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. 2023. Understanding Before Recommendation: Semantic Aspect-Aware Review Exploitation via Large Language Models. *arXiv preprint arXiv:2312.16275* (2023).
- [44] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems. *Transactions of the Association for Computational*

- Linguistics* 11 (2023), 1553–1571.
- [45] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. 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*. 452–461.
 - [46] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *Comput. Surveys* 57, 2 (2024), 1–17.
 - [47] Qidong Liu, Zhaopeng Qiu, Xiangyu Zhao, Xian Wu, Zijian Zhang, Tong Xu, and Feng Tian. 2025. A Contrastive Pretrain Model with Prompt Tuning for Multi-center Medication Recommendation. *ACM Transactions on Information Systems* 43, 3 (2025), 1–29.
 - [48] Qidong Liu, Xian Wu, Wanyu Wang, Yejing Wang, Yuanshao Zhu, Xiangyu Zhao, Feng Tian, and Yefeng Zheng. 2025. LLMEmb: Large Language Model Can Be a Good Embedding Generator for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12183–12191.
 - [49] Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. 2024. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 26701–26727.
 - [50] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1114.
 - [51] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024. Large language model distilling medication recommendation model. *arXiv preprint arXiv:2402.02803* (2024).
 - [52] Qidong Liu, Fan Yan, Xiangyu Zhao, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Feng Tian. 2023. Diffusion augmentation for sequential recommendation. In *Proceedings of the 32nd ACM International conference on information and knowledge management*. 1576–1586.
 - [53] Qidong Liu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Howard Zhong, Chong Chen, Xiang Li, Wei Huang, and Feng Tian. 2025. Bridge the Domains: Large Language Models Enhanced Cross-domain Sequential Recommendation. *arXiv preprint arXiv:2504.18383* (2025).
 - [54] Zhenghao Liu, Zulong Chen, Moufeng Zhang, Shaoyang Duan, Hong Wen, Liangyue Li, Nan Li, Yu Gu, and Ge Yu. 2024. Modeling User Viewing Flow using Large Language Models for Article Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*. 83–92.
 - [55] Jing Long, Liang Qu, Guanhua Ye, Tong Chen, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2024. Unleashing the Power of Large Language Models for Group POI Recommendations. *arXiv preprint arXiv:2411.13415* (2024).
 - [56] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.
 - [57] Sichun Luo, Jiansheng Wang, Aojun Zhou, Li Ma, and Linqi Song. 2024. Large Language Models Augmented Rating Prediction in Recommender System. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7960–7964.
 - [58] Sichun Luo, Yuxuan Yao, Bowei He, Yinya Huang, Aojun Zhou, Xinyi Zhang, Yuanzhang Xiao, Mingjie Zhan, and Linqi Song. 2024. Integrating large language models into recommendation via mutual augmentation and adaptive aggregation. *arXiv preprint arXiv:2401.13870* (2024).
 - [59] Weiqing Luo, Chonggang Song, Lingling Yi, and Gong Cheng. 2024. KELLMRec: Knowledge-Enhanced Large Language Models for Recommendation. *arXiv preprint arXiv:2403.06642* (2024).
 - [60] Xinchun Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, et al. 2024. QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou. *arXiv preprint arXiv:2411.11739* (2024).
 - [61] Yucong Luo, Qitao Qin, Hao Zhang, Mingyue Cheng, Ruiran Yan, Kefan Wang, and Jie Ouyang. 2024. Molar: Multimodal LLMs with Collaborative Filtering Alignment for Enhanced Sequential Recommendation. *arXiv preprint arXiv:2412.18176* (2024).
 - [62] Hanjia Lyu, Ryan Rossi, Xiang Chen, Md Mehrab Tanjim, Stefano Petrangeli, Somdeb Sarkhel, and Jiebo Luo. 2024. X-Reflect: Cross-Reflection Prompting for Multimodal Recommendation. *arXiv preprint arXiv:2408.15172* (2024).
 - [63] Qiya Ma, Menglin Yang, Mingxuan Ju, Tong Zhao, Neil Shah, and Rex Ying. 2024. Harec: Hyperbolic graph-llm alignment for exploration and exploitation in recommender systems. *arXiv preprint arXiv:2411.13865* (2024).
