Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems

Peng Liu[*] and **Lemei Zhang**[*] and **Jon Atle Gulla**
Department of Computer Science
Norwegian University of Science and Technology, Norway
{peng.liu, lemei.zhang, jon.atle.gulla}@ntnu.no

Abstract

The emergence of Pre-trained Language Models (PLMs) has achieved tremendous success in the field of Natural Language Processing (NLP) by learning universal representations on large corpora in a self-supervised manner. The pre-trained models and the learned representations can be beneficial to a series of downstream NLP tasks. This training paradigm has recently been adapted to the recommendation domain and is considered a promising approach by both academia and industry. In this paper, we systematically investigate how to extract and transfer knowledge from pre-trained models learned by different PLM-related training paradigms to improve recommendation performance from various perspectives, such as generality, sparsity, efficiency and effectiveness. Specifically, we propose a comprehensive taxonomy to divide existing PLM-based recommender systems w.r.t. their training strategies and objectives. Then, we analyze and summarize the connection between PLM-based training paradigms and different input data types for recommender systems. Finally, we elaborate on open issues and future research directions in this vibrant field.

1 Introduction

As an important part of the online environment, Recommender Systems (RSs) play a key role in discovering users' interests and alleviating information overload in their decision-making process. Recent years have witnessed tremendous success in recommender systems empowered by deep neural architectures and increasingly improved computing infrastructures. However, deep recommendation models are inherently data-hungry with an enormous amount of parameters to learn, which are likely to overfit and fail to generalize well in practice when their training data (i.e., user-item interactions) are insufficient. Such scenarios widely exist in practical RSs when a large number of new users join in but have fewer interactions. Consequently, the data sparsity issue becomes a major performance bottleneck of the current deep recommendation models.

With the thriving of pre-training in NLP (Qiu et al., 2020), many language models have been pre-trained on large-scale unsupervised corpora and then fine-tuned in various downstream supervised tasks to achieve state-of-the-art results, such as GPT (Brown et al., 2020), and BERT (Devlin et al., 2019). One of the advantages of this pre-training and fine-tuning paradigm is that it can extract informative and transferrable knowledge from abundant unlabelled data through self-supervision tasks such as masked LM (Devlin et al., 2019), which will benefit downstream tasks when the labelled data for these tasks is insufficient and avoid training a new model from scratch. A recently proposed paradigm, prompt learning (Liu et al., 2023b), further unifies the use of pre-trained language models (PLMs) on different tasks in a simple yet flexible manner. In general, prompt learning relies on a suite of appropriate prompts, either hard text templates (Brown et al., 2020), or soft continuous embeddings (Qin and Eisner, 2021), to reformulate the downstream tasks as the pre-training task. The advantage of this paradigm lies in two aspects: (1) It bridges the gap between pre-training and downstream objectives, allowing better utilization of the rich knowledge in pre-trained models. This advantage will be multiplied when very little downstream data is available. (2) Only a small set of parameters are needed to tune for prompt engineering, which is more efficient.

Motivated by the remarkable effectiveness of the aforementioned paradigms in solving data sparsity and efficiency issues, adapting language

---

[*]Equal contribution.

modeling paradigms for recommendation is seen as a promising direction in both academia and industry, which has greatly advanced the state-of-the-art in RSs. Although there have been several surveys on pre-training paradigms in the fields of CV (Long et al., 2022), NLP (Liu et al., 2023b), and graph learning (Liu et al., 2023d), only a handful of literature reviews are relevant to RSs. Zeng et al. (2021) summarize some research on the pre-training of recommendation models and discusses knowledge transfer methods between different domains. But this paper only covers a small number of BERT-like works and does not go deep into the training details of pre-trained recommendation models. Yu et al. (2023) give a brief overview of the advances of self-supervised learning in RSs. However, its focus is on a purely self-supervised recommendation setting, which means the supervision signals used to train the model are semi-automatically generated from the raw data itself. Our work does not strictly focus on the self-supervised training strategies but also incorporates the adaptation and exploration of supervised signals and data augmentation techniques in the pre-training, fine-tuning, and prompting process for various recommendation purposes. Furthermore, none of them systematically analyzed the relationship between different data types and training paradigm choices in RSs. To the best of our knowledge, our survey is the first work that presents an up-to-date and comprehensive review of **L**anguage **M**odeling Paradigm Adaptations for **R**ecommender **S**ystems (LMRS).[1] The main contributions of this paper are summerized as follows:

- We survey the current state of PLM-based recommendation from perspectives of training strategy, learning objective and related data types, and provide the first systematic survey, to the best of our knowledge, in this nascent and rapidly developing field.

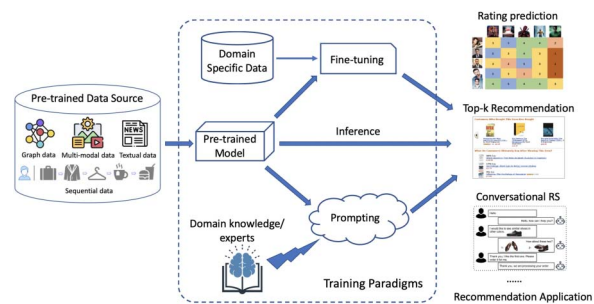- We comprehensively review existing research work on adapting language model-

Figure 1: A generic architecture of language modeling paradigm for recommendation purpose.

ing paradigms to recommendation tasks by systematically categorizing them from two perspectives: pre-training & fine-tuning and prompting. For each category, several sub-categories are provided and explained along with their concepts, formulations, involved methods, and their training and inferencing process for recommendations.

- We shed light on limitations and possible future research directions to help beginners and practitioners interested in this field learn more effectively with the shared integrated resources.

2 Generic Architecture of LMRS

LMRS provides a new way to conquer the data sparsity problem via knowledge transfer from Pre-trained models (PTMs). Figure 1 shows a high-level overview of the LMRS, highlighting the data input, pre-training, fine-tuning/prompting and inference stages for various recommendation tasks. In general, the types of input data objects can be relevant w.r.t. both the training and inference stages. After preprocessing the input into desired forms such as graphs, ordered sequences, or aligned text-image pairs, the training process takes in the preprocessed data and performs either ''*pre-train, fine-tune*'' or ''*pre-train, prompt*'' flow. If the inference is solely based upon the pre-trained model, it can be seen as an end-to-end approach leveraging LM-based learning objectives. The trained model can then be used to infer different recommendation tasks.

