Personalized News Recommendation Towards the Era of LLMs: Review and Prospect

Jie Li, Zeyi Liu, Linmei Hu, Yunbo Rao, Bo Liu, Bo Fang, and Liqiang Nie, *Senior Member, IEEE*

(Survey Paper)

***Abstract*—With the prevalence of online news services, personalized news recommendation (PNR) has played an indispensable role in meeting users' needs and mitigating information overload, with the aim of providing news articles that cater to user preferences. Despite significant progress made in the field of PNR over the past few decades, their performances are still hindered by some limitations, such as insufficient news modeling, difficulties in effectively modeling diverse user interests, and ignorance of fine-grained matching signals. It is fortunate that the emergence of large language models (LLMs) provides a promising insight into empowering the capabilities of news recommendation. Known for their impressive capabilities of natural language understanding and generation, LLMs have achieved disruptive achievements in various natural language processing (NLP) tasks, which motivates us to integrate LLMs into news recommendation and benefits from them to make up existing deficiencies. In this paper, we conduct a comprehensive review of current efforts made towards utilizing LLMs for PNR, with a focus on three core modules involved in the news recommendation process, i.e., news modeling, user modeling, and accurate matching. We systematically discuss and analyze relevant works under each focus. In addition, we point out several potential research directions to provide more inspiration for future investigation in this thriving field.**

***Index Terms*—Personalized news recommendation, large language models, news modeling, user modeling, accurate matching.**

I. Introduction

WITH the rapid development of the Internet, people are inclined to access daily information from online news platforms, such as MSN News.[1] As huge amounts of news always make users exhaust their patience to find the real-needed information, personalized news recommendation (PNR), which aims to pick up news based on user preferences, has played a vital role in meeting users' needs and alleviating information overload [1], [2]. Traditional PNR methods leverage manual feature engineering to model news and users for matching [3], [4], [5], [6], [7], [8]. Later with the emergence of deep learning techniques, several neural networks are introduced to automatically learn news and user representations [9], [10], [11], [12], [13], [14], [15], as well as the subsequent matching process [16], [17], [18].

Despite the significant progress made by conventional PNR techniques over the past few decades, their recommendation performances are still hampered by three major limitations as follows: First, conventional news recommendation extracts news features merely from in-domain data, which lacks general-world knowledge to enhance the comprehension of deep semantic information within news texts [19], [20], [21], especially in cold-start scenarios. Second, compared to news content that is generally static, user interests are often dynamic and constantly evolving over time. Conventional news recommendation struggles to fully understand inherent correlations between different user behaviors in various contexts [22]. Meanwhile, it's also hard for them to infer accurate user interests when user behaviors are sparse [2]. Third, most conventional news recommendation relies on explicit semantic representations to match users and candidate news, which neglects fine-grained matching signals with low-level interactions [23]. In addition, users are unable to actively guide conventional models to meet their specific requirements via detailed instructions in natural language, which restricts them from achieving more accurate recommendation results on downstream PNR tasks [24].

Fortunately, the advent of large language models (LLMs) has brought a promising insight into the field of news recommendation. LLMs that are pre-trained on vast corpora of textual data, have demonstrated impressive general intelligence on various natural language processing (NLP) tasks due to their strong language understanding and generation abilities [25], [26], [27], [28], [29]. These capabilities empower them to deliver more context-aware and personalized recommendations. Therefore, the booming of LLMs suggests their potential to promote the

Received 18 July 2024; revised 20 March 2025; accepted 10 June 2025. Date of publication 20 June 2025; date of current version 24 July 2025. This work was supported in part by the National Science Foundation of China under Grant 62276029, in part by Beijing Institute of Technology Research Fund Program for Young Scholars under Grant 6120220261, and in part by CIPSC-SMP-Zhipu Large Model Cross-Disciplinary Fund. Recommended for acceptance by X. Zhu.
(Corresponding author: Linmei Hu.)

Jie Li and Zeyi Liu are with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jli@bupt.edu.cn; iliuzeyi@163.com).

Linmei Hu is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: hulinmei@bit.edu.cn).

Yunbo Rao is with the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: uestc2008@126.com).

Bo Liu is with China CSCEC Real Estate Company, Ltd., Shanghai 200001, China (e-mail: 13801093053@139.com).

Bo Fang is with the Network department, China Mobile Communications Corporation, Beijing 100033, China (e-mail: fangbo@chinamobile.com).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China (e-mail: nieliqiang@gmail.com).

Digital Object Identifier 10.1109/TKDE.2025.3581806

[1]https://www.msn.com/en-us/news

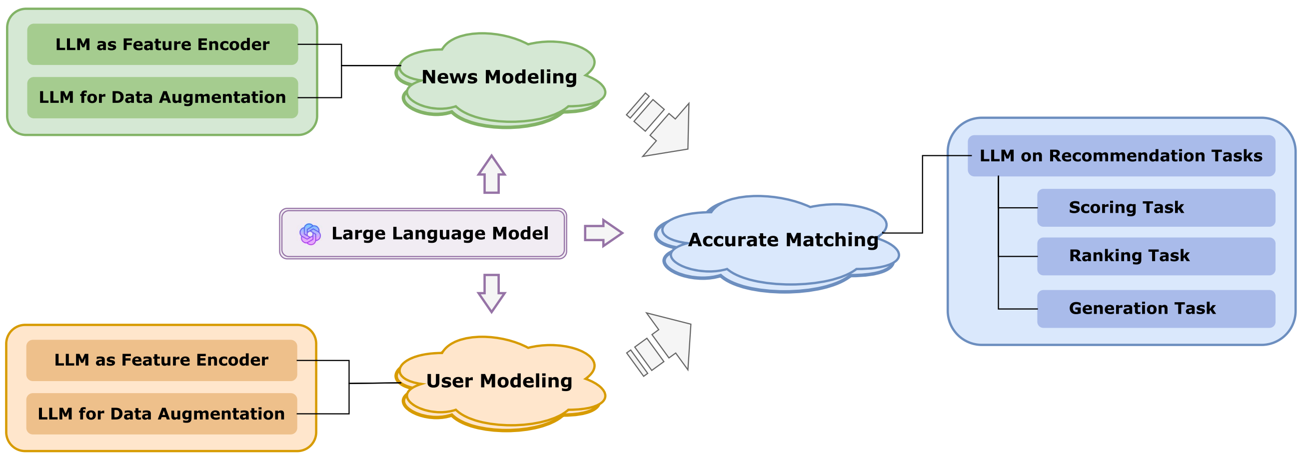

Fig. 1. The illustrative dissection of utilizing LLMs in personalized news recommendation. We exhibit that LLMs can be adopted to three core modules involved in the news recommendation process, i.e., news modeling, user modeling, and accurate matching.

innovation in conventional paradigms of news recommendation, i.e., we can incorporate LLMs into PNR and benefit from their emergent abilities to compensate for the aforementioned limitations. More specifically, LLMs with powerful language understanding ability and profound world knowledge can be exploited to enhance deep semantic comprehension of news texts, and explore the intrinsic relations between user behaviors to facilitate full comprehension of the evolution on diverse user interests. Their zero/few-shot learning capabilities make it possible to generate synthetic content that may be of interest to new users with limited interaction data. LLMs can also be leveraged to learn fine-grained semantic matching between news and users with their exceptional understanding abilities. Besides, LLMs allow news recommendation to grasp user intents in the form of natural language. Although a few recent surveys have summarized employing LLMs for general recommendation systems [22], [30], [31], [32], there still lacks an in-depth review of the integration of LLMs in the the news recommendation field.

In this paper, we provide a comprehensive and systematic review of current efforts made on utilizing LLMs for PNR, with a focus on three core modules involved in the news recommendation process. (1) News modeling, that is, how to model the rich information from news texts. We discuss the utilization of LLMs as feature encoders to enhance semantic comprehension of news texts, and as auxiliary textual feature generators for data augmentation. (2) User modeling, that is, how to model diverse user interests from user behaviors. We explore the use of LLMs as feature encoders to fully analyze user interests derived from click behaviors. Additionally, we consider utilizing LLMs to generate supplementary user features, thereby enriching the training data. (3) Accurate matching, that is, how to accurately match user interests with candidate news. We concentrate on three kinds of recommendation tasks that LLMs solve to achieve ultimate recommendation results, i.e., scoring task, ranking task, and generation task. The illustrative diagram of our research scheme is shown in Fig. 1. In general, these three core modules constitute the foundation of news recommendation, making it accessible for researchers to trace the technical advancements of this thriving field from the side of PNR. To the best of our knowledge, this survey is the first work that concludes related advancements of news recommendation utilizing LLMs.

The rest of the paper is arranged as follows. In Section II, we first briefly conduct an overview of personalized news recommendation based on its framework and research status. In Section III, we outline the background of large language models and pay attention to aspects where PNR can benefit from the utilization of LLMs. Then in Sections IV, V, and VI, we thoroughly analyze the current efforts spent on utilizing LLMs for PNR, focusing on three core modules that exist in the news recommendation process, i.e., news modeling, user modeling, and accurate matching. Detailed discussions of the latest research progress are provided under each focus. From these three perspectives, we propose feasible and instructive insights into the investigation of news recommendation towards the era of LLMs. We also introduce several existing datasets for news recommendation in Section VII. Finally, we present some potential directions that are worth further exploration in Section VIII and conclude in Section IX, hoping to draw a hopeful vision for researchers in this thriving field.

II. Overview of News Recommendation

Personalized news recommendation has received extensive concern of academia and industry due to its effectiveness in meeting user needs and solving information overload. As shown in Fig. 2, the general framework of personalized news recommendation system can be characterized as a workflow with three key stages. Given a user with a certain context (usually his/her historical behaviors) and a set of candidate news, the system first (1) transforms news features into news representations and (2) learns user representations from user context via specific encoders, and then (3) match news and user representations based on their relevance to provide tailored recommendation results.

The foregoing three stages correspond to three core modules that exist in the personalized news recommendation, i.e., news modeling, user modeling, and accurate matching. In this section,

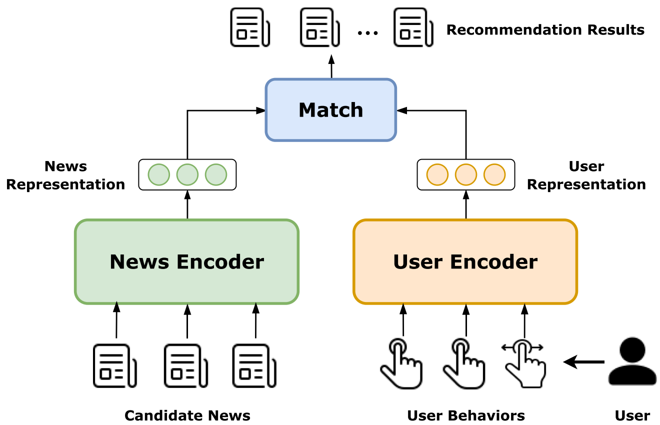

Fig. 2. A general framework of personalized news recommendation system.

we conduct a systematic review of these modules to summarize the efforts made for news recommendation before the emergence of LLMs.

A. News Modeling

News modeling stands for the core techniques of capturing characteristics and understanding the content of news articles. Earlier works tend to utilize manual feature engineering to model news, which heavily depends on handcrafted informative features. Collaborative filtering (CF) based methods usually build news embeddings based on descriptive features such as news IDs [3], [33], [34], [35], [36], [37]. However, these methods are exposed to suffer from severe cold-start problems due to the rapid updates and short timelines of news. Therefore, several methods attempt to incorporate content features extracted from news texts into the news modeling process, including but not limited to semantic features [4], [6], [7], [8], [38], [39], [40], [41], [42], topic distribution [5], emotion features [43], [44], [45], and multimodal information [46]. In addition, many other meta features of news are also taken into consideration by news modeling, such as categories [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], locations [59], [60], [61], [62], publishers [57], [60], [63], time stamp [34], [48], [55], [60], [64], as well as popularity [38], [48], [52], [53], [54], [58], [59], [60], [65] and recency [38], [47], [52], [53], [57], [58], [59], [60]. Though the above methods have covered a wide range of news information, they still require specific expertise for feature designment.