 - [64] Shutong Qiao, Chen Gao, Yong Li, and Hongzhi Yin. 2024. LLM-assisted Explicit and Implicit Multi-interest Learning Framework for Sequential Recommendation. *arXiv preprint arXiv:2411.09410* (2024).
 - [65] Shutong Qiao, Chen Gao, Junhao Wen, Wei Zhou, Qun Luo, Peixuan Chen, and Yong Li. 2024. LLM4SBR: A Lightweight and Effective Framework for Integrating Large Language Models in Session-based Recommendation. *arXiv preprint arXiv:2402.13840* (2024).
 - [66] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3464–3475.
 - [67] Yankun Ren, Zhongde Chen, Xinxing Yang, Longfei Li, Cong Jiang, Lei Cheng, Bo Zhang, Linjian Mo, and Jun Zhou. 2024. Enhancing Sequential Recommenders with Augmented Knowledge from Aligned Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 345–354.
 - [68] Rong Shan, Jianghao Lin, Chenxu Zhu, Bo Chen, Menghui Zhu, Kangning Zhang, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. An Automatic Graph Construction Framework based on Large Language Models for Recommendation. *arXiv preprint arXiv:2412.18241* (2024).
 - [69] Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. 2024. Language Models Encode Collaborative Signals in Recommendation. *arXiv preprint arXiv:2407.05441* (2024).
 - [70] Chonggang Song, Chunxu Shen, Hao Gu, Yaoming Wu, Lingling Yi, Jie Wen, and Chuan Chen. 2024. PRECISE: Pre-training Sequential Recommenders with Collaborative and Semantic Information. *arXiv preprint arXiv:2412.06308* (2024).
 - [71] Tianrui Song, Wenshuo Chao, and Hao Liu. 2024. Large Language Model Enhanced Hard Sample Identification for Denoising Recommendation. *arXiv preprint arXiv:2409.10343* (2024).
 - [72] Yuqi Sun, Qidong Liu, Haiping Zhu, and Feng Tian. 2025. LLMSeR: Enhancing Sequential Recommendation via LLM-based Data Augmentation. *arXiv preprint arXiv:2503.12547* (2025).
 - [73] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, and Jun Xu. 2024. Large Language Models Enhanced Collaborative Filtering. *arXiv preprint arXiv:2403.17688* (2024).
 - [74] Jiayi Tang and Ke Wang. 2018. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2289–2298.
 - [75] Jiahao Tian, Jinman Zhao, Zhenkai Wang, and Zhicheng Ding. 2024. Mmrec: Llm based multi-modal recommender system. *arXiv preprint arXiv:2408.04211* (2024).
 - [76] Arpita Vats, Vinija Jain, Rahul Raja, and Aman Chadha. 2024. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv preprint arXiv:2402.18590* (2024).
 - [77] Zhizhong Wan, Bin Yin, Junjie Xie, Fei Jiang, Xiang Li, and Wei Lin. 2024. LARR: Large Language Model Aided Real-time Scene Recommendation with Semantic Understanding. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 23–32.
 - [78] Bohao Wang, Feng Liu, Jiawei Chen, Yudi Wu, Xingyu Lou, Jun Wang, Yan Feng, Chun Chen, and Can Wang. 2024. LLM4DSR: Leveraging Large Language Model for Denoising Sequential Recommendation. *arXiv preprint arXiv:2408.08208* (2024).
 - [79] Hangyu Wang, Jianghao Lin, Xiangyang Li, Bo Chen, Chenxu Zhu, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. FLIP: Fine-grained Alignment between ID-based Models and Pretrained Language Models for CTR Prediction. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 94–104.
 - [80] Jianling Wang, Haokai Lu, James Caverlee, Ed H Chi, and Minmin Chen. 2024. Large Language Models as Data Augmenters for Cold-Start Item Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*. 726–729.
 - [81] Menghan Wang, Yuchen Guo, Duanfeng Zhang, Jianian Jin, Minnie Li, Dan Schonfeld, and Shawn Zhou. 2024. Enabling Explainable Recommendation in E-commerce with LLM-powered Product Knowledge Graph. *arXiv preprint arXiv:2412.01837* (2024).