3 Data Types

Encoding input data as embeddings is usually the first step in recommendations. However, the

---

[1]It is worth noting that most of the existing literature reviews on pre-trained models focus on the architecture of large-scale language models (such as Bert, T5, UniLMv2, etc.), while our survey mainly discusses training paradigms, which are not limited to pre-trained language model architectures. It can also be other neural networks, such as CNN (Chen et al., 2023), and GCN (Liu et al., 2023c).
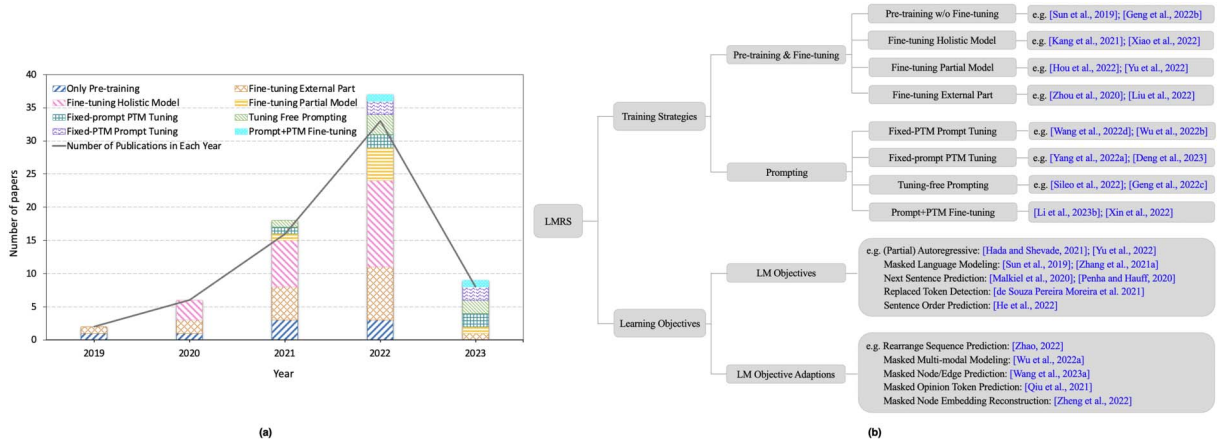
Figure 2: LMRS structure with representatives and statistics on different training strategies and the total number of publications per year.

input for recommender systems is more diverse than most NLP tasks, and therefore encoding techniques and processes may need to be adjusted to align with different input types. **Textual data** as a powerful medium of spreading and transmitting knowledge are commonly used as input for modeling user preferences. Examples of textual data include reviews, comments, summaries, news, conversations, and codes. Note that we also consider item metadata and user profiles as a kind of textual data for simplicity. **Sequential data**, such as user-item interactions strictly arranged chronologically or in a specific order, are used as sequential input for sequential and session-based recommender systems. **Graphs**, usually containing different semantic information from other types of data inputs such as the user-user social graph or heterogeneous knowledge graph, are also commonly used to extract structural knowledge to improve recommendation performance. The diversity of online environments promotes the generation of massive multimedia content, which has been shown to improve recommendation performance in numerous research works. Therefore, **multi-modal data** such as images, videos and audios can also be importance sources for LMRS. Multi-modal data plays a crucial role in recommendation systems. However, the utilization of multi-modal data in LMRS papers is scarce, possibly due to the absence of accessible datasets. A few scholars have gathered their individual datasets to facilitate text-video-audio tri-modal music recommendations (Long et al., 2023) or to establish benchmarks for shopping scenarios (Long et al., 2023).

4   Training Strategies of LMRS

Given the significant impact that PLMs have had on NLP tasks in the pre-train and fine-tune paradigm, there has been a surge recently in adapting such paradigms to multiple recommendation tasks. As illustrated in Figure 1, there are mainly two classes regarding different training paradigms: pre-train, fine-tune paradigm and prompt learning paradigm. Each class is further classified into subclasses regarding different training efforts on different parts of the recommendation model. This section will go through various training strategies w.r.t. specific recommendation purposes. Figure 2(a) presents the statistics of recent publications of LMRSs grouped by different training strategies and the total number of published research works each year. Figure 2(b) shows the taxonomy and some corresponding representative LMRSs.

4.1   Pre-train, Fine-tune Paradigm for RS

The ''*pre-train, fine-tune*'' paradigm attracts increasing attention from researchers in the recommendation field due to several advantages: 1) Pre-training provides a better model initialization, which usually leads to better generalization on different downstream recommendation tasks, improves recommendation performance from various perspectives, and speeds up convergence on the fine-tuning stage; 2) Pre-training on huge source corpus can learn universal knowledge which can be beneficial for the downstream recommenders; 3) Pre-training can be regarded as a kind of regularization to avoid overfitting on

low-resource, and small datasets (Erhan et al., 2010).

Pre-train This training strategy can be seen as traditional end-to-end training with domain input. Differently, we only focus on research adapting LM-based learning objectives into the training phase. Many typical LM-based RSs fall into this category, such as BERT4Rec (Sun et al., 2019), which models sequential user behavior with a bidirectional self-attention network through Cloze task, and Transformers4Rec (de Souza Pereira Moreira et al., 2021) which adopts a Hugging-Face transformer-based architecture as the base model for next-item prediction and explores four different LM tasks, namely, Causal LM, MLM, Permutation LM, and Replacement Token Detection, during training. These two models laid the foundation for LM-based recommender systems and have become popular baselines for their successors.

Pre-train, Fine-tune Holistic Model Under this category, the model is pre-trained and fine-tuned with different data sources, and the fine-tuning process will go through adjusting the whole model parameters. The learning objectives can also vary between the pre-training and fine-tuning stages. Pre-training and fine-tuning with different domains of data sources, also called cross-domain recommendation, can refer to the works of Kang et al. (2021) and Qiu et al. (2021). Kang et al. (2021) pre-trained a GPT model using segmented source API code and fine-tuned it with API code snippets from another library for cross-library recommendation. Wang et al. (2022a) fine-tuned the pre-trained DialoGPT model on domain-specific datasets for conversational recommendation together with an R-GCN model to inject knowledge from DBpedia to enhance recommendation performance. Xiao et al. (2022) fine-tuned the PTM to learn news embedding together with a user embedding part in an auto-regressive manner for news recommendation. They also explored different fine-tuning strategies like tuning part of the PTM and tuning the last layer of the PTM but empirically found fine-tuning the whole model resulted in better performance, which gives us an insight into balancing the recommendation accuracy and training efficiency.

Pre-train, Fine-tune Partial Model Since fine-tuning the whole model is usually time-consuming

and less flexible, many LMRSs choose to fine-tune partial parameters of the model to achieve a balance between training overhead and recommendation performance (Hou et al., 2022; Yu et al., 2022; Wu et al., 2022a). For instance, to deal with the domain bias problem that BERT induces a non-smooth anisotropic semantic space for general texts resulting in a large language gap for texts from different domains of items, Hou et al. (2022) applied a linear transformation layer to transform BERT representations of items from different domains followed by an adaptive combination strategy to derive a universal item representation. Meanwhile, considering the seesaw phenomenon that learning from multiple domain-specific behavioural patterns can be a conflict, they proposed sequence-item and sequence-sequence contrastive tasks for multi-task learning during the pre-training stage. They found only fine-tuning a small proportion of model parameters could quickly adapt the model to unseen domains with cold-start or new items.

Pre-train, Fine-tune Extra Part of the Model With the increase in the depth of PTMs, the representation they capture makes the downstream recommendation easier. Apart from the aforementioned two fine-tuning strategies, some works leverage a task-specific layer on top of the PTMs for recommendation tasks. Fine-tuning only goes through such extra parts of the PTMs by optimizing the parameters of the task-specific layer. Shang et al. (2019) pre-trained a GPT and a BERT model to learn patient visit embeddings, which were then used as input to fine-tune the extra prediction layer for medication recommendation. Another approach is to use the PTM to initialize a new model with a similar architecture in the fine-tuning stage, and the fine-tuned model is used for recommendations. In Zhou et al. (2020), a bidirectional transformer-based model was first pre-trained on four different self-supervised learning objectives (associated attribute prediction, masked item prediction, masked attribute prediction and segment prediction) to learn item embeddings. Then, the learned model parameters were adopted to initialize a unidirectional transformer-based model for fine-tuning with pairwise rank loss for recommendation. In McKee et al. (2023), the authors leveraged the pre-trained BLOOM-176B to generate natural languages descriptions of music given a set of music tags.