With the development of deep learning techniques, neural networks are widely applied in modeling feature representations of news by a series of works [9], [10], [11], [12], [13], [14], [15], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78]. These methods automatically learn news representations from their original texts, rather than using handcrafted features. For example, Okura et al. [66] employed denoising autoencoders to learn news representations from its textual content. Wang et al. [9] and Zhu et al. [10] presented to leverage convolutional neural network (CNN) to extract news representations. Wu et al. [13] utilized multi-head self-attention networks to learn contextual representations of news titles. Wu et al. [14] treated titles, bodies, and categories as news information from different views, and developed a multi-view attentive network to learn the unified representation of news. Among them, CNN and attention mechanism are the two most commonly used architectures, since CNN is effective in capturing local contexts of news texts and varient attention mechanisms are efficient in selecting important features. Besides, some methods are not only limited to leveraging semantic information extracted from news texts, but also introduce knowledge or commonsense information to assist text modeling [9], [10], [73], [79], [80]. For example, Wang et al. [9] employed knowledge graph embedding to represent entities identified from news texts, and combined them with word semantic representations via knowledge-aware CNN. Zhu et al. [10] treated entities extracted from news articles as additional textual features and modeled them via CNN.

In addition to the sequence modeling techniques mentioned above, several works have also explored to capture high-order structural interactions [81], [82], [83], [84], [85], [86], [87] or semantic-augmented information [88] in the form of graphs to enhance news modeling. For example, Hu et al. [82], [83] constructed heterogeneous graphs with news and user interactions, and applied graph convolutional network (GCN) to learn news representations. Yang et al. [87] first built a global news graph where two news are connected only if they are clicked by the same user in sequence, and then adopted graph neural networks to encode news interactions. Mao et al. [88] constructed a news-graph for each candidate news with its most semantic-relevant articles as neighbors, and then aggregated information on graphs to enrich semantic representations of news. All of these methods attempt to enhance news modeling with graph information for better recommendations.

B. User Modeling

Similar to news modeling, user modeling also lies in the core place of personalized news recommendation, with the duty of inferring users' personal interests in news. Earlier methods leverage manual feature engineering to represent user preferences, such as CF-based methods that generally derive user embeddings from their IDs [3], [33]. However, these approaches often suffer from the issue of data sparsity, since the huge number of news always results in a situation where each news has only been clicked by a few users. Several methods incorporate user behaviors, especially the historical interaction sequences, into the user modeling process to infer user preferences by aggregating features of clicked news [4], [5], [6]. For example, Garcin et al. [5] aggregated LDA features of all the clicked news to obtain user vectors. It is noteworthy that these methods are hard to model accurate user interests when behaviors are sparse. So except for the news that users have interactions with, some works also explore to introduce additional user features and other kinds of behaviors (e.g., dwell time [89] and access pattern [52], [90]) as supplements. Common user features, including interest clusters [47], [57], [58], as well as meta features, including demographic characteristics (e.g., age, gender, profession, economic grade, etc.) [47], [48], [60], [91] and geographic location information [61], [64], are taken into account.

Compared with traditional feature engineering methods that need handcrafted rules or heuristic patterns, deep learning-based works focus on utilizing neural networks to infer user preference patterns from their historical behaviors in an end-to-end approach. Several methods directly aggregate representations of clicked news to acquire user representations [9], [14], [23], [71], [72], [79], [92], [93], [94]. For example, Wu et al. [14] developed a news-level attention network to model user interests via the aggregation of browsed news. Though these methods could select important behaviors to learn informative user representations, they overlook modeling the sequential relations among click news which provide rich context information. As a result, several works have attempted to leverage sequence models, such as GRU [12], [66], [73], [75], [77], [95], [96] and LSTM [10], [67], [83], [97], [98], [99], [100] to model the strong sequential dependency between clicked news. For example, Zhu et al. [10] utilized an attention-based LSTM to capture sequential features of users' clicked news to obtain user representations. Besides, other commonly used deep neural networks, such as CNN [15], [16], [69], [70], [101] and self-attention [10], [13], [76], [85], [95], [102], [103], [104], [105], [106] are also applied broadly in modeling users' local or global contexts. For example, Qi et al. [15] proposed candidate-aware CNN and self-attention networks to respectively learn short-term and long-term user representations from contexts of clicked news.

Instead of modeling user behaviors as sequences, a few works model each user's behavior contexts as a personalized user graph and apply graph neural networks to encode high-order relations among different click behaviors [73], [74], [86], [88], [107], [108]. For example, Sheu et al. [73] employed GCN to uncover the structural information among news articles within a user session. Wu et al. [86] built a personalized heterogeneous graph for each user based on their click behaviors, and developed a graph pooling method to learn the user representation via aggregating node features. Besides, user-news graphs constructed from collaborative information between users and news are also considered by several works [81], [82], [83], [84], [85], [109], [110] to learn graph-based user representations. For instance, Hu et al. [82] exploited the high-order relations on global user-news graphs to encode user representations via feature propagation. Hu et al. [83] constructed a user-news-topic graph and took advantage of GNN to model user representations. Their approach also alleviated the problem of cold-start and user-item interactions sparsity, as they incorporated non-clicked news into graph construction through topic connections.

C. Accurate Matching

In general, the ultimate goal of personalized news recommendation is the accurate matching between user interest and candidate news for click prediction. Previous works typically adopt the two-tower architectures, which model news and user representation vectors separately, and then directly match these two vectors based on their semantic relevance [4], [5], [111]. For example, Okura et al. [66] leveraged denoising autoencoders to encode news from its content and GRU to encode user interests from their browsed news independently, and then matched candidate news for users based on inner product operations. Wang et al. [9] employed knowledge-aware CNN to model news and attention network to characterize user interests. The final user and news embeddings were matched through deep neural network (DNN). Zhu et al. [10] computed cosine similarity between user and candidate news embeddings for personalized matching. After this, several methods utilized GRU [12], [77], CNN [12], [15], multi-head self-attention [13], [15], [76], [77] and variant attention mechanisms [11], [14] to learn news and user representations, and then matched them via inner product to obtain matching scores used for ranking candidate news.

However, users always have diverse interests and news articles always contain multiple aspects, simply compressing all the information into one single vector may lose fine-grained matching signals and make it insufficient to accurately match users with news [16], [17], [18], [112]. More specially, only part of the user interests or news aspects can be matched with specific candidate news or user. Therefore, several methods have considered modeling the interactions between user interests and candidate news to achieve fine-grained matching [16], [17], [18], [82]. For example, Wang et al. [16] first learned multi-level representations for each news, and then performed matching between representation pairs of each browsed news and candidate news at each level to acquire fine-grained information, rather than just fusing users' browsed news into a single vector. Hu et al. [82] introduced a graph neural network with preference disentanglement to explore user and news interactions under different latent preference spaces on user-news graphs. Li et al. [18] developed a poly attention scheme that learns multiple interest vectors from different aspects for each user, and aggregated matching scores between each interest vector with candidate news vectors to acquire the final click score.

III. LLMs in News Recommendation

In order to present a thorough overview for PNR in the era of LLMs, in this section, we first introduce the basic concepts of large language models, and then pay attention to how personalized news recommendation can benefit from the utilization of LLMs.

A. Large Language Models

Language model (LM) plays a vital role in understanding human languages, as it aims to model the probability distribution of word sequences and predict the likelihood of specific tokens appearing in a given context. With the rise of Transformer architecture [113], language models are first pre-trained on extensive textual data and then fine-tuned via task-specific objectives to adapt to various downstream tasks. These pre-trained language models (PLMs) mainly fall into three classical categories: encoder-only models (e.g., BERT [114]), decoder-only models (e.g., GPT [115]) and encoder-decoder models (e.g., T5 [116] and BART [117]).

As the scale of models continues to increase, large language models (LLMs) with billions or trillions of parameters have demonstrated improved capacity on downstream tasks, such as GPT-3 [25], GPT-3.5, LLaMA [118], PaLM [119] and ChatGLM [120]. These models are generally built upon Transformer architecture and trained on more massive textual corpora. But compared with traditional PLMs, LLMs exhibit some emergent abilities (e.g., in-context learning, instruction following, and step-by-step reasoning) that are not present in previous smaller ones, especially when their parameter scale exceeds a certain threshold [27]. Equipped with these powerful abilities, LLMs could better understand natural languages and generate human-like texts, which revolutionize the research paradigm of NLP. Moreover, LLMs also show remarkable potential in a wide range of real-world applications such as healthcare [121], [122], education [123], [124], finance [125], [126], and law [127], [128].

B. LLMs for News Recommendation

With the prosperous development of large language models, recent studies have explored to harness LLMs as advanced techniques for recommendation systems. A few surveys outline the current research state of the recommendation field [22], [30], [31]. For example, Lin et al. [22] reviewed LLM-enhanced recommendation systems from the perspective of where and how to adapt LLMs. As an indispensable scenario of general recommendations, news recommendation has also attracted growing interest in combining LLMs, since they are effective in boosting recommendation performance and enhancing user experience. We believe that PNR could benefit from the following perspectives of LLMs:

First, LLMs possess powerful contextual understanding abilities. Since they are pre-trained on a quantity of mixture-of-source corpora, LLMs excel in understanding a wide range of general-world knowledge. Aided by their powerful knowledge and representation capability, news recommendation utilizing LLMs can uncover semantic nuances within news texts and deeply comprehend users' personal interests.

Second, LLMs demonstrate excellent content generation abilities and flexible interaction patterns. These enable news recommendation system to acquire users' preferences or needs through natural language instructions that are more human-like in nature, and even through real-time interactions such as dialogue and chitchat. Due to the generation ability of LLMs, they can create or adjust recommendation suggestions to improve the interactivity of news recommendation.

Third, LLMs exhibit unprecedented emergent abilities that are not present in previous smaller models, such as strong reasoning abilities. News recommendation with LLMs could infer user interests more accurately from their historical behavior sequences via step-by-step reasoning, and thus provide more explainable recommendation results. Moreover, in-context learning capabilities in zero/few-shot settings offer a new paradigm to evoke the adaptation of LLMs on downstream recommendation tasks.

From the above, we can see that LLMs offer a promising alternative for news recommendation. In this survey, we attempt to delve into the booming works of PNR utilizing LLMs,[2] and summarize them from the perspective of three key modules: (1) how to model the textual information of news content, i.e., news modeling, (2) how to model user interests based on their past behaviors, i.e., user modeling, and (3) how to accurately match user interest with candidate news, i.e., accurate matching. The specific classification of applying LLMs in PNR as well as the relevant works under each category are shown in Fig. 3.

[2]Here, we broaden the boundary of LLMs and take relatively smaller models into consideration, to provide a more comprehensive review.

IV. LLMs for News Modeling

News modeling focuses on representing the rich textual information of news articles. Equipped with powerful natural language understanding and generation abilities, LLMs can be exploited to enhance semantic comprehension of news texts and generate auxiliary textual features as supplements. In this section, we roughly divided related works into two categories: (1) LLM as feature encoder and (2) LLM for data augmentation.

A. LLMs as Feature Encoder

Extracting rich textual information within news content is crucial for understanding news articles. Recent works have delved into leveraging LLMs as news encoders to learn high-quality feature representations for news text modeling [19], [20], [21], [129], [130], [131], [132], [133], [134]. For example, UNBERT [19] employs relatively small pre-trained language models to enhance the deep semantic comprehension of news titles. LLMs with implicit general-world knowledge can be utilized to mitigate the cold-start problem. PLM-NR [20] explores to mine the underlying contexts of news texts via LLMs with strong text modeling ability, and demonstrates the effective performance improvement of news recommendation empowered by LLMs. LKPNR [21] presents an LLM-augmented news encoder, which leverages the powerful text understanding ability of LLMs to learn news representations that encode rich semantic information.