 - [82] Xingmei Wang, Weiwen Liu, Xiaolong Chen, Qi Liu, Xu Huang, Defu Lian, Xiangyang Li, Yasheng Wang, Zhenhua Dong, and Ruiming Tang. 2024. CELA: Cost-Efficient Language Model Alignment for CTR Prediction. *arXiv preprint arXiv:2405.10596* (2024).
 - [83] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Zhang, Qing Cui, et al. 2024. LLMRG: Improving Recommendations through Large Language Model Reasoning Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19189–19196.
 - [84] Yuhao Wang, Junwei Pan, Xiangyu Zhao, Pengyue Jia, Wanyu Wang, Yuan Wang, Yue Liu, Dapeng Liu, and Jie Jiang. 2024. Pre-train, Align, and Disentangle: Empowering Sequential Recommendation with Large Language Models. *arXiv preprint arXiv:2412.04107* (2024).
 - [85] Yu Wang, Lei Sang, Yi Zhang, and Yiwen Zhang. 2025. Intent Representation Learning with Large Language Model for Recommendation. *arXiv preprint arXiv:2502.03307* (2025).
 - [86] Yuling Wang, Changxin Tian, Binbin Hu, Yanhua Yu, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Liang Pang, and Xiao Wang. 2024. Can Small Language Models be Good Reasoners for Sequential Recommendation?. In *Proceedings of the ACM on Web Conference 2024*. 3876–3887.
 - [87] Yuhao Wang, Yichao Wang, Zichuan Fu, Xiangyang Li, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. LLM4MSR: An LLM-Enhanced Paradigm for

- Multi-Scenario Recommendation. *arXiv preprint arXiv:2406.12529* (2024).
- [88] Yejing Wang, Dong Xu, Xiangyu Zhao, Zhiren Mao, Peng Xiang, Ling Yan, Yao Hu, Zijian Zhang, Xuetao Wei, and Qidong Liu. 2024. GPRC: Bi-level User Modeling for Deep Recommenders. *arXiv preprint arXiv:2410.20730* (2024).
- [89] Yuhao Wang, Xiangyu Zhao, Bo Chen, Qidong Liu, Huifeng Guo, Huanshuo Liu, Yichao Wang, Rui Zhang, and Ruiming Tang. 2023. PLATE: A prompt-enhanced paradigm for multi-scenario recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1498–1507.
- [90] Yejing Wang, Xiangyu Zhao, Tong Xu, and Xian Wu. 2022. Autofield: Automating feature selection in deep recommender systems. In *Proceedings of the ACM Web Conference 2022*. 1977–1986.
- [91] Azmine Tushik Wasi. 2024. HRGraph: Leveraging LLMs for HR Data Knowledge Graphs with Information Propagation-based Job Recommendation. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*. 56–62.
- [92] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 806–815.
- [93] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.
- [94] Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 12–22.
- [95] Yunjia Xi, Weiwen Liu, Jianghao Lin, Muyan Weng, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Yong Yu, et al. 2024. Efficient and Deployable Knowledge Infusion for Open-World Recommendations via Large Language Models. *arXiv preprint arXiv:2408.10520* (2024).
- [96] Lanling Xu, Zhen Tian, Bingqian Li, Junjie Zhang, Jinpeng Wang, Mingchen Cai, and Wayne Xin Zhao. 2024. Sequence-level Semantic Representation Fusion for Recommender Systems. *arXiv preprint arXiv:2402.18166* (2024).
- [97] Wujiang Xu, Zujie Liang, Jiaojiao Han, Xuying Ning, Wenfang Lin, Linxun Chen, Feng Wei, and Yongfeng Zhang. 2024. SLMRec: Empowering Small Language Models for Sequential Recommendation. *arXiv preprint arXiv:2405.17890* (2024).
- [98] Wentao Xu, Qianqian Xie, Shuo Yang, Jiangxia Cao, and Shuchao Pang. 2024. Enhancing Content-based Recommendation via Large Language Model. *arXiv preprint arXiv:2404.00236* (2024).