1556

Subsequently, two distinct pre-trained models, namely, CLIP and the D2T pipeline, were employed to initialize textual, video, and audio representations of the provided music content. Following this, a transformer-based architecture model was fine-tuned for multi-modal music recommendation.

4.2 Prompting Paradigm for RSs

Instead of adapting PLMs to different downstream recommendation tasks by designing specific objective functions, a rising trend in recent years is to use the ''*pre-train, prompt, and inference*'' paradigm to reformulate downstream recommendations through hard/soft prompts. In this paradigm, fine-tuning can be avoided, and the pre-trained model itself can be directly employed to predict item ratings, generate top-k item ranking lists, make conversations, recommend similar libraries for programmers while coding, or even output subtasks related to recommendation targets such as explanations (Li et al., 2023b). Prompt learning breaks through the problem of data constraints and bridges the gap of objective forms between pre-training and fine-tuning.

Fixed-PTM Prompt Tuning Prompt-tuning only requires tuning a small set of parameters for the prompts and labels, which is especially efficient for few-shot recommendation tasks. Despite the promising results achieved through constructing prompt information without significantly changing the structure and parameters of PTMs, it also calls for the necessity of choosing the most appropriate prompt template and verbalizer, which can greatly impact recommendation performance. Prompt tuning can be both in the form of discrete textual templates (Penha and Hauff, 2020), which are more human-readable, and soft continuous vectors (Wang et al., 2022d; Wu et al., 2022b). For instance, Penha and Hauff (2020) manually designed several prompt templates to test the performance of movie/book recommendations on a pre-trained BERT model with a similarity measure. Wu et al. (2022b) proposed a personalized prompt generator tuned to generate a soft prompt as a prefix before the user behaviour sequence for sequential recommendation.

Fixed-prompt PTM Tuning Fixed-prompt PTM tuning tunes the parameters of PTMs sim-

ilarly to the ''pre-train, fine-tune'' strategy but additionally uses prompts with fixed parameters to steer the recommendation task. One natural way is to use artificially designed discrete prompt to specify recommendation items. For instance, Zhang et al. (2021b) designed a prompt, ''A user watched item A, item B, and item C. Now the user may want to watch () '' to reformulate the recommendation as a multi-token cloze task during fine-tuning of the LM-based PTM. The prompts can also be one or several tokens/words to seamlessly shift/lead the conversations from various tasks. Deng et al. (2023) concatenate input sequences with special designed prompts, such as *[goal]*, *[topic]*, *[item]*, and *[system]*, to indicate different tasks: goal planning, topic prediction, item recommendation, and response generation in conversations. The model is trained using a multi-task learning scheme, and the parameters of the PTM are optimized with the same objective. Yang et al. (2022a) designed a *[REC]* token as a prompt to indicate the start of the recommendation process and to summarize the dialogue context for the conversational recommendation.

Tuning-free Prompting This training strategy can be referred to as *zero-shot recommendations*, which directly generate recommendations and/or related subtasks without changing the parameters of the PTMs but based only on the input prompts. Zero-shot recommendation has been shown to be effective in dealing with new users/ items in one domain or cross-domain settings (Sileo et al., 2022; Geng et al., 2022c), compared to state-of-the-art baselines. Specifically, Geng et al. (2022c) learned multiple tasks, such as sequential recommendation, rating prediction, explanation generation, review summarization and direct recommendation, in a unified way with the same Negative Log-likelihood (NLL) training objectives during pre-training. At the inference stage, a series of carefully designed discrete textual template prompts were taken as input, including prompts for recommending items in the new domain (not appearing in the pre-training phase), and the trained model outputs the preferable results without a fine-tuning stage. The reason for the effectiveness of zero-shot recommendation is that the training data and pre-training tasks are able to distil rich knowledge of semantics and correlations from diverse modalities into user and

item tokens, which can comprehend user preference behaviours w.r.t. item characteristics (Geng et al., 2022c). Building upon this research, Geng et al. (2023) extended their efforts to train an adapter for diverse multimodal assignments, including sequential recommendations, direct recommendations, and the generation of explanations. In particular, they utilized the pre-trained CLIP component to convert images into image tokens. These tokens were added to the textual tokens of an item to create a personalized multimodal soft prompt. This combined prompt was then used as input to fine-tune the adapter in an autoregressive manner.

Prompt+PTM Tuning In this setting, the parameters include two parts: prompt-relevant parameters and model parameters. The tuning phase involves optimizing all parameters for specific recommendation tasks. Prompt+PTM tuning differs from the ''pre-train, fine-tune the holistic model'' strategy by providing additional prompts that can provide additional bootstrapping at the start of model training. For example, Li et al. (2023b) proposed a continuous prompt learning approach by first fixing the PTM, tuning the prompt to bridge the gap between the continuous prompts and the loaded PTM, and then fine-tuning both the prompt and PTM, resulting in a higher BLUE score in empirical results. They combined both discrete prompts (three user/item feature keywords, such as gym, breakfast, and Wi-Fi) and soft prompts (user/item embeddings) to generate recommendation explanations. Case studies showed improvements in the readability and fluency of generated explanations using the proposed prompts. Note that the Prompt+PTM tuning stage does not necessarily mean the fine-tuning stage but can be any possible stage for tuning parameters from both sides for specific data input. Xin et al. (2022) adapted a reinforcement learning framework as a Prompt+PTM tuning strategy by learning reward-state pairs as soft prompt encodings w.r.t. observed actions during training. At the inference stage, the trained prompt generator can directly generate soft prompt embeddings for the recommendation model to generate actions (items).

5 Learning Objectives of LMRS

This section will overview several typical learning tasks and objectives of language models and their adaptations for different recommendation tasks.

5.1 Language Modeling Objectives to Recommendation

The expensive manual efforts required for annotated datasets have led many language learning objectives to adopt self-supervised labels, converting them to classic probabilistic density estimation problems. Among language modeling objectives, autoregressive, reconstruction, and auxiliary are three categories commonly used (Liu et al., 2023b). Here, we only introduce several language modeling objectives used for RSs.

Partial/ Auto-regressive Modeling (P/AM) Given a text sequence $\mathbf{X}_{1:T} = [x_1, x_2, \cdots x_T]$, the training objective of AM can be summarized as a joint negative log-likelihood of each variable given all previous variables:

$$\mathcal{L}_{AM} = -\sum_{t=1}^{T} log\, p(x_t|\mathbf{X}_{<t-1}) \qquad (1)$$

Modern LMRS typically utilize popular pre-trained left-to-right LMs such as GPT-2 (Hada and Shevade, 2021) and DialoGPT (Wang et al., 2022a,d) as the backbone for explainable and conversational recommendations, respectively, to avoid the laborious task of pre-training from scratch. While auto-regressive objectives can effectively model context dependency, the modeling context can only be accessed from one direction, primarily left-to-right. To address this limitation, PAM is introduced, which extends AM by enabling the factorization step to be a span. For each input $\mathbf{X}$, one factorization order M is sampled. One popular PTM that includes PAM as an objective is UniLMv2 (Bao et al., 2020). The pre-trained UniLMv2 model can be utilized to initialize the news embedding model for news recommendation (Yu et al., 2022).