It is clear that equipped with exceptional language modeling capacity and extensive world knowledge, LLMs could better facilitate the deep contextual comprehension of news content, especially in such text-abundant scenarios. On the one hand, LLMs are mainly built based on transformer architectures, which makes them perform well in solving long-distance dependency problems in news text modeling. On the other hand, LLMs introduce out-domain knowledge beyond the recommendation field to reveal intricate linguistic relations and subtle semantic distinctions within news texts, which avoids being confined to specific in-domain data. Furthermore, Li et al. [135] have suggested that increasing the size of LLMs could achieve higher recommendation accuracy when they are employed as text encoders. And their performance might be further improved if more powerful LLMs are developed.

The ultimate goal of news encoders is to obtain a unified news representation. As shown in Fig. 4, this can only be achieved through open-source LLMs such as ChatGLM and LLaMA, since closed-source LLMs such as ChatGPT are only accessible to token outputs. Most of these methods first initialize feature encoders with pre-trained language models, and then align the general language space with the recommendation space via

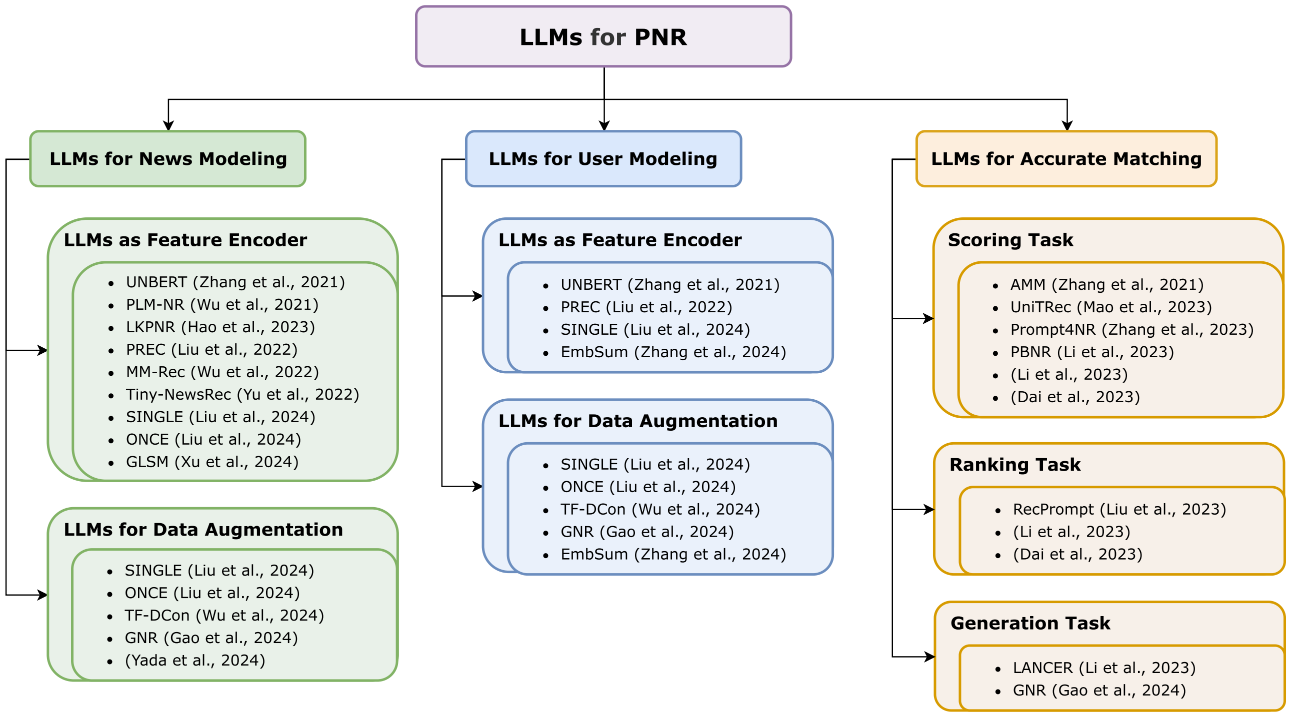

Fig. 3. The category of relevant works on utilizing LLMs for PNR.

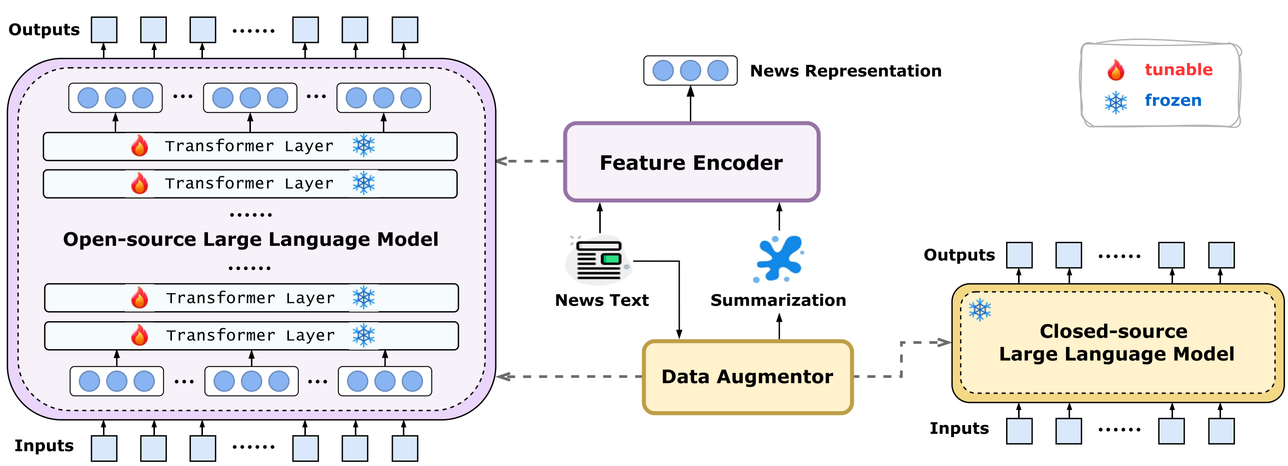

Fig. 4. A general paradigm of LLMs for news modeling. The feature encoder can be implemented with open-source LLMs, as its goal is to acquire a unified news representation. Note that, the parameters of each layer in open-source LLMs can be tunable or frozen during the training process. The data augmentor can be achieved through either open-source or closed-source LLMs to summarize various textual contents as additional news features.

fine-tuning strategy. During this process, some adopt full fine-tuning which updates all parameters of the entire model [19], while others only fine-tune the last few layers and keep the rest frozen (a type of parameter-efficient fine-tuning) [20], [130], [133], [134]. For example, UNBERT [19] fine-tune all model hyper-parameters on the MIND dataset, and PLM-NR [20] tune only the last two layers of Transformer. Besides, few methods even directly utilize pre-trained models to obtain news representations without tuning any parameters [21], [132], and further transform these representations to be compatible with news recommendation space via customized projector.

B. LLMs for Data Augmentation

Most of the existing methods utilize only titles to model news due to the limited input length of language models. However, since different genres of information (e.g., titles, abstracts, categories, etc.) often contain different aspects of news, all these contents are expected to be covered as much as possible. Recent works have borrowed the strong generation ability of LLMs to summarize some of them into condensed news texts, which are then provided as auxiliary textual features for data augmentation [132], [133], [136], [137]. For example, TF-DCon [136] designs content-level prompts to guide LLMs in condensing various contents of each news into a succinct yet informative title. And they would like to use this data condensation method to relieve the strain of computational resources. SINGLE [132] designs article summarization prompts and instructs LLMs to generate summarized bodies with fewer words for each article. These summarized ones then replace the original bodies to represent news, which effectively evades the problem of long text modeling. Note that these methods make efforts to balance data augmentation with resource-intensive training, which aim at compressing various news contents into small but informative summary texts to reduce textual data load while maximizing the information utilization.

Besides, commonsense knowledge of LLMs can also be infused along with the summarization process to explore implicit real-world relations between entities within news articles, which we believe could bring fresh insights into alleviating long-tail and cold-start problems. For example, ONCE [133] offers titles, abstracts, and categories as inputs of LLMs to produce concise titles with more informative contents, which associates long-tail entities within news texts with the knowledge of LLMs. GNR [137] uses commonsense knowledge of LLMs to summarize relevant themes for each news, and makes them as supplements to original news texts. Yada et al. [138] employ LLMs with extensive knowledge to generate detailed category descriptions for news, and integrate them with original news texts to enhance news modeling.

Regardless of the purpose of introducing LLMs for data augmentation, all the above methods do not tune the parameters of LLMs, and attempt to trigger their few/zero-shot abilities by constructing recommendation-specific prompts. As shown in Fig. 4, either open-source or closed-source LLMs can be used to summarize news inputs into more informative textual features, as they both generate tokens as outputs. The difference is that we can also acquire hidden states of open-source LLMs as news representations.

In addition to processing information that news already has, leveraging LLMs to generate additional textual features that are not included in the original news can also be adopted for data augmentation. For example, in general recommendation, KAR [139] regards LLM as a knowledge provider to generate factual knowledge for items. TagGPT [140] draws support from LLMs to infer concise descriptions of semantic contents as tags from given textual data. Brinkmann et al. [141] explore how LLMs generate attributes from item titles for e-commerce systems. Similarly in news recommendation, LLMs can also be exploited to produce extra features such as background about news, associated figures, and relevant events. But it needs to be noticed that the produced news features should be aligned with specific user preference to avoid generating correct but useless information. This research perspective is still under investigation at present.

V. LLMs for User Modeling

User modeling focuses on inferring personalized user interests from their historical interaction behaviors. Since user interests are often diverse and changeable, LLMs become the dawn of effective modeling over volatile user behaviors and underlying inclinations. Similar to news modeling, we also divided related works on user modeling into two categories, i.e., (1) LLM as feature encoder and (2) LLM for data augmentation.

A. LLMs as Feature Encoder

User interests are usually reflected through their historical interaction sequences in personalized news recommendation, as most users on the news platform are anonymous [142] and only quite few users are willing to provide their personal information such as gender or age. So that in order to encode diverse and ever-changing user interests, recent works have adopted LLMs as user encoders to learn user representations from their click behaviors [19], [129], [132], [143]. For example, UNBERT [19] and PREC [129] implement several Transformer layers of LLMs to capture interactive relations between different clicked news. The final output of user encoders is ordinarily a unified user representation.

B. LLMs for Data Augmentation

As we mentioned before, users in news recommendation usually do not offer detailed information like gender, location, topics, and channels as user profiles to exhibit their precise preferences or personal interests. Thus analyzing and understanding users from their historical click behaviors has become the most prevalent approach at the moment. But most previous studies just acquire a unified user vector via aggregating all browsed news, which weakens the multifaceted and protean nature of user interests. Benefiting from in-context learning, recent works have introduced LLMs to generate some of the textual user features inferred from clicked news to explicitly indicate user interests [132], [133], [136], [137], [143]. For example, SINGLE [132] prompts LLMs to summarize keywords from historical clicked news as user characteristics via elaborate instructions. ONCE [133] employs LLMs to infer topics and regions of interest as user profiles when given users' click histories. It then fuses the extracted user profile into an interest vector to enhance the conventional user vector encoded from click histories. TF-DCon [136] also extracts primary interests of users via LLMs to minimize historical redundancies. Note that we can employ both open-source and closed-source LLMs in this process, and devise specific prompts to instruct the summarization of user features.

In addition, several methods explore to generate synthetic instances via LLMs, which can enrich the training data especially for cold-start users. For example, ONCE [133] generates synthetic news pieces for users with limited interactions to expand their historical behaviors. This makes sense since LLMs are gifted in few-shot learning. And more abundant data enables the better extraction of user features, which allows new users to access more personalized recommendation results that closely resonate with their interests, rather than just popular news.