- [99] Yuki Yada and Hayato Yamana. 2024. News Recommendation with Category Description by a Large Language Model. *arXiv preprint arXiv:2405.13007* (2024).
- [100] Shenghao Yang, Weizhi Ma, Peijie Sun, Qingyao Ai, Yiqun Liu, Mingchen Cai, and Min Zhang. 2024. 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*. 335–344.
- [101] Shenghao Yang, Weizhi Ma, Peijie Sun, Min Zhang, Qingyao Ai, Yiqun Liu, and Mingchen Cai. 2024. Common Sense Enhanced Knowledge-based Recommendation with Large Language Model. *arXiv preprint arXiv:2403.18325* (2024).
- [102] Xihong Yang, Heming Jing, Zixing Zhang, Jindong Wang, Huakang Niu, Shuaiqiang Wang, Yu Lu, Junfeng Wang, Dawei Yin, Xinwang Liu, et al. 2024. DaRec: A Disentangled Alignment Framework for Large Language Model and Recommender System. *arXiv preprint arXiv:2408.08231* (2024).
- [103] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* (2024), nwae403.
- [104] Chiyu Zhang, Yifei Sun, Minghao Wu, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Rong Jin, Angli Liu, Ji Zhu, Sem Park, et al. 2024. EmbSum: Leveraging the Summarization Capabilities of Large Language Models for Content-Based Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1010–1015.
- [105] Chao Zhang, Shiwei Wu, Haoxin Zhang, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024. NoteLLM: A Retrievable Large Language Model for Note Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*. 170–179.
- [106] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024. NoteLLM-2: Multimodal Large Representation Models for Recommendation. *arXiv preprint arXiv:2405.16789* (2024).
- [107] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 52, 1 (2019), 1–38.
- [108] Wenlin Zhang, Chuhan Wu, Xiangyang Li, Yuhao Wang, Kuicai Dong, Yichao Wang, Xinyi Dai, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. LLMTreeRec: Unleashing the Power of Large Language Models for Cold-Start Recommendations. *arXiv preprint arXiv:2404.00702* (2024).
- [109] Weizhi Zhang, Liangwei Yang, Woosong Yang, Henry Peng Zou, Yuqing Liu, Ke Xu, Sourav Medya, and Philip S Yu. 2025. LLMInit: A Free Lunch from Large Language Models for Selective Initialization of Recommendation. *arXiv preprint arXiv:2503.01814* (2025).
- [110] Zijian Zhang, Shuchang Liu, Ziru Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. 2025. Llm-powered user simulator for recommender system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13339–13347.
- [111] Hongke Zhao, Songming Zheng, Likang Wu, Bowen Yu, and Jing Wang. 2024. Lane: Logic alignment of non-tuning large language models and online recommendation systems for explainable reason generation. *arXiv preprint arXiv:2407.02833* (2024).
- [112] Qian Zhao, Hao Qian, Ziqi Liu, Gong-Duo Zhang, and Lihong Gu. 2024. Breaking the Barrier: Utilizing Large Language Models for Industrial Recommendation Systems through an Inferential Knowledge Graph. *arXiv preprint arXiv:2402.13750* (2024).
- [113] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [114] Xiangyu Zhao, Maolin Wang, Xinjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. 2023. Embedding in Recommender Systems: A Survey. *arXiv:2310.18608 [cs.LG]*
- [115] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [116] Ziwei Zhao, Fake Lin, Xi Zhu, Zhi Zheng, Tong Xu, Shitian Shen, Xueying Li, Zikai Yin, and Enhong Chen. 2024. DynLLM: When Large Language Models Meet Dynamic Graph Recommendation. *arXiv preprint arXiv:2405.07580* (2024).
- [117] Bowen Zheng, Zihan Lin, Enze Liu, Chen Yang, Enyang Bai, Cheng Ling, Wayne Xin Zhao, and Ji-Rong Wen. 2024. A Large Language Model Enhanced Sequential Recommender for Joint Video and Comment Recommendation. *arXiv preprint arXiv:2403.13574* (2024).
- [118] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.