Besides directly leveraging PTMs trained on textual inputs, some researchers apply this objective to train inputs with sequential patterns, such as graphs (Geng et al., 2022b) and user-item interactions (Zheng et al., 2022). These patterns serve as either scoring functions to select suitable paths from the start node/user to the end node/item or detectors to explore novel user-item pairs.

1558

Masked Language Modeling (MLM) Taking a sequence of textual sentences as input, MLM first masks a token or multi-tokens with a special token such as $[MASK]$. Then the model is trained to predict the masked tokens taking the rest of the tokens as context. The objective is as follows:

$$\mathcal{L}_{MLM} = - \sum_{\hat{x} \in M(\mathbf{X})} log\, p(\hat{x}|\mathbf{X}_{M(\mathbf{X})}) \quad (2)$$

where $M(\mathbf{X})$ and $\mathbf{X}_{M(\mathbf{X})}$ represent the masked tokens in the input sequence $\mathbf{X}$ and the rest of the tokens in $\mathbf{X}$, respectively. A typical example of MLM training strategy can be found on BERT, which is leveraged as backbone in Zhang et al. (2021a) to capture user-news matching signals for news recommendation.

Concurrently, some research works propose multiple enhanced versions of MLM. RoBERTa (Liu et al., 2019) improves BERT by dynamic masking instead of in a static manner and can be used to initiate word embedding for conversations (Wang et al., 2022d) and news articles (Wu et al., 2021) for different recommendation scenarios.

Next Sentence Prediction (NSP) It is a binary classification loss for predicting whether two segments follow each other in the original text. The training can be performed in a self-supervised way by taking positive examples from consecutive sentences from the input text corpus and creating negative examples by pairing segments from different documents. A general loss of the NSP is as follows:

$$\mathcal{L}_{NSP} = - log\, p(c|\mathbf{x}, \mathbf{y}) \quad (3)$$

where $\mathbf{x}$ and $\mathbf{y}$ represent two segments from the input corpus, and $c = 1$ if $\mathbf{x}$ and $\mathbf{y}$ are consecutive, otherwise $c = 0$. The NSP objective involves reasoning about the relationships between pairs of sentences and can be utilized for better representation learning of textual items such as news articles, item descriptions, and conversational data for recommendation purposes. Moreover, it can be employed to model the intimate relationships between two components. Malkiel et al. (2020) used the NSP to capture the relationship between the title and description of an item for next-item prediction. Furthermore, models pre-trained with NSP (such as BERT) can be leveraged for probing

the learned knowledge with prompts, which are then infused in the fine-tuning stage to improve model training on adversarial data for conversational recommendation (Penha and Hauff, 2020). Sentence Order Prediction (SOP) as a variation of the NSP takes two consecutive segments from the same document as positive examples, which are then swapped in order as negative examples. SOP has been used to learn the inner coherence of title, description, and code for tag recommendation on StackOverflow (He et al., 2022).

Nevertheless, some researchers have questioned the necessity and effectiveness of NSP and SOP for downstream tasks (He et al., 2022), which highlights the need for further investigation in recommendation scenarios.

Replaced Token Detection (RTD) It is used to predict whether a token is replaced given its surrounding context:

$$\mathcal{L}_{RTD} = - \sum_{t=1}^{T} log\, p(y_t|\hat{\mathbf{X}}) \quad (4)$$

where $y_t = \mathbf{1}(\hat{x}_t = x_t)$, and $\hat{\mathbf{X}}$ is corrupted from the input sequence $\mathbf{X}$. de Souza Pereira Moreira et al. (2021) trained a transformer-based model with RTD objective for session-based recommendations, which achieved the best performance among MLM and AM objectives. This is probably because RTD takes the whole user-item interaction sequence as input and model the context from the bidirectional way.

5.2 Adaptive Objectives to Recommendation

Numerous pre-training or fine-tuning objectives draw inspiration from LM objectives and have been effectively applied to specific downstream tasks based on the input data types and recommendation goals. In sequential recommendations, there is a common interest in modeling an ordered input sequence in an auto-regressive manner from left to right.

Analogous to text sentences, Zheng et al. (2022) and Xiao et al. (2022) treated the user's clicked news history as input text and proposed to model user behavior in an auto-regressive manner for next-click prediction. However, as the sequential dependency may not always hold strictly in terms of user preference for recommendations (Yuan et al., 2020a), MLM objectives can be modified accordingly. Yuan et al. (2020b) randomly

masked a certain percentage of historical user records and predicted the masked items during training. Auto-regressive learning tasks can also be adapted to other types of data. Geng et al. (2022b) modeled a series of paths sampled from a knowledge graph in an auto-regressive manner for recommendation by generating the end node from the pre-trained model. Zhao (2022) proposed pre-training the Rearrange Sequence Prediction task to learn the sequence-level information of the user's entire interaction history by predicting whether the user interaction history had been rearranged, which is similar to Permuted Language Modeling (PerLM) (Yang et al., 2019).

MLM, also known as Cloze Prediction, can be adapted to learn graph representations for different recommendation purposes. Wang et al. (2023a) proposed pre-training a transformer model on a reconstructed subgraph from a user-item-attribute heterogeneous graph, using Masked Node Prediction (MNP), Masked Edge Prediction (MEP), and meta-path type prediction as objectives. Specifically, MNP was performed by randomly masking a proportion of nodes in a heterogeneous subgraph and then predicting the masked nodes based on the remaining contexts by maximizing the distance between the masked node and the irrelevant node. Similarly, MEP was used to recover the masked edge of two adjacent nodes based on the surrounding context. Apart from that, MLM can also be adapted to multi-modal data called Masked Multi-modal Modeling (MMM) (Wu et al., 2022a). MMM was performed by predicting the semantics of masked news and news image regions given the unmasked inputs and indicating whether a news image and news content segment correspond to each other for news recommendation purposes.

The NSP/SOP can be adapted for CTR prediction as Next K Behaviors Prediction (NBP). NBP was proposed to learn user representations in the pre-training stage by inferring whether a candidate behavior is the next i-th behavior of the target user based on their past N behaviors. NBP can also capture the relatedness between past and multiple future behaviors.

6   Formulating Training with Data Types

To associate training strategy, learning objectives with different input data types, we summarize representative works in this domain in Table 1.

The listed training strategies and objectives are carefully selected and are typical in existing work. For the page limit, we only selected part of recent research on LMRS. For more research progress and related resources, please refer to `https://github.com/SmartmediaAI/LMRS`.