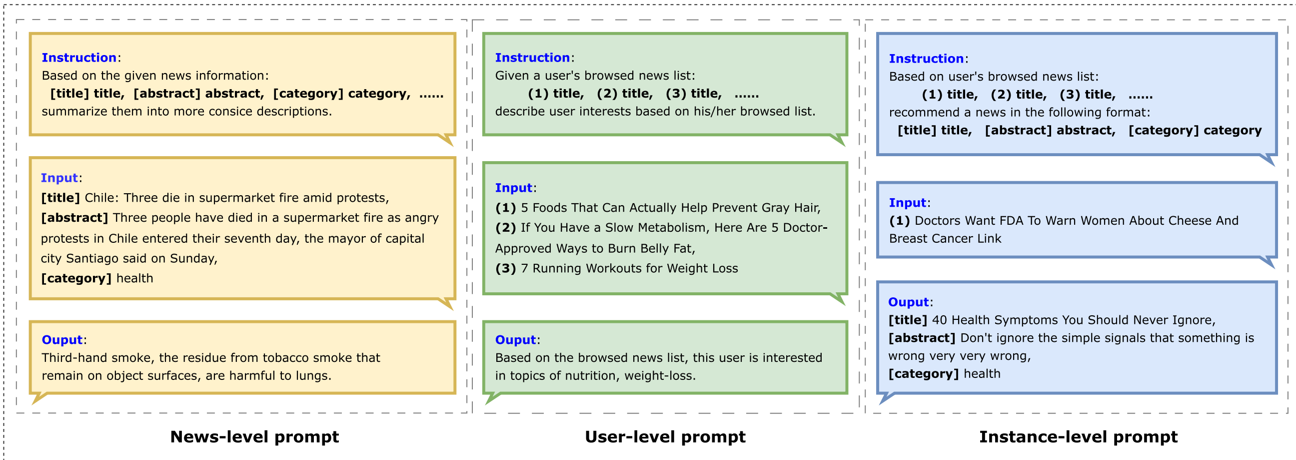

Fig. 5. A demonstration of news-level, user-level and instance-level prompts for data augmentation.

In summary, designing appropriate prompts to instruct LLMs to generate auxiliary textual features from raw input texts is currently the most common way for news/user data augmentation. We demonstrate three kinds of prompts at news-level, user-level, and instance-level as shown in Fig. 5. After constructing prompt templates, each training sample is provided as input of LLMs to generate desired outputs.

VI. LLMs for Accurate Matching

Accurate matching user interests with candidate news lies in the core of personalized news recommendation to help users find their news of interests. In this section, we view matching as a process of selecting or ranking candidate news to satisfy user interests, with the general purpose of providing a ranked list of news for the target user. Based on the different news recommendation tasks that LLMs solve to achieve such a purpose, we classify relevant studies into three categories, i.e., scoring task, ranking task, and generation task.

A. Scoring Task

The scoring task aims to estimate the matching score of each candidate news with the target user for click prediction. Some studies concentrate on measuring the relevance between users and news, since matching scores are typically computed based on their relevance. Traditional approaches often first produce unified news and user representations respectively, and then perform operations like dot product to match them. This is known as the typical two-tower architecture, which relies on the semantic similarity between the final user and news representations. However, considering that doing so may lose fine-grained matching signals, representative work such as AMM [23] develops a multi-field matching framework with LLMs to learn textual matching signals within each pair of users' clicked and candidate news. User-news matching representations of each news pair are finally obtained to calculate the click probability rather than separate news and user vectors. We summarize the differences between this fine-grained matching framework and the two-tower architecture, as shown in Fig. 6. (a) General two-tower architecture matches two unified user and news vectors learned from independent encoders separately. (b) Fine-grained matching framework matches each vector pair of clicked and candidate news for one user with LLMs. Methods with this framework can excavate more fine-grained relatedness between user behaviors and candidate news, thereby offering news that better targets user interests.

Apart from exploiting LLMs only as match encoders for relevance modeling, recent works have completely revolutionized the conventional recommendation paradigm by directly transferring LLMs as recommenders. Pairs of user behaviors and candidate news are input into LLMs in textual formats to output token sequences that are expected to offer the final recommendation results. For example, UniTRec [144] borrows a text-to-text Transformer with encoder-decoder architecture, where user history and candidate news texts are input into the encoder and decoder respectively to predict the user-news matching score. Since match scores should be real numbers rather than discrete tokens generated by LLMs, it abandons the original language modeling decoder head and sends the last hidden states of the decoder into a specifically designed score head to calculate the final score. Compared to this method that customizes specific objectives to adapt LLMs to downstream tasks, a few discriminative methods introduce prompt tuning to reformulate scoring into the pre-training task with prompt templates and the label word verbalizer. For example, PBNR [145] transforms scoring into a text-to-text language task with devised input templates. The matching score for personalized ranking is estimated by the probability of the target template (with a standard of "yes/no") generated from the decoder. Prompt4NR [146] designs a prompt template with a [MASK] token, and converts the scoring task into a cloze-style mask prediction task to produce the probability distribution of answer words (e.g., "yes" and "no") for [MASK]. The probability is then mapped to label words through a verbalizer. Besides, Li et al. [147] and Dai et al. [148] tailor distinct prompts to instruct LLMs to generate numerical strings within a range of 1 to 5, which uncover the scoring ability of LLMs in a generative manner.

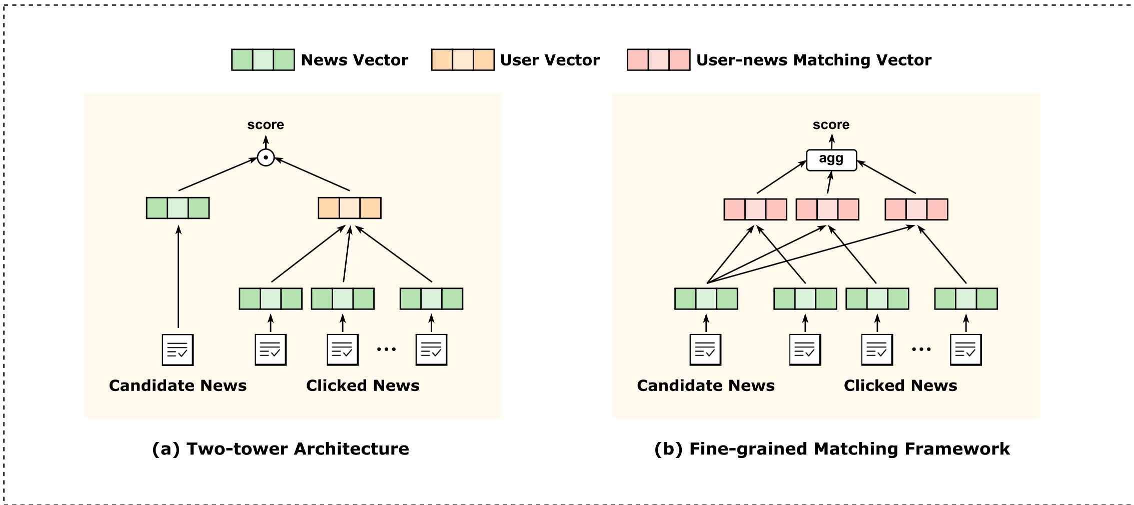

Fig. 6. A Demonstration of the two-tower architecture and fine-grained matching framework. (a) The two-tower architecture matches between unified user and news vectors. (b) The fine-grained matching framework matches each pair of clicked and candidate news vectors for one user.

B. Ranking Task

Personalized ranking before the era of LLMs is conducted based on their match scores with the target user. In contrast, ranking task with LLMs intends to directly generate the ranked news list for the target user, without the need of calculating each match score of candidate news and then sort them to acquire final ranking results. In most cases, a small set of candidate news is first retrieved from the universal news pool, and then wrapped into input prompts designed for LLMs to generate the final ranked list. For example, Li et al. [147] employ prompts tailored for ranking task to ask fine-tuned LLMs to directly sort the given candidate news based on users' recent interactions. RecPrompt [24] incorporates an additional LLM-based prompt optimizer beyond the general LLM recommender, which iteratively improves the prompt template to obtain more effective ranked results for the given candidate news set. Dai et al. [148] explore three ranking approaches via specially designed input prompts of LLMs, namely point-wise ranking (equivalent to scoring) that predicts the matching score between a specific news and the target user, pair-wise ranking that selects one piece of news the target user would prefer from a pair of news, and list-wise ranking that sorts the given list of news. Li et al. [149] and Xu et al. [150] also explore the fairness of ranked news lists generated by LLMs.

C. Generation Task

Compared to the above tasks that score or rank from the provided candidate news set, we classify approaches that generate recommendation results without any candidate news into the type of generation task. LLMs in this situation are unaware of the universal news pool, and could just rely on their significant reasoning abilities to generate news or news lists from user profiles and behavior sequences. But this easily leads to the problem of generative hallucination [27], [151], [152], since the generated news might fail to match an exact one in the news pool. Representative work such as LANCER [153] exploits item mapping operations after generating the next item that a user might have interest in. Specifically, LANCER [153] takes the first new item (the one that never appeared in the user history) in each output sequence of LLMs as the prediction result, and matches its encoded representations with each item in the item pool via cosine similarity. Through these post-processing operations for matching, the newly generated news will concretely match with a certain one in the universal pool, which is available to overcome the generative hallucination problem.

Apart from the above methods that generate the next predicted item in textual format, a few studies also explore to generate item IDs via aligning ID-based recommendation space with natural language space [152], [154], [155]. For example, LightLM [152] customizes a lightweight LLM architecture with collaborative ID indexing, and develops a constrained generation method to construct a Tire[3] structure that contains only valid ID tokens, eliminating the hallucination issue that may create non-existent IDs. But similar approaches have not yet appeared in news recommendation field, possibly because the rapid updates and short lifetime of news make it impractical to deploy unique and efficient IDs.

So far, regardless of whether candidate news is included in prompt templates to instruct LLMs' generation, existing methods can only recommend news in its original form and cannot provide a concise overview of news articles. Considering that users may become exhausted from reading lengthy news articles, sketching several relevant news into a concise narrative provides a new perspective to better save users' reading time and enhance overall comprehension, especially when they are interested in fresh events or new horizons. GNR [137] pioneer the generation of the coherent narrative, which fuses several inter-connected news into a logically coherent narrative centered around one theme. Specifically, it first selects the top-1 news with the highest matching score as the focal news, and filters out a reference news set associated with it from the whole news pool. LLMs are then

[3]https://en.wikipedia.org/wiki/Trie

TABLE I
DATA STATISTICS OF EXISTING PUBLIC DATASETS FOR NEWS RECOMMENDATION

Dataset	#News	#Users	#Clicks	Language
Plista	70,353	14,897,978	1,095,323	German
Adressa	48,486	3,083,438	27,223,576	Norwegian
Globo	46,000	314,000	3,000,000	Portuguese
Yahoo!	14,180	-	34,022	English
MIND	161,013	1,000,000	24,155,470	English
EB-NeRD	125,541	1,103,602	37,966,985	Danish

used to summarize the news set into a multi-news narrative, fusing key facts aligned with user interests of the focal ones.

VII. NEWS RECOMMENDATION DATASETS

Most research on news recommendation is typically conducted on proprietary datasets tailored for specific user cases. Only a few news recommendation datasets are publicly available, whose detailed information is shown in Table I. Here, we present a brief overview of these datasets as follows.

Plista [156] dataset is composed of 70,353 news articles, 14,897,978 users, and 1,095,323 news click records collected from 13 German news portals in June 2013.