Considering that datasets are another important factor for empirical analysis of LMRS approaches, in Table 2, we also list several representative publicly available datasets taking into account the popularity of data usage and the diversity of data types, as well as their corresponding recommendation tasks, training strategies, and adopted data types. From Table 2, we draw several observations: First, datasets can be converted into different data types, which can then be analyzed from various perspectives to enhance downstream recommendations. The integration of different data types can also serve different recommendation goals more effectively (Geng et al., 2022c; Liu et al., 2021). For instance, Liu et al. (2021) transformed user-item interactions and multimodal item side information into a homogeneous item graph. A sampling approach was introduced to select and prioritize neighboring nodes around a central node. This process effectively translated the graph data structure into a sequential format. The subsequent training employed a self-supervised signal within a transformer framework, utilizing an objective for reconstructing masked node features. The resultant pre-trained node embeddings could be readily applied for recommendation purposes, or alternatively, fine-tuned to cater to specific downstream objectives. Second, some training strategies can be applied to multiple downstream tasks by fine-tuning a few parameters from the pre-trained model, adding an extra component, or using different prompts. Geng et al. (2022c) designed different prompt templates for five different tasks to train a transformer-based model with a single objective, and achieved improvements on multiple tasks with zero-shot prompting. Deng et al. (2023) unified the multiple goals of conversational recommenders into a single sequence-to-sequence task with textual input, and designed various prompts to shift among different tasks. We further observe that prompting methods are primarily used in LMRS with textual and sequential data types, but there has been a lack of exploration for multi-modal or graph data. This suggests that investigating additional data

Training Strategy	Paper	Learning Objective	Recommendation Task	Data Type	Source Code
Pre-training & Fine-tuning					
Pre-training w/o Fine-tuning	(Sun et al., 2019)	Pre-train: MLM	Sequential RS	Sequential data	Link
	(Geng et al., 2022b)	Pre-train: AM	Explainable RS	Graph	N/A
	(de Souza Pereira Moreira et al., 2021)	Pre-train: AM + MLM + PerLM + RTD	Session-based RS	Textual + Sequential data	Link
Fine-tuning Holistic Model	(Kang et al., 2021)	Pre-train: cross-entropy / Fine-tune: cross-entropy	Cross-library API RS	Textual data (code)	Link
	(Wang et al., 2022a)	Pre-train: AM / Fine-tune: AM + cross-entropy	Conversational RS	Textual data + Graph	Link
	(Xiao et al., 2022)	Pre-train: AM + MLM / Fine-tune: Negative Sampling Loss	News RS	Textual + Sequential data	Link
	(Zhang et al., 2023)	Pre-train: MLM + NT-Xent / Fine-tune: Negative Sampling Loss	Social RS	Textual data	Link
	(Wang et al., 2023a)	Pre-train: MNP + MEP + cross-entropy + Contrastive Loss; Fine-tune: cross-entropy	Top-N RS	Graph	N/A
Fine-tuning Partial Model	(Hou et al., 2022)	Pre-train: Contrastive Loss / Fine-tune: cross-entropy	Cross-domain RS Sequential RS	Textual + Sequential data	Link
	(Yu et al., 2022)	Pre-train: MLM + AM / Fine-tune: cross-entropy + MSE + InfoNCE	News RS	Textual + Sequential data	Link
	(Wu et al., 2022a)	Pre-train: MMM + MAP / Fine-tune: cross-entropy	News RS	Sequential + Multi-modal data	Link
Fine-tuning External Part	(Zhou et al., 2020)	Pre-train: MIM / Fine-tune: Pairwise Ranking Loss	Sequential RS	Textual + Sequential data	Link
	(Liu et al., 2022)	Pre-train: MTP + cross-entropy / Fine-tune: cross-entropy	News RS	Textual + Sequential data	Link
	(Shang et al., 2019)	Pre-train: binary cross-entropy / Fine-tune: cross-entropy	Medication RS	Graph	Link
	(Liu et al., 2023c)	Pre-train: binary cross-entropy / Fine-tune: BPR + binary cross-entropy	Top-N RS	Textual data + Graph	Link
Prompting					
Fixed-PTM Prompt Tuning	(Wang et al., 2022d)	Pre-train: AM + MLM + cross-entropy / Prompt-tuning: AM + cross-entropy	Conversational RS	Textual data	Link
	(Wu et al., 2022b)	Pre-train: Pairwise Ranking Loss / Prompt-tuning: Pairwise Ranking Loss + Contrastive Loss	Cross-domain RS Sequential RS	Textual + Sequential data	N/A
Fixed-prompt PTM Tuning	(Yang et al., 2022a)	Pre-train: AM + MLM / PTM Fine-tune: AM + cross-entropy	Conversational RS	Textual data	Link
	(Deng et al., 2023)	Pre-train: AM; PTM Fine-tune: AM	Conversational RS	Textual data	Link
Tuning-free Prompting	(Sileo et al., 2022)	Pre-train: AM	Zero-Shot RS	Textual data	Link
	(Geng et al., 2022c)	Pre-train: AM	Zero-Shot RS Cross-domain RS	Textual + Sequential data	Link
Prompt+PTM Tuning	(Li et al., 2023b)	Pre-train: AM; Prompt-tuning: NLL / Prompt+PTM tuning: NLL + MSE	Explainable RS	Textual data	Link
	(Xin et al., 2022)	Prompt+PTM tuning: cross-entropy	Next Item RS	Sequential data	N/A

Note. NT-Xent: Normalized Temperature-scaled Cross Entropy Loss; MMM: Masked Multi-modal Modeling; MAP: Multi-modal Alignment Prediction; MIM: Mutual Information Maximization Loss; MTP: Masked News/User Token Prediction; NLL: Negative Log-likelihood Loss.

Table 1: A list of representative LMRS methods with open-source code.

types may be a future direction for research in prompting-based LMRS.

7 Evaluation

7.1 Evaluation Metrics

As an essential aspect of recommendation design, evaluation can provide insights on recommendation quality from multiple dimensions. Apart from well-known metrics such as RMSE, MAP, AUC, MAE, Recall, Precision, MRR, NDCG, F1-score, and HitRate in offline mode, some works define Group AUC (Zhang et al., 2022) or User Group AUC (Zheng et al., 2022) to evaluate the utility of group recommendations. JIANG et al. (2022)

and Liu et al. (2022) conducted A/B testing to evaluate performance with online users using Conversion rate or CTR.

The integration of generative modules such as GPT and T5 into existing recommender systems offers additional possibilities for recommender systems, such as generating free-form textual explanations for recommendation results or simulating more realistic real-life dialogue scenarios during conversational recommendations to enhance users' experience. In such cases, BLEU and ROUGE are commonly adopted to automatically evaluate the relevance of generated text based on lexicon overlap. Additionally, Perplexity (PPL), Distinct-n, and Unique Sentence Ratio

Dataset	Data Source	Recommendation Task	Training Strategy	Data Type
MovieLens	Link	Rating Prediction	Tuning-free Prompting (Gao et al., 2023)	Textual data (Zhang et al., 2021b; Sileo et al., 2022; Penha and Hauff, 2020; Xie et al., 2023; Gao et al., 2023); Sequential data (Yuan et al., 2020a; Liu et al., 2021; Zhao, 2022) Graph (Liu et al., 2023c, 2021; Wang et al., 2023a); Multi-modal data (Liu et al., 2021)
		Explainable RS	Fine-tuning Holistic Model (Xie et al., 2023)	
		Sequential RS	Pre-training w/o Fine-tuning (Yuan et al., 2020a), Fine-tuning Holistic Model (Zhao, 2022)	
		Conversational RS	Fine-tuning Holistic Model (Penha and Hauff, 2020), Tuning-free Prompting (Gao et al., 2023)	
		Top-N RS	Fine-tuning Holistic Model (Wang et al., 2023a), Fine-tuning External Part (Liu et al., 2023c), Fixed-prompt PTM Tuning (Zhang et al., 2021b), Tuning-free Prompting (Zhang et al., 2021b; Sileo et al., 2022)	
		CTR Prediction	Fine-tuning External Part (Liu et al., 2021)	
Amazon Review Data	Link	Rating Prediction	Fine-tuning External Part (Hada and Shevade, 2021), Tuning-free Prompting (Geng et al., 2022c)	Textual data (Hada and Shevade, 2021; Qiu et al., 2021; Li et al., 2023b; Geng et al., 2022c; Zhou et al., 2020; Penha and Hauff, 2020; Xie et al., 2023; Zhao, 2022; Hou et al., 2023; Li et al., 2023a); Sequential data (Sun et al., 2019; Geng et al., 2022c; Zhou et al., 2020; Geng et al., 2022b; Liu et al., 2021; Hou et al., 2023; Guo et al., 2023); Graph (Geng et al., 2022b; Liu et al., 2021); Multi-modal data (Liu et al., 2021)
		Cross-domain RS	Fine-tuning Holistic Model (Qiu et al., 2021), Fine-tuning Partial Model (Hou et al., 2023), Fixed-PTM Prompt Tuning (Guo et al., 2023)	
		Explainable RS	Pre-training w/o Fine-tuning (Geng et al., 2022b), Fine-tuning Holistic Model (Xie et al., 2023), Fixed-PTM Prompt Tuning (Li et al., 2023b), Fixed-prompt PTM Tuning (Li et al., 2023a), Tuning-free Prompting (Geng et al., 2022c)	
		Zero-Shot RS	Tuning-free Prompting (Geng et al., 2022c)	
		Sequential RS	Pre-training w/o Fine-tuning (Sun et al., 2019), Fine-tuning Holistic Model (Zhao, 2022), Fine-tuning Partial Model (Hou et al., 2023), Fine-tuning External Part (Zhou et al., 2020), Fixed-PTM Prompt Tuning (Guo et al., 2023), Tuning-free Prompting (Geng et al., 2022c)	
		Conversational RS	Fine-tuning Holistic Model (Penha and Hauff, 2020)	
		Top-N RS	Fine-tuning External Part (Liu et al., 2021)	
Yelp	Link	Rating Prediction	Fine-tuning Holistic Model (Xie et al., 2023), Fine-tuning External Part (Hada and Shevade, 2021; Geng et al., 2022a), Tuning-free Prompting (Geng et al., 2022c)	Textual data (Hada and Shevade, 2021; Qiu et al., 2021; Li et al., 2023b; Geng et al., 2022c; Xiao et al., 2021; Zhou et al., 2020; Xie et al., 2023); Sequential data (Geng et al., 2022c; Xiao et al., 2021; Zhou et al., 2020; Sankar et al., 2021); Graph Xiao et al., 2021; Zheng et al., 2022; Wang et al., 2023a); Multi-modal data (Geng et al., 2022a)
		Cross-domain RS	Fine-tuning Holistic Model (Qiu et al., 2021)	
		Explainable RS	Fine-tuning Holistic Model (Xie et al., 2023), Fine-tuning External Part (Geng et al., 2022a), Fixed-PTM Prompt Tuning (Li et al., 2023b), Tuning-free Prompting (Geng et al., 2022c)	
		Zero-Shot RS	Tuning-free Prompting (Geng et al., 2022c)	
		Sequential RS	Fine-tuning Holistic Model (Xiao et al., 2021), Fine-tuning External Part (Zhou et al., 2020), Tuning-free Prompting (Geng et al., 2022c)	
		Top-N RS	Pre-training w/o Fine-tuning (Zheng et al., 2022), Fine-tuning Holistic Model (Wang et al., 2023a), Fine-tuning External Part (Sankar et al., 2021)	
TripAdvisor	Link	Rating Prediction	Fine-tuning Holistic Model (Xie et al., 2023), Fine-tuning External Part (Geng et al., 2022a)	Textual data (Li et al., 2023b; Xie et al., 2023); Multi-modal data (Geng et al., 2022a)
		Explainable RS	Fine-tuning Holistic Model (Xie et al., 2023), Fine-tuning External Part (Geng et al., 2022a), Fixed-PTM Prompt Tuning (Li et al., 2023b)	
MIND	Link	Top-N RS	Fine-tuning Holistic Model (Xiao et al., 2022), Fine-tuning Partial Mode (Yu et al., 2022), Fine-tuning External Part (Yu et al., 2022), Fixed-prompt PTM Tuning (Zhang and Wang, 2023)	Textual data (Xiao et al., 2022; Yu et al., 2022; Zhang and Wang, 2023); Sequential data (Xiao et al., 2022; Yu et al., 2022)
ReDial	Link	Conversational RS	Fine-tuning Holistic Model (Li et al., 2022), Fixed-PTM Prompt Tuning (Wang et al., 2022d), Fixed-prompt PTM Tuning (Yang et al., 2022a)	Textual data (Wang et al., 2022d; Yang et al., 2022a; Li et al., 2022; Graph (Li et al., 2022)
Polyvore Outfits	Link	Fashion RS	Fine-tuning Partial Model + External Part (Sarkar et al., 2022)	Multi-modal data (Sarkar et al., 2022)
MIMIC-III	Link	Medication RS	Fine-tuning External Part (Shang et al., 2019)	Graph (Shang et al., 2019)
Stackoverflow	Link	Top-N RS	Fine-tuning Holistic Mode (He et al., 2022)	Textual data (He et al., 2022)
Online Retail	Link	Cross-domain RS	Fine-tuning Partial Model (Hou et al., 2022)	Textual + Sequential data (Hou et al., 2022)

Table 2: A list of commonly used and publicly accessible real-world datasets for LMRS.

(USR) are also widely used metrics to measure fluency, diversity, and informativeness of generated texts. Other evaluation metrics are leveraged with respect to special requests in LMRSs. For instance, Xie et al. (2023) adopted Entailment Ratio and MAUVE to measure if the generated explanations are factually correct and how close the generated contents are to the ground truth corpus, respectively. Geng et al. (2022a) adopted Feature Diversity (DIV) and CLIPScore (CS) to measure the generated explanations and text-image alignment. Besides, to assess the system's capability to provide item recommendations during conversations, Wang et al. (2022a) computed the Item Ratio within the final generated responses. They evaluated the recommendation
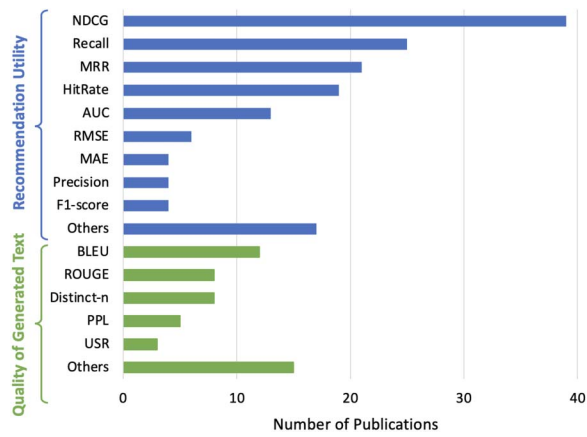
Figure 3: The statistics of evaluation metrics on recommendation utility and generated text quality in LMRS.

performance in an end-to-end manner to prevent the inappropriate insertion of recommended items into dialogues.