Adressa [157] dataset is constructed based on over three months of news logs from Adresseavisen website, a Norwegian news publisher. It comes in two versions, where the large one consists of 3,083,438 users, 48,486 articles, and 27,223,576 clicks within ten weeks, and the small one comprises 561,733 users, 11,207 articles, and 2,286,835 clicks in one week.

Globo [158] dataset is retrieved from Globo.com, a popular news portal in Brazil. It includes about 314,000 users, 46,000 news articles, and 3,000,000 click records. The dataset is in Portuguese, and only provides pre-trained word embeddings of each news without original news text.

Yahoo![4] dataset consists of around 14,180 news articles and 34,022 click events provided by Yahoo! Front Page Today. The dataset is in English, with original news texts not provided. The number of users is also unknown.

MIND [159] dataset is a large-scale English dataset constructed from real behavior logs of 1,000,000 users on Microsoft News website in six weeks. It contains 161,013 news articles, 15,777,377 impressions, and 24,155,470 news clicks. A small version of MIND dataset has also been released, which involves 50,000 users randomly sampled from the original large version, and their 230,117 behavior logs with 93,698 news articles.

EB-NeRD [160] dataset is a large-scale Danish dataset created from user behavior logs of Ekstra Bladet over six weeks. It contains 1,103,602 unique users, 125,541 distinct news articles, and 37,966,985 impression logs.

We report the datasets leveraged by the aforementioned works of PNR utilizing LLMs in Table II. It can be seen that almost all works conduct experiments on the MIND dataset, with some of them also taking Adressa into account. Since most datasets such as Adressa are in low-source languages and studies on them often adopt different data processing approaches, the MIND dataset with given training, validation, and test splits has gradually

[4]https://webscope.sandbox.yahoo.com/

TABLE II
SUMMARY OF WORKS ON PNR UTILIZING LLMS

Method	NM	UM	AM	Datasets
UNBERT [19]	✓	✓		MIND
PLM-NR [20]	✓			MIND
AMM [23]			✓	MIND, Adressa
PREC [130]	✓	✓		MIND, Adressa
MM-Rec [131]	✓			- (private)
Tiny-NewsRec [132]	✓			MIND, Feeds, News
LKPNR [21]	✓			MIND
SINGLE [133]	✓	✓		MIND
TF-DCon [137]	✓	✓		MIND
UniTRec [145]			✓	MIND
Prompt4NR [147]			✓	MIND
PBNR [146]			✓	MIND
LANCER [154]			✓	MIND
RecPrompt [24]			✓	MIND
ONCE [134]	✓	✓		MIND
GLSM [135]	✓			MIND, Adressa
GNR [138]	✓	✓	✓	MIND
EmbSum [144]		✓		MIND
Li et al. [148]			✓	MIND
Dai et al. [149]			✓	MIND
Yada et al. [139]	✓			MIND

The modules where LLMs are employed (columns 2-4) and the datasets leveraged (columns 5) in each work are annotated. Here, NM, UM and AM refer to news modeling, user modeling, and accurate matching, respectively. Only news recommendation datasets are taken into consideration without other types of recommendation dataset used by several works.

become to serve as a standard testbed for news recommendation in the era of LLMs.

VIII. FUTURE PROSPECTS

So far, we have thoroughly reviewed current news recommendation methods utilizing LLMs, and observed their substantial progress in achieving better performance. But there still remain challenges and opportunities in the field of personalized news recommendation. We would like to discuss several potential directions that need to be further explored in this section.

A. Generative Recommendation

News recommendation typically retrieves available news from the universal news pool for personalized recommendations. However, existing news in the universal pool may be insufficient to satisfy users' diverse and personalized preferences, and users are unable to express their information needs explicitly except for passive click behaviors. AI-Generated Content (AIGC) offers the potential to revolutionize the traditional retrieval-based recommendation paradigm into the generative recommendation paradigm [161]. Specifically, generative AI can create totally fresh news content in real-time based on main news events to meet personalized user preferences. Their flexible interaction patterns also allow users to deliver their explicit information needs via natural language instructions. Combined with powerful generative AI, generative news recommendation will be more intelligent and flexible to better meet diverse and personalized user needs.

Despite the gifts of LLMs in generating fluent natural language texts, they could suffer from a challenging phenomenon called hallucination, where the generated contents seem reasonable but in fact conflict with user inputs or existing facts [162]. For instance, LLMs may mistakenly fabricate news events that have not actually occurred when they are asked to recommend something of a user's interests. Multiple reasons can lead to this problem, such as source-reference divergence in the dataset as well as defective training and modeling selection for neural network models [163]. The erroneous information caused by hallucination problem is often unfaithful and nonsensical, which poses a negative impact on users' reading experience. Additionally, if the generated false information is recommended to users and further spread over the internet, more severe consequences will occur in the realistic society and even affect the trend of public opinion. Hence, it is critical to verify the reliability and availability of contents generated by LLMs, and draw support from external resources such as knowledge graphs [164], [165], [166] to supplement factual knowledge when necessary.

B. Privacy

A paramount concern when recommendation systems are discussed is to protect users' personal information from unauthorized access or leakage. This is momentous for offering users of online services with data confidentiality [167], as data privacy is critical for developing more responsible recommendation systems. So far, only a few studies have investigated the preservation of privacy in news recommendation [95], [168], [169], [170]. Among them, Qi et al. [95] developed a federated learning-based framework to learn the privacy-preserving news recommendation model, which stores user data on local devices rather than intensively on a central server. They also employed local differential privacy techniques to reduce the private information contained in model gradients.

Though previous efforts strive to balance the personal data mining and privacy preservation, we argue that the privacy issue of news recommendation in the era of LLMs faces more severe challenges. On the one hand, domain-specific corpora collected from various news platforms are utilized to fine-tune modern LLMs to improve their performance, but some of them may contain sensitive attributes such as user's gender, age and location. As Carlini et al. [171] have shown, LLMs can implicitly memorize such sensitive information within the datasets, which pose the risk of information disclosure. On the other hand, LLMs are commonly used to analyze and mine personalized user preferences, like news reading habits and point of interests, from user historical behaviors. Unless proper protection, LLMs' accessibility to these sensitive user resources may expose news systems to malicious attacks or exploitation, causing the infringement of user privacy. Moreover, since news recommendation is highly temporal sensitive, the real-time data collection and processing of constantly evolving user interests and online behaviors will further exacerbate continuous privacy leaks. As there is currently no work on privacy-preserving news recommendation with LLMs, effective and efficient methods on this privacy issue are still under exploration.

C. Bias and Fairness

Bias and fairness have emerged as notable topics of research interest in machine learning. And with the pursuit of responsible recommendations in recent years, these concepts also attract wide attentions in recommendation systems with multi-stakeholder natures [172], [173]. Researchers have found that LLMs might be misled into producing harmful or toxic content due to potential biases and stereotypes within the training data [150]. This phenomenon can trigger a series of social and ethical concerns, such as exhibiting discriminatory behaviors towards specific user groups [174], [175], [176]. Existing works on news recommendation generally investigate fairness problems from both user-side and provider-side. For user-side fairness, various methods propose to provide unbiased recommendation services for individual users or user groups with different sensitive attributes [177], [178]. For instance, Wu et al. [177] developed an approach based on decomposed adversarial learning to mitigate unfairness news recommendation caused by the bias of sensitive user attributes like gender. For provider-side fairness, several methods are designed to ensure each news provider has an equal recommendation opportunity [149], [179]. For instance, Qi et al. [179] learned provider-fair representations from biased user data to construct a recommendation model that is fair for different news providers. Also in general recommendations, Hou et al. [180] observed the bias of LLMs while ranking, and designed special prompts to alleviate this phenomenon. In sum, works on fairness-aware news recommendation are still in the elementary stage and need further study.

D. Explainability

Making recommendation explainable and reasoning process transparent appears to be the breakthrough point of recommendation systems, as they are vital for boosting user confidence and promoting informed decisions [181]. On one hand, LLMs have demonstrated powerful generative ability in natural language, which enables them to generate reasonable explanations for recommendation results. For instance, Rahdari et al. [182] combined chain-of-thought prompting into their framework to output explanations through intermediate reasoning steps. While on the other hand, leveraging LLMs to generate user-friendly explanations still faces challenges, since the models themselves are perceived as black boxes for the recommendation field. The complex internal mechanisms of LLMs remain unknown, especially when it comes to closed-source LLMs like ChatGPT. If the behaviors of LLMs are unexplainable and uncontrollable, it is difficult to understand exactly how they reach specific explanations, which lacks enough transparency to earn users' trust.

Up to now, only a very few studies have focused on this explainability issue in news recommendation [183], [184]. Based on the two insights above, we claim that there are two possible research directions for developing explainable news recommendation systems with LLMs. One is to elaborate prompt templates or tuning strategies to instruct LLMs in producing contextually and syntactically coherent explanations. This could involve explaining the intricate patterns of user-news interactions, for example, the underlying factors that encourage a user to click on a particular news article. Quality control mechanisms for the generated explanations, such as human inspection, are required if necessary. Another is to seek the internal working mechanism of LLMs themselves. A possible option is to combine external knowledge sources such as knowledge graphs, so that the decision-making process can be aligned with explicit paths on the graphs for explanation. But this direction remains largely unexplored, and needs breakthrough innovations and ideas from the entire community.

E. Safety and Robustness

Most existing research on news recommendation tends to build algorithms under the assumption of a trusted environment [2]. However, there supposes to be numerous threats posed by malicious users and platforms in real-world scenarios. With the pervasive integration of LLMs into news recommendation, concerns about safety and robustness have grown, as they are crucial to ensure the trustworthiness in recommendation results. Studies have shown that LLMs are vulnerable to adversarial perturbations or noises on inputs [185], which significantly limits their utilization in high-security applications. For example, attackers could juggle part of the content or metadata of news articles, such as inserting several deliberate words that transform the originally mild news into a violent or radical one, to prevent systems from making accurate decisions. These tampered articles may even be manipulated to further promote, and cause the systems to misinterpret such fabricated popularity as actual user interests. Given that loopholes towards noisy inputs are often elicited out of malice, ensuring that LLMs can produce stable output even with small changes in the input is crucial for improving the robustness of news recommendation systems.

To boost the safety and robustness, models such as GPT-4 [186] incorporate safety-relevant prompts during the reinforcement learning from human feedback (RLHF). But this seems impractical in reality since it requires a lot of experts for manual labeling. A more likely solution might be to pre-process the prompts tailored for news recommendation tasks before inputting them to LLMs, such as malice detection or input standardization [31]. Moreover, adversarial training can also be employed to improve the robustness of news recommendation systems with LLMs [187]. Unfortunately, despite the vital importance of safety and robustness in developing responsible news recommendation systems, investigations in this area are still quite limited. Unless effective approaches are explored to resist potential threats, such unstable and unreliable news recommendation systems will struggle to gain users' trust and retain them in the long term.

F. Multimodal Recommendation

In addition to texts, many news websites also include multimodal information such as images, audio, and videos. In fact, users might be attracted to click the news not only for its title or abstract, but also for its multimodal information especially images. Incorporating these kinds of information into news modeling could enhance the comprehensive understanding of news profiles, and provide auxiliary features as supplements. For example, Wu et al. [130] developed multimodal news recommendation that integrates visual information besides news texts to learn multimodal news representations. Moreover, in general recommendations, several methods explore to directly generate multimodal information through LLMs. For instance, Geng et al. [188] generated image explanations for recommendation results with the help of vision-language models. Videos and audio generation with LLMs [189], [190] also provide promising research direction for multimodal recommendation. In news recommendation, we believe that the generation of images with highlights might draw users' attention more easily. But what needs to be noted here is that the generated content of LLMs should be aligned with existing facts or inputs, to prevent negative impacts on recommendation systems.

G. Content Moderation

Content moderation plays a critical role in purifying online environments. News platforms often face a huge flow of news articles each day, some of which might be clickbait, toxicity, or misinformation. If recommended, these low-quality and harmful contents could ruin platforms' reputations, reduce users' experience, and even bring negative impacts to society. Previous works mainly rely on human moderators to manually moderate online content. But humans tend to get exhausted quickly due to the high ratio of contents to moderators. Thus, researchers have been looking for solutions to assist moderator decisions. On the one hand, platforms need to be safeguarded against hazardous and toxic content. On the other hand, platforms are encouraged to recommend what they deem good and high-quality content to users. For example, Waterschoot et al. [191] presented a hybrid moderation method that supports moderators to filter news with a rank-based recommender. Other alternatives such as fake news detection and clickbait detection can also be incorporated into news recommendation to adjust recommendation results according to news quality. Besides, LLMs equipped with powerful natural language comprehension abilities have opened up new possibilities for online moderation. For instance, Kolla et al. [192] integrated LLMs into content moderation to identify posts that are against the rules. This issue is still rarely studied in news recommendation, but we believe that strengthening news moderation can avoid recommending negative or harmful content to users, thus creating a healthy and upward network environment.

H. Cold-Start Problem

Personalized news recommendation can become challenging if there lack historical interactions between users and news [193]. And this is the so-called cold-start problem, especially when the system encounters new users or new news. Previous studies have explored to solve this problem since it may lead to low accuracy of recommendation that prevents users from sticking to news platforms. For instance, Alshehri et al. [194] leveraged a framework based on generative adversarial zero-shot learning to address the cold-start problem caused by new users/articles via generating virtual interaction behaviors for them. Recently, LLMs have exhibited their strong capabilities in recommendation tasks even without being fine-tuned on recommendation-specific datasets, which can be attributed to their learning of general-world knowledge during the pre-training stage. For instance, Liu et al. [195] showed that ChatGPT achieves promising results on recommendation tasks that are transformed into natural language tasks with few-shot prompting. In consequence, we can infer that news recommendation based on LLMs could mitigate the cold-start problem with limited or without user-news interactions. Besides, LLMs can also utilized to generate synthetic user interactions or news facts to expand insufficient users/news data for cold-start alleviation.

IX. Conclusion

Personalized news recommendation has played a positive role in alleviating information overload and improving user experiences. With the advancement of AI techniques, large language models which possess remarkable capabilities in natural language understanding and generation, have demonstrated their potential to promote recommendation performance and opened up a new research direction for conventional news recommendation. Though a few recent studies have reviewed the current state of adapting LLMs into general recommendation systems, there still lacks in-depth overview of the integration of LLMs in the specific news recommendation field. So in this survey, we conduct a comprehensive and systematic overview of personalized news recommendation in the era of LLMs, with a focus on three core problems involved in news recommendation: news modeling, user modeling, and accurate matching. Representative works and detailed development paths are discussed under each focus. We hope to provide perceptive insights into extending the frontier progress of this field, and facilitate future research on personalized news recommendation.

References

[1] M. Li and L. Wang, "A survey on personalized news recommendation technology," *IEEE Access*, vol. 7, pp. 145861–145879, 2019.

[2] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized news recommendation: Methods and challenges," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, pp. 24:1–24:50, 2023.

[3] A. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: Scalable online collaborative filtering," in *Proc. Int. Conf. World Wide Web*, 2007, pp. 271–280.

[4] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak, "News personalization using the CF-IDF semantic recommender," in *Proc. Int. Conf. Web Intell. Mining Semantics*, 2011, Art. no. 10.

[5] F. Garcin, K. Zhou, B. Faltings, and V. Schickel, "Personalized news recommendation based on collaborative filtering," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2012, pp. 437–441.

[6] M. Capelle, F. Frasincar, M. Moerland, and F. Hogenboom, "Semantics-based news recommendation," in *Proc. Int. Conf. Web Intell. Mining Semantics*, 2012, pp. 27:1–27:9.

[7] M. Moerland, F. Hogenboom, M. Capelle, and F. Frasincar, "Semantics-based news recommendation with SF-IDF+," in *Proc. Int. Conf. Web Intell. Mining Semantics*, 2013, Art. no. 22.

[8] F. Hogenboom, M. Capelle, M. Moerland, and F. Frasincar, "Bing-SF-IDF+: Semantics-driven news recommendation," in *Proc. Int. Conf. World Wide Web*, 2014, pp. 291–292.

[9] H. Wang, F. Zhang, X. Xie, and M. Guo, "DKN: Deep knowledge-aware network for news recommendation," in *Proc. Int. Conf. World Wide Web*, 2018, pp. 1835–1844.

[10] Q. Zhu, X. Zhou, Z. Song, J. Tan, and L. Guo, "DAN: Deep attention neural network for news recommendation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5973–5980.

[11] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "NPA: Neural news recommendation with personalized attention," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2576–2584.

[12] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, and X. Xie, "Neural news recommendation with long- and short-term user representations," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 336–345.

[13] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, and X. Xie, "Neural news recommendation with multi-head self-attention," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Proces.*, 2019, pp. 6388–6393.

[14] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "Neural news recommendation with attentive multi-view learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3863–3869.

[15] T. Qi, F. Wu, C. Wu, and Y. Huang, "News recommendation with candidate-aware user modeling," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1917–1921.

[16] H. Wang, F. Wu, Z. Liu, and X. Xie, "Fine-grained interest matching for neural news recommendation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 836–845.

[17] T. Qi, F. Wu, C. Wu, and Y. Huang, "Personalized news recommendation with knowledge-aware interactive matching," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 61–70.

[18] J. Li et al., "MINER: Multi-interest matching network for news recommendation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2022, pp. 343–352.

[19] Q. Zhang et al., "UNBERT: User-news matching BERT for news recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 3356–3362.

[20] C. Wu, F. Wu, T. Qi, and Y. Huang, "Empowering news recommendation with pre-trained language models," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1652–1656.

[21] C. Hao et al., "LKPNR: LLM and KG for personalized news recommendation framework," 2023, *arXiv:2308.12028*.

[22] J. Lin et al., "How can recommender systems benefit from large language models: A survey," 2023, *arXiv:2306.05817*.

[23] Q. Zhang, Q. Jia, C. Wang, J. Li, Z. Wang, and X. He, "AMM: Attentive multi-field matching for news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1588–1592.

[24] D. Liu et al., "RecPrompt: A self-tuning prompting framework for news recommendation using large language models," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 3902–3906.

[25] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 159.

[26] J. Huang and K. C. Chang, "Towards reasoning in large language models: A survey," in *Proc. Conf. Assoc. Comput. Linguistics*, 2023, pp. 1049–1065.

[27] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[28] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," 2023, *arXiv:2302.09419*.

[29] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 1413–1430, Apr. 2024.

[30] L. Wu et al., "A survey on large language models for recommendation," *World Wide Web*, vol. 27, no. 5, 2024, Art. no. 60.

[31] Z. Zhao et al., "Recommender systems in the era of large language models (LLMs)," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6889–6907, Nov. 2024.

[32] L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," in *Proc. Joint Int. Conf. Comput. Linguistics Lang. Resour. Eval.*, 2024, pp. 10146–10159.

[33] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: An open architecture for collaborative filtering of netnews," in *Proc. Conf. Comput. Supported Cooperative Work*, 1994, pp. 175–186.

[34] Y. Xiao, P. Ai, C.-H. Hsu, H. Wang, and X. Jiao, "Time-ordered collaborative filtering for news recommendation," *China Commun.*, vol. 12, no. 12, pp. 53–62, 2015.

[35] S. Liu, Y. Dong, and J. Chai, "Research of personalized news recommendation system based on hybrid collaborative filtering algorithm," in *Proc. 2nd IEEE Int. Conf. Comput. Commun.*, 2016, pp. 865–869.

[36] Y. Ji, W. Hong, Y. Shangguan, H. Wang, and J. Ma, "Regularized singular value decomposition in news recommendation system," in *Proc. Int. Conf. Comput. Sci. Educ.*, 2016, pp. 621–626.

[37] K. Han, "Personalized news recommendation and simulation based on improved collaborative filtering algorithm," *Complexity*, vol. 2020, pp. 8834908:1–8834908:12, 2020.

[38] A. Gershman, T. Wolfe, E. Fink, and J. G. Carbonell, "News personalization using support vector machines," Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2011.

[39] F. Hogenboom, M. Capelle, and M. Moerland, "News recommendation using semantics with the Bing-SF-IDF approach," in *Proc. Int. Conf. Conceptual Model.*, 2013, pp. 160–169.

[40] M. Capelle, M. Moerland, F. Hogenboom, F. Frasincar, and D. Vandic, "Bing-SF-IDF+: A hybrid semantics-driven news recommender," in *Proc. 30th Annu. ACM Symp. Appl. Comput.*, 2015, pp. 732–739.

[41] E. de Koning, F. Hogenboom, and F. Frasincar, "News recommendation with CF-IDF+," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, 2018, pp. 170–184.

[42] E. Brocken et al., "Bing-CF-IDF+: A semantics-driven news recommender system," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, 2019, pp. 32–47.

[43] A. H. Parizi and M. Kazemifard, "Emotional news recommender system," in *Proc. 6th Int. Conf. Cogn. Sci.*, 2015, pp. 37–41.

[44] A. H. Parizi, M. Kazemifard, and M. Asghari, "Emonews: An emotional news recommender system," *J. Digit. Inf. Manage.*, vol. 14, no. 6, pp. 392–402, 2016.

[45] N. Babanejad, A. Agrawal, H. Davoudi, A. An, and M. Papagelis, "Leveraging emotion features in news recommendations," in *Proc. 7th Int. Workshop News Recommendation Analytics Conjunction 13th ACM Conf. Recommender Syst.*, 2019, pp. 70–78.

[46] H. Luo, J. Fan, and D. A. Keim, "Personalized news video recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 1001–1002.

[47] H. J. Lee and S. J. Park, "MONERS: A news recommender for the mobile web," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 143–150, 2007.

[48] W. Chu and S. Park, "Personalized recommendation on dynamic content using predictive bilinear models," in *Proc. Int. Conf. World Wide Web*, 2009, pp. 691–700.

[49] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proc. Int. Conf. Intell. User Interfaces*, 2010, pp. 31–40.

[50] M. Kompan and M. Bieliková, "Content-based news recommendation," in *Proc. Int. Conf. Electron. Commerce Web Technol.*, 2010, pp. 61–72.

[51] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 661–670.

[52] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, "SCENE: A scalable two-stage personalized news recommendation system," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 125–134.

[53] L. Li, L. Zheng, and T. Li, "LOGO: A long-short user interest integration in personalized news recommendation," in *Proc. ACM Conf. Recommender Syst.*, 2011, pp. 317–320.

[54] N. Jonnalagedda and S. Gauch, "Personalized news recommendation using Twitter," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, 2013, pp. 21–25.

[55] E. V. Epure, B. Kille, J. E. Ingvaldsen, R. Deneckère, C. Salinesi, and S. Albayrak, "Recommending personalized news in short user sessions," in *Proc. ACM Conf. Recommender Syst.*, 2017, pp. 121–129.

[56] G. Sottocornola, P. Symeonidis, and M. Zanker, "Session-based news recommendations," in *Proc. Int. Conf. World Wide Web*, 2018, pp. 1395–1399.

[57] G. Zheng et al., "DRN: A deep reinforcement learning framework for news recommendation," in *Proc. Int. Conf. World Wide Web*, 2018, pp. 167–176.

[58] A. Darvishy, H. Ibrahim, F. Sidi, and A. Mustapha, "HYPNER: A hybrid approach for personalized news recommendation," *IEEE Access*, vol. 8, pp. 46877–46894, 2020.

[59] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, J. E. Ingvaldsen, G. Nygreen, and E. Berg, "Tailored news in the palm of your hand: A multi-perspective transparent approach to news recommendation," in *Proc. Int. Conf. World Wide Web*, 2013, pp. 305–308.