Human evaluation complements objective evaluation, as automatic metrics may not match subjective feedback from users. Liu et al. (2023a) pointed out that human subjective and automatic objective evaluation measurements may yield opposite results, which underscores the limitations of existing automatic metrics for evaluating generated explanations and dialogues in LMRSs. Figure 3 displays usage frequency statistics for different evaluation metrics in their respective tasks.

7.2 Discussion on Evaluation Across Datasets

In this section, we compare the results obtained from various models using commonly used datasets. Specifically, based on the reported results in the paper, we measured the improvement achieved by different models compared to a shared baseline and evaluated them using the same metrics on the same dataset. The comparisons are presented in Table 3–6. Most improvements are highlighted in bold, and N@k denotes NDCG@k, H@k denotes HitRate@k. It's important to recognize that a comprehensive and precise assessment cannot be achieved without a carefully designed platform and thoughtful settings for conducting the experiments. Various factors, such as diverse training platforms, parameter settings, and data split strategy, can lead to fluctuations in the results. Hence, it is essential to consider the analysis solely for reference purposes. From the tables, we can ob-

serve that: First, among the four conversational recommender systems assessed using the ReDial dataset, *fixed prompt PTM tuning* paradigm Yang et al. (2022a) demonstrate the most significant improvements compared to the shared baselines. Second, on the Amazon dataset, *zero-shot* and *few-shot learning* of ChatGPT underperformed the supervised recommendation baselines (Liu et al., 2023a). This could be due to language models' strength in capturing language patterns rather than effectively collaborating to suggest similar items based on user preferences (Zhang et al., 2021b). Additionally, Liu et al. (2023a) pointed out that the position of candidate items in the item pool can also affect the direct recommendation performance. Another *prompting-based* model, P5, showed the most improvements for both Amazon and Yelp datasets (Geng et al., 2022c), which verifies the need for more guidance when using large pre-trained language models for recommendations. Finally, for news recommendation on the MIND dataset, Xiao et al. (2022) introduced a model-agnostic *fine-tuning* framework with cache management, which can accelerate the model training process and yield the most improvements over the baselines.

8 Discussion and Future Directions

Despite the effectiveness of LM training paradigms has been verified in various recommendation tasks, there are still several challenges that could be the future research directions.

Language Bias and Fact-consistency in Language Generation Tasks of Recommendation. While generating free-form responses of conversational recommender systems or explanations of the recommended results, the generative components of existing LMRSs tend to predict generic tokens to ensure sentences fluency or repeat certain universally applicable "safe" sentences (e.g., "the hotel is very nice" generated from PETER [Li et al., 2021]). Therefore, one future research direction is to enhance the diversity and pertinence of generated explanations and replies while maintaining language fluency, rather than resorting to "Tai Chi" responses. Additionally, generating factually consistent sentences is also an urgent research problem that needs to be addressed but has not received sufficient attention (Xie et al., 2023).

Metrics	Fine-tune Holistic Model						Fixed-PTM Prompt Tuning			Fixed-prompt PTM Tuning		
	(Wang et al., 2022b)			(Li et al., 2022)			(Wang et al., 2022d)			(Wang et al., 2022c)		
	ReDial	KBRD	KGSF	ReDial	KBRD	KGSF	ReDial	KBRD	KGSF	ReDial	KBRD	KGSF
Recall@1	**1.458**	**0.903**	**0.513**	–	–	–	1.217	0.545	0.457	1.333	0.860	0.436
Recall@10	0.174	0.6	0.311	0.307	0.219	0.115	0.736	0.28	0.266	**0.829**	**0.707**	**0.399**
Recall@50	0.291	0.229	0.093	0.268	0.154	0.043	**0.439**	0.248	0.128	0.422	**0.354**	**0.204**
Distinct-2	1.031	0.738	0.581	0.541	0.149	0.159	1.187	0.751	0.629	**2.653**	**2.125**	**1.844**
Distinct-3	1.767	0.774	0.505	1.408	0.492	0.204	1.746	0.71	0.497	**3.881**	**2.13**	**1.654**
Distinct-4	2.338	0.799	0.466	1.524	0.7	0.225	2.649	0.9	0.597	**4.759**	**2.104**	**1.530**

Table 3: LMRSs performance comparison using common benchmarks on the ReDial dataset.

Training Strategy	Paper	Caser				GRU4Rec				SASRec			
		H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
Pre-train	(Sun et al., 2019)	0.3582	0.5229	0.168	0.3691	0.2392	0.3643	0.1398	0.2815	0.1412	0.1135	0.1402	0.1402
Fine-tune Extra Part	(Zhou et al., 2020)	0.4848	0.5354	0.4857	0.3968	0.4406	0.5022	0.341	0.4443	0.2034	0.1963	**0.1725**	0.1825
Tuning-free Prompt	(Geng et al., 2022c)	**1.478**	**1.8931**	**0.9135**	**1.4375**	**2.0976**	**2.8283**	**1.3463**	**2.1314**	**0.3127**	**0.5221**	0.0975	**0.3491**
	(Liu et al., 2023a)	−0.3415	0.0305	−0.611	−0.233	–	–	–	–	−0.6512	−0.4578	−0.7769	−0.5755

Table 4: LMRSs performance comparison using common benchmarks on the Amazon Beauty dataset.

Training Strategy	Paper	Caser				SASRec				BERT4Rec				GRU4Rec			
		H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
Fine-tune Holistic Model	(Xiao et al., 2021)	0.2097	0.1953	0.2078	0.1966	0.2581	0.2380	0.2811	0.2533	0.0666	0.087	0.617	0.081	0.3022	0.3961	0.2.26	0.3153
Fine-tune Extra Part	(Zhou et al., 2020)	0.1906	0.178	0.1597	0.1753	0.0592	0.07	0.0477	0.0629	0.0182	0.035	0.0168	0.0326	0.1192	0.1631	0.0633	0.1278
Tuning-free Prompt	(Geng et al., 2022c)	**2.8013**	**3.1979**	**1.7945**	**2.4651**	**2.5215**	**3.03**	**1.5803**	**2.2868**	**10.2549**	**11.2121**	**6.8556**	**8.9333**	**2.7763**	**3.0707**	**1.6882**	**2.3358**

Table 5: LMRSs performance comparison using common benchmarks on the Yelp dataset.

Training Strategy	Paper	NAML				NPA				LSTUR				NRMS			
		AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10
Fine-tune	(Zhang et al., 2021a)	0.0635	0.0895	0.0973	0.0816	0.0722	0.1126	0.127	0.1092	0.0537	0.1026	0.1132	0.0941	0.0446	0.0731	0.0786	0.0667
Holistic Model	(Xiao et al., 2022)	**0.0913**	**0.1784**	**0.1974**	**0.1713**	**0.1343**	**0.2855**	**0.32**	**0.2793**	**0.1456**	**0.3018**	**0.3448**	**0.2906**	**0.0746**	**0.1612**	**0.1825**	**0.1575**
Fine-tune Partial/	(Wu et al., 2021)	0.0401	0.0608	0.0666	0.0553	0.039	0.063	0.0654	0.0538	0.037	0.0594	0.0659	0.0525	0.0361	0.0631	0.0661	0.0517
Extra Part	(Shin et al., 2023)	–	–	–	–	0.0772	0.1416	0.1557	0.1231	0.0572	0.1131	0.1281	0.1041	0.0611	0.1066	0.1222	0.094

Table 6: LMRSs performance comparison using common benchmarks on the MIND dataset.