[60] I. Ilievski and S. Roy, "Personalized news recommendation based on implicit feedback," in *Proc. Int. News Recommender Syst. Workshop Challenge*, 2013, pp. 10–15.

[61] Y. Noh, Y. Oh, and S. Park, "A location-based personalized news recommendation," in *Proc. Int. Conf. Big Data Smart Comput.*, 2014, pp. 99–104.

[62] G. Kazai, I. Yusof, and D. Clarke, "Personalised news and blog recommendations based on user location, Facebook and Twitter user profiling," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 1129–1132.

[63] Y. Liang, B. Loni, and M. A. Larson, "CLEF NewsREEL 2017: Contextual bandit news recommendation," in *Proc. Conf. Labs Eval. Forum*, 2017.

[64] B. Fortuna, C. Fortuna, and D. Mladenic, "Real-time news recommender system," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discov. Databases: Part III*, 2010, pp. 583–586.

[65] N. Jonnalagedda, S. Gauch, K. Labille, and S. Alfarhood, "Incorporating popularity in a personalized news recommender system," *PeerJ Comput. Sci.*, vol. 2, 2016, Art. no. e63.

[66] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 1933–1942.

[67] V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma, "Deep neural architecture for news recommendation," in *Proc. Conf. Labs Eval. Forum*, 2017.

[68] V. Kumar, D. Khattar, S. Gupta, and V. Varma, "Word semantics based 3-D convolutional neural networks for news recommendation," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 761–764.

[69] D. Khattar, V. Kumar, V. Varma, and M. Gupta, "Weave&Rec: A word embedding based 3-D convolutional network for news recommendation," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1855–1858.

[70] H. Zhang, X. Chen, and S. Ma, "Dynamic news recommendation with hierarchical attention network," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 1456–1461.

[71] C. Wu, F. Wu, M. An, Y. Huang, and X. Xie, "Neural news recommendation with topic-aware news representation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 1154–1159.

[72] D. Liu et al., "KRED: Knowledge-aware document representation for news recommendations," in *Proc. ACM Conf. Recommender Syst.*, 2020, pp. 200–209.

[73] H. Sheu and S. Li, "Context-aware graph embedding for session-based news recommendation," in *Proc. ACM Conf. Recommender Syst.*, 2020, pp. 657–662.

[74] Z. Mao, X. Zeng, and K. Wong, "Neural news recommendation with collaborative news encoding and structural user encoding," in *Proc. Findings Assoc. Comput. Linguistics: Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 46–55.

[75] Y. Sun et al., "A hybrid approach to news recommendation based on knowledge graph and long short-term user preferences," in *Proc. IEEE Int. Conf. Serv. Comput.*, 2021, pp. 165–173.

[76] T. Qi, F. Wu, C. Wu, and Y. Huang, "FUM: Fine-grained and fast user modeling for news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1974–1978.

[77] R. Wang, S. Wang, W. Lu, and X. Peng, "News recommendation via multi-interest news sequence modelling," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 7942–7946.

[78] D. Wang, X. Xiong, Y. Li, J. Wang, and Q. Tan, "HCURec: Hierarchical candidate-aware user modeling for news recommendation," *Expert Syst. Appl.*, vol. 229, no. Part A, 2023, Art. no. 120468.

[79] D. Liu et al., "News graph: An enhanced knowledge graph for news recommendation," in *Proc. 2nd Workshop Knowl.-Aware Conversational Recommender Syst. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1–7.

[80] D. H. Tran, Q. Z. Sheng, W. E. Zhang, N. H. Tran, and N. L. D. Khoa, "CupMar: A deep learning model for personalized news recommendation based on contextual user-profile and multi-aspect article representation," *World Wide Web*, vol. 26, no. 2, pp. 713–732, 2023.

[81] Y. Qian et al., "Interaction graph neural network for news recommendation," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2019, pp. 599–614.

[82] L. Hu et al., "Graph neural news recommendation with unsupervised preference disentanglement," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 4255–4264.

[83] L. Hu, C. Li, C. Shi, C. Yang, and C. Shao, "Graph neural news recommendation with long-term and short-term interest modeling," *Inf. Process. Manage.*, vol. 57, no. 2, 2020, Art. no. 102142.

[84] S. Ge, C. Wu, F. Wu, T. Qi, and Y. Huang, "Graph enhanced representation learning for news recommendation," in *Proc. Int. Conf. World Wide Web*, 2020, pp. 2863–2869.

[85] T. Y. S. S. Santosh, A. Saha, and N. Ganguly, "MVL: Multi-view learning for news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1873–1876.

[86] C. Wu, F. Wu, Y. Huang, and X. Xie, "User-as-graph: User modeling with heterogeneous graph pooling for news recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1624–1630.

[87] B. Yang, D. Liu, T. Suzumura, R. Dong, and I. Li, "10024 going beyond local: Global graph-enhanced personalized news recommendations," in *Proc. ACM Conf. Recommender Syst.*, 2023, pp. 24–34.

[88] Z. Mao, J. Li, H. Wang, X. Zeng, and K. Wong, "DIGAT: Modeling news recommendation with dual-graph interaction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6595–6607.

[89] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan, "Beyond clicks: Dwell time for personalization," in *Proc. ACM Conf. Recommender Syst.*, 2014, pp. 113–120.

[90] K. Saranya and G. S. Sadhasivam, "A personalized online news recommendation system," *Int. J. Comput. Appl.*, vol. 57, no. 18, pp. 6–14, 2012.

[91] K. F. Yeung and Y. Yang, "A proactive personalized mobile news recommendation system," in *Proc. Develop. E-Syst. Eng.*, 2010, pp. 207–212.

[92] V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma, "User profiling based deep neural network for temporal news recommendation," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 765–772.

[93] D. Lee, B. Oh, S. Seo, and K. Lee, "News recommendation with topic-enriched knowledge graphs," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 695–704.

[94] T. Qi et al., "HieRec: Hierarchical user interest modeling for personalized news recommendation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5446–5456.

[95] T. Qi, F. Wu, C. Wu, Y. Huang, and X. Xie, "Privacy-preserving news recommendation model learning," in *Proc. Findings Assoc. Comput. Linguistics: Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1423–1432.

[96] S. Shi et al., "WG4Rec: Modeling textual content with word graph for news recommendation," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1651–1660.

[97] K. Park, J. Lee, and J. Choi, "Deep neural networks for news recommendations," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2017, pp. 2255–2258.

[98] L. Zhang, P. Liu, and J. A. Gulla, "A deep joint network for session-based news recommendations with contextual augmentation," in *Proc. 29th Hypertext Social Media*, 2018, pp. 201–209.

[99] L. Meng, C. Shi, S. Hao, and X. Su, "DCAN: Deep co-attention network by modeling user preference and news lifecycle for news recommendation," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2021, pp. 100–114.

[100] Z. Huang et al., "Personal or general? A hybrid strategy with multi-factors for news recommendation," *ACM Trans. Inf. Syst.*, vol. 41, no. 2, 2023, Art. no. 44.

[101] P. Zhang, Z. Dou, and J. Yao, "Learning to select historical news articles for interaction based neural news recommendation," 2021, *arXiv:2110.06459*.

[102] Q. Chu, G. Liu, H. Sun, and C. Zhou, "Next news recommendation via knowledge-aware sequential model," in *Proc. 18th China Nat. Conf. Chin. Comput. Linguistics*, 2019, pp. 221–232.

[103] J. Xun et al., "Why do we click: Visual impression-aware news recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 3881–3890.

[104] C. Wu, F. Wu, T. Qi, and Y. Huang, "Two birds with one stone: Unified model learning for both recall and ranking in news recommendation," in *Proc. Findings Assoc. Comput. Linguistics: ACL*, 2022, pp. 3474–3480.

[105] C. Wu et al., "FeedRec: News feed recommendation with various user feedbacks," in *Proc. Int. Conf. World Wide Web*, 2022, pp. 2088–2097.

[106] Y. Zhang, G. Chen, L. Wang, X. Guo, and L. Ren, "MIRec: Neural news recommendation with multi-interest and popularity-aware modeling," in *Proc. 47th IEEE Annu. Comput. Softw. Appl. Conf.*, 2023, pp. 458–465.

[107] Y. Tian et al., "Joint knowledge pruning and recurrent graph convolution for news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 51–60.

[108] X. Zhang, Q. Yang, and D. Xu, "Combining explicit entity graph with implicit text information for news recommendation," in *Proc. Int. Conf. World Wide Web*, 2021, pp. 412–416.

[109] Z. Ji, M. Wu, J. Liu, and J. E. Armendáriz-Iñigo, "Attention-based graph neural network for news recommendation," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[110] M. Ma, S. Na, H. Wang, C. Chen, and J. Xu, "The graph-based behavior-aware recommendation for interactive news," *Appl. Intell.*, vol. 52, no. 2, pp. 1913–1929, 2022.

[111] J. Lian, F. Zhang, X. Xie, and G. Sun, "Towards better representation learning for personalized news recommendation: A multi-channel deep fusion approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3805–3811.

[112] Z. Wang and X. Fu, "Enhanced semantic matching with topic-aware and fine-grained user modeling for news recommendation," in *Proc. 2nd Int. Conf. Algorithms Data Mining Inf. Technol.*, 2023, pp. 189–195.

[113] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[114] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[115] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Accessed: Feb. 15, 2024. [Online]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[116] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[117] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[118] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[119] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2023.

[120] A. Zeng et al., "GLM-130B: An open bilingual pre-trained model," 2022, *arXiv:2210.02414*.

[121] K. He et al., "A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics," *Inf. Fusion*, vol. 118, 2025, Art. no. 102963.

[122] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *Informatics*, vol. 11, no. 3, 2024, Art. no. 57.

[123] Q. Li et al., "Adapting large language models for education: Foundational capabilities, potentials, and challenges," 2023, *arXiv:2401.08664*.

[124] S. Wang et al., "Large language models for education: A survey and outlook," 2024, *arXiv:2403.18105*.

[125] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *Proc. 4th ACM Int. Conf. AI Finance*, 2023, pp. 374–382.

[126] J. Lee, N. Stevens, S. Caren Han, and M. Song, "A survey of large language models in finance (FinLLMs)," 2024, *arXiv:2402.02315*.

[127] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large language models in law: A survey," *AI Open*, vol. 5, pp. 181–196, 2024.

[128] Z. Zoey Chen et al., "A survey on large language models for critical societal domains: Finance, healthcare, and law," 2024, *arXiv:2405.01769*.

[129] Q. Liu, J. Zhu, Q. Dai, and X. Wu, "Boosting deep CTR prediction with a plug-and-play pre-trainer for news recommendation," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2823–2833.

[130] C. Wu, F. Wu, T. Qi, C. Zhang, Y. Huang, and T. Xu, "MM-Rec: Visiolinguistic model empowered multimodal news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2560–2564.

[131] Y. Yu, F. Wu, C. Wu, J. Yi, and Q. Liu, "Tiny-NewsRec: Effective and efficient PLM-based news recommendation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2022, pp. 5478–5489.

[132] Z. Liu et al., "Modeling user viewing flow using large language models for article recommendation," in *Proc. Int. Conf. World Wide Web*, 2024, pp. 83–92.

[133] Q. Liu, N. Chen, T. Sakai, and X. Wu, "ONCE: Boosting content-based recommendation with both open- and closed-source large language models," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2024, pp. 452–461.

[134] H. Xu, Q. Peng, H. Liu, Y. Sun, and W. Wang, "Group-based personalized news recommendation with long- and short-term fine-grained matching," *ACM Trans. Inf. Syst.*, vol. 42, no. 1, pp. 3:1–3:27, 2024.

[135] 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," 2023, *arXiv:2305.11700*.

[136] J. Wu et al., "TF-DCon: Leveraging large language models (LLMs) to empower training-free dataset condensation for content-based recommendation," 2023, *arXiv:2310.09874*.

[137] S. Gao et al., "Generative news recommendation," in *Proc. Int. Conf. World Wide Web*, 2024, pp. 3444–3453.

[138] Y. Yada and H. Yamana, "News recommendation with category description by a large language model," 2024, *arXiv:2405.13007*.

[139] Y. Xi et al., "Towards open-world recommendation with knowledge augmentation from large language models," in *Proc. ACM Conf. Recommender Syst.*, 2024, pp. 12–22.

[140] C. Li, Y. Ge, J. Mao, D. Li, and Y. Shan, "TagGPT: Large language models are zero-shot multimodal taggers," 2023, *arXiv:2304.03022*.

[141] A. Brinkmann, R. Shraga, R. Chiz Der, and C. Bizer, "Product information extraction using ChatGPT," 2023, *arXiv:2306.14921*.

[142] S. Gong and K. Q. Zhu, "Positive, negative and neutral: Modeling implicit feedback in session-based news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1185–1195.

[143] C. Zhang et al., "EmbSum: Leveraging the summarization capabilities of large language models for content-based recommendations," in *Proc. ACM Conf. Recommender Syst.*, 2024, pp. 1010–1015.

[144] Z. Mao, H. Wang, Y. Du, and K. Wong, "UniTRec: A unified text-to-text transformer and joint contrastive learning framework for text-based recommendation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2023, pp. 1160–1170.

[145] X. Li, Y. Zhang, and E. C. Malthouse, "PBNR: Prompt-based news recommender system," 2023, *arXiv:2304.07862*.

[146] Z. Zhang and B. Wang, "Prompt learning for news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 227–237.

[147] X. Li, Y. Zhang, and E. C. Malthouse, "Exploring fine-tuning ChatGPT for news recommendation," 2023, *arXiv:2311.05850*.

[148] S. Dai et al., "Uncovering ChatGPT's capabilities in recommender systems," in *Proc. ACM Conf. Recommender Syst.*, 2023, pp. 1126–1132.

[149] X. Li, Y. Zhang, and E. C. Malthouse, "A preliminary study of ChatGPT on news recommendation: Personalization, provider fairness, and fake news," in *Proc. Int. Workshop News Recommendation Analytics Conjunction ACM Conf. Recommender Syst.*, 2023.

[150] C. Xu, W. Wang, Y. Li, L. Pang, J. Xu, and T.-S. Chua, "Do LLMs implicitly exhibit user discrimination in recommendation? An empirical study," 2023, *arXiv:2311.07054*.

[151] Y. Chang et al., "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 39:1–39:45, 2024.

[152] K. Mei and Y. Zhang, "LightLM: A lightweight deep and narrow language model for generative recommendation," 2023, *arXiv:2310.17488*.

[153] J. Jiang et al., "Reformulating sequential recommendation: Learning dynamic user interest with content-enriched language modeling," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2024, pp. 353–362.

[154] W. Hua, S. Xu, Y. Ge, and Y. Zhang, "How to index item IDs for recommendation foundation models," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval Asia Pacific Region*, 2023, pp. 195–204.

[155] J. Liao et al., "LLaRA: Large Language-recommendation assistant," 2023, *arXiv:2312.02445*.

[156] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz, "The plista dataset," in *Proc. 2013 Int. News Recommender Syst. Workshop Challenge*, 2013, pp. 16–23.

[157] J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, and X. Su, "The adressa dataset for news recommendation," in *Proc. Int. Conf. Web Intell.*, 2017, pp. 1042–1048.

[158] G. de Souza Pereira Moreira, F. Ferreira, and A. M. da Cunha, "News session-based recommendations using deep neural networks," in *Proc. 3rd Workshop Deep Learn. Recommender Syst.*, 2018, pp. 15–23.

[159] F. Wu et al., "MIND: A large-scale dataset for news recommendation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 3597–3606.

[160] J. Kruse et al., "EB-NeRD a large-scale dataset for news recommendation," in *Proc. Recommender Syst. Challenge*, 2024, pp. 1–11.

[161] W. Wang, X. Lin, F. Feng, X. He, and T.-S. Chua, "Generative recommendation: Towards next-generation recommender paradigm," 2023, *arXiv:2304.03516*.

[162] Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," 2023, *arXiv:2309.01219*.

[163] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 248:1–248:38, 2023.

[164] Q. Guo et al., "A survey on knowledge graph-based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3549–3568, Aug. 2022.

[165] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3091–3110, Jul. 2024.

[166] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024.

[167] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," 2023, *arXiv:2312.02003*.

[168] T. Qi, F. Wu, C. Wu, Y. Huang, and X. Xie, "Uni-FedRec: A unified privacy-preserving news recommendation framework for model training and online serving," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 1438–1448.

[169] J. Yi, F. Wu, C. Wu, R. Liu, G. Sun, and X. Xie, "Efficient-FedRec: Efficient federated learning framework for privacy-preserving news recommendation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2814–2824.

[170] X. Huang, Y. Luo, L. Liu, W. Zhao, and S. Fu, "Randomization is all you need: A privacy-preserving federated learning framework for news recommendation," *Inf. Sci.*, vol. 637, 2023, Art. no. 118943.

[171] N. Carlini et al., "Extracting training data from large language models," in *Proc. USENIX Secur. Symp.*, 2021, pp. 2633–2650.

[172] R. Burke, N. Sonboli, and A. Ordonez-Gauger, "Balanced neighborhoods for multi-sided fairness in recommendation," in *Proc. Conf. Fairness Accountability Transparency*, 2018, pp. 202–214.

[173] Y. Deldjoo, "Understanding Biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency," 2024, *arXiv:2401.10545*.

[174] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5356–5371.

[175] H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang, "Does gender matter? Towards fairness in dialogue systems," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4403–4416.

[176] 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 *Proc. ACM Conf. Recommender Syst.*, 2023, pp. 993–999.

[177] C. Wu, F. Wu, X. Wang, Y. Huang, and X. Xie, "Fairness-aware news recommendation with decomposed adversarial learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 4462–4469.

[178] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, "User-oriented fairness in recommendation," in *Proc. Int. Conf. World Wide Web*, 2021, pp. 624–632.

[179] T. Qi et al., "ProFairRec: Provider fairness-aware news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1164–1173.

[180] Y. Hou et al., "Large language models are zero-shot rankers for recommender systems," in *Proc. Eur. Conf. Inf. Retrieval*, 2024, pp. 364–381.

[181] O. Tal, Y. Liu, J. X. Huang, X. Yu, and B. Aljbawi, "Neural attention frameworks for explainable recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2137–2150, May 2021.

[182] B. Rahdari et al., "Logic-scaffolding: Personalized aspect-instructed recommendation explanation generation using LLMs," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2024, pp. 1078–1081.

[183] Z. Kurt, T. Köllmer, and P. Aichroth, "An explainable knowledge graph-based news recommendation system," in *Proc. Int. Joint Conf. Knowl. Discov. Knowl. Eng. Knowl. Manage.*, 2023, pp. 214–221.

[184] H. Jiang, C. Li, J. Cai, and J. Wang, "RCENR: A reinforced and contrastive heterogeneous network reasoning model for explainable news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 1710–1720.

[185] Z. Zhang et al., "Certified robustness for large language models with self-denoising," 2023, *arXiv:2307.07171*.

[186] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[187] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, and T. Chua, "Adversarial training towards robust multimedia recommender system," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 5, pp. 855–867, May 2020.

[188] S. Geng, Z. Fu, Y. Ge, L. Li, G. de Melo, and Y. Zhang, "Improving personalized explanation generation through visualization," in *Proc. Conf. Assoc. Comput. Linguistics*, 2022, pp. 244–255.

[189] P. K. Rubenstein et al., "AudioPaLM: A large language model that can speak and listen," 2023, *arXiv:2306.12925*.

[190] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "VideoGPT: Video generation using VQ-VAE and transformers," 2021, *arXiv:2104.10157*.

[191] C. Waterschoot and A. van den Bosch, "Hybrid moderation in the newsroom: Recommending featured posts to content moderators," 2023, *arXiv:2307.07317*.

[192] M. Kolla, S. Salunkhe, E. Chandrasekharan, and K. Saha, “LLM-Mod: Can large language models assist content moderation?,” in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 217:1–217:8.

[193] X. Lin et al., “Towards flexible and adaptive neural process for cold-start recommendation,” *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 1815–1828, Apr. 2024.

[194] M. A. Alshehri and X. Zhang, “Generative adversarial zero-shot learning for cold-start news recommendation,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 26–36.

[195] J. Liu, C. Liu, P. Zhou, R. Lv, K. Zhou, and Y. Zhang, “Is ChatGPT a good recommender? A preliminary study,” 2023, *arXiv:2304.10149*.

Jie Li received the PhD degree from the Beijing University of Posts and Telecommunications, in 2009. He is the vice dean with the Center for Data Science and Service, and a distinguished associate researcher with the School of Computer Science, Beijing University of Posts and Telecommunications. His research interests include natural language processing, knowledge graph, and swarm intelligence. He has published some papers in top conferences and journals, such as ACL, AAAI, EMNLP, and *IEEE Transactions on Knowledge and Data Engineering*.

Zeyi Liu received the BE degree from the University of Science and Technology Beijing, in 2023. She is currently working toward the master’s degree in computer science and technology with the Beijing University of Posts and Telecommunications.

Linmei Hu received the PhD degree from Tsinghua University, in 2018. She works as an associate professor with the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include knowledge graph, natural language processing, and multimodal content analysis. She has published some papers in top conferences and journals, such as ACL, AAAI, EMNLP, ACM SIGKDD, MM, *ACM Transactions on Information Systems*, and *IEEE Transactions on Knowledge and Data Engineering*.

Yunbo Rao received the BS degree from Sichuan Normal University, Chengdu, China, in 2003, and the ME and PhD degrees from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2012, respectively. He is an associate professor with the University of Electronic Science and Technology of China. His research interests include artificial intelligence, image segmentation, 3-D reconstruction, video enhancement, and medical image processing. He has published more than 100 research papers in top international journals and conferences such as *Transactions on Internet and Information Systems*, *Transactions on Multimedia*, *Multimedia Tools and Applications*, *World Wide Web*, *Optical Engineering*, *Knowledge-Based System*, *Visual Computer*, ICDE, COMPSAC, etc.

Bo Liu received the PhD degree from the Beijing University of Posts and Telecommunications, in 2007. He is a division general manager with China CSCEC Real Estate Co., Ltd., and an adjunct professor with the Beijing University of Posts and Telecommunications. His research interests include wireless communication, Big Data, and artificial intelligence.

Bo Fang received the PhD degree from the Beijing University of Posts and Telecommunications, in 2005. He is a data analysis expert with the Network Department of China Mobile Communications Corporation. His research interests include data analysis and wireless network management.

Liqiang Nie (Senior Member, IEEE) received the BEng degree from Xi’an Jiaotong University, and the PhD degree from the National University of Singapore (NUS). He is currently the dean with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). His research interests lie primarily in multimedia content analysis and information retrieval. He has co-/authored more than 100 CCF-A papers and five books, with 15 k plus Google Scholar citations. He is an AE of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Science*. Meanwhile, he is the regular area chair or SPC of ACM MM, NeurIPS, IJCAI, and AAAI. He is a member of ICME steering committee. He has received many awards over the past three years, like ACM MM and SIGIR best paper honorable mention in 2019, the AI 2000 most influential scholars 2020, SIGMM rising star in 2020, MIT TR35 China 2020, DAMO Academy Young Fellow in 2020, SIGIR best student paper in 2021, first price of the provincial science and technology progress award in 2021 (rank 1), and provincial youth science and technology award in 2022. Some of his research outputs have been integrated into the products of Alibaba, Kwai, and other listed companies. He is a fellow of the IAPR.