Knowledge Transmission and Injection for Downstream Recommendations. Improper training strategies may cause varying degrees of problems when transferring knowledge from pre-trained models. Zhang et al. (2022) have pointed out the catastrophic forgetting problem in continuously-trained industrial recommender systems. The degree of domain knowledge pre-trained models possess and the effective ways to transfer and inject it for recommendation purposes are both open questions. For example, Zhang et al. (2021b) experimented with a simple approach to injecting knowledge through domain-adaptive pre-training, resulting in only limited improvements. Furthermore, questions about maximizing knowledge transfer to different recommendation tasks, quantifying the degree of transferred knowledge, and whether an upper bound for knowledge transfer exists are all valuable issues that need to be studied and explored in the AI community.

Scalability of Pre-training Mechanism in Recommendation. As model parameters growing larger and larger, the knowledge stored in them is also increasing. Despite the great success of pre-trained models in multiple recommendation tasks, how to maintain and update such complex and large-scale models without affecting the efficiency and accuracy of recommendations in reality needs more attention. Some works have proposed improving model updating efficiency by fine-tuning a partial pre-trained model or an extra part with far fewer parameters than the model's magnitude. However, Yuan et al. (2020b) empirically found that fine-tuning only the output layer often resulted in poor performance in recommendation scenarios. While properly fine-tuning the last few layers sometimes offered promising performance, the improvements were quite unstable and depended on the pre-trained model and tasks. Yu et al. (2022) proposed compressing large pre-trained language models into student

models to improve recommendation efficiency, while Yang et al. (2022b) focused on accelerating the fine-tuning of pre-trained language models and reducing GPU memory footprint for news recommendation by accumulating the gradients of redundant item encodings. Despite all these achievements, efforts are still needed in this rapidly developing field.

Balancing Multiple Objectives in Pre-training. Much research uses multi-task learning objectives to better apply the knowledge learned in the pre-training phase to downstream tasks (Geng et al., 2022c; Wang et al., 2023a). The primary objective of multi-task learning for recommendation is to enhance recommendation accuracy and/or other related aspects by promoting interactions among related tasks. The learning optimization process requires trade-offs among different objectives. For instance, Wang et al. (2023b) fine-tuned parameters to optimize and balance the overarching goals of topic-level recommendation, semantic-level recommendation, and a specific aspect of topic learning. Similarly in Wang et al. (2022c), the authors employed a parameter that required learning to achieve a balance between conversation generation objective and quotation recommendation objective. Yang et al. (2022a) proposed a conversational recommendation framework that contain a generation module and a recommendation module. The overall objectives were designed to balance these two modules with a parameter learned through a fine-tuning process. However, improper optimization can lead to other problems, as pointed out by Deng et al. (2023) that ''Error Propagation'' may occur when solving multiple tasks in sequential order, leading to a decrease in performance with the sequential completion of each task. Although some potential solutions to this issue (Deng et al., 2023; Li et al., 2022; Geng et al., 2022a) were suggested, further verification is still needed.

Multiple Choices of PLM as Recommendation Bases. With the advances in variational PLMs, including ChatGPT, and their success in various downstream tasks, researchers have started exploring the potential of ChatGPT in conversational recommendation tasks. For example, Liu et al. (2023a) and Gao et al. (2023) have investigated the ability of GPT-3/GPT-3.5-based ChatGPT in zero-shot scenarios, using human-designed prompts to assess its performance in rating prediction, sequential recommendation, direct recommendation, and explanation generation. However, these studies are just initial explorations, and more extensive research is required on different recommendation tasks based on various pre-trained language models. This includes prompt design and performance evaluation in diverse domains. Moreover, recent LMRS studies have yet to explore instruction tuning, which could be a promising direction for future research.

Privacy Issue. The study conducted by Yuan et al. (2020b) revealed that pre-trained models can infer user profiles (such as gender, age, and marital status) based on learned user representations, which raises concerns about privacy protection. The pre-training process is often performed on large-scale web-crawled corpus without fine-grained filtering, which may expose users' sensitive information. Therefore, developing LMRS that strike a balance between privacy and high-performance recommendation algorithms remains an open issue.

Acknowledgments

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 642–652. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter,

Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lei Chen, Fajie Yuan, Jiaxi Yang, Xiangnan He, Chengming Li, and Min Yang. 2023. User-specific adaptive fine-tuning for cross-domain recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):3239–3252. `https://doi.org/10.1109/TKDE.2021.3119619`

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems*, 41(3):1–25. `https://doi.org/10.1145/3570640`

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N19-1423`

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208, Chia Laguna Resort, Sardinia, Italy. PMLR.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards interactive and explainable LLMs-augmented recommender system. *arXiv preprint arXiv:2303.14524v2*.

Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022a. Improving personalized explanation generation through visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 244–255. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.20`

Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard De Melo, and Yongfeng Zhang. 2022b. Path language modeling over knowledge graphs for explainable recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 946–955. Association for Computing Machinery. `https://doi.org/10.1145/3485447.3511937`

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022c. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315. Association for Computing Machinery. `https://doi.org/10.1145/3523227.3546767`

Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. VIP5: Towards multimodal foundation models for recommendation. *arXiv preprint arXiv:2305.14302v1*.

Lei Guo, Chunxiao Wang, Xinhua Wang, Lei Zhu, and Hongzhi Yin. 2023. Automated prompting for non-overlapping cross-domain sequential recommendation. *arXiv preprint arXiv:2304.04218v1*.

Deepesh V. Hada and Shirish K. Shevade. 2021. ReXPlug: Explainable recommendation using plug-and-play language model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–91. Association for Computing Machinery. `https://doi.org/10.1145/3404835.3462939`

Junda He, Bowen Xu, Zhou Yang, DongGyun Han, Chengran Yang, and David Lo. 2022. PTM4Tag: Sharpening tag recommendation of stack overflow posts with pre-trained models. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, pages 1–11. Association for Computing Machinery. `https://doi.org/10.1145/3524610.3527897`

Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation

for transferable sequential recommenders, pages 1162–1171. Association for Computing Machinery. https://doi.org/10.1145/3543507.3583434

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 585–593. Association for Computing Machinery. https://doi.org/10.1145/3534678.3539381

Caigao Jiang, Siqiao Xue, James Y. Zhang, Lingyue Liu, Zhibo Zhu, and Hongyan Hao. 2022. Learning large-scale universal user representation with sparse mixture of experts. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

Yuning Kang, Zan Wang, Hongyu Zhang, Junjie Chen, and Hanmo You. 2021. APIRecX: Cross-library API recommendation via pre-trained language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3425–3436. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.275

Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gérard Medioni. 2023a. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. In *SIGIR 2023 Workshop on eCommerce*.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.383

Lei Li, Yongfeng Zhang, and Li Chen. 2023b. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26. https://doi.org/10.1145/3580488

Shuokai Li, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Zhenwei Tang, Wayne Xin Zhao, and Qing He. 2022. Self-supervised learning for conversational recommendation. *Information Processing & Management*, 59(6):103067. https://doi.org/10.1016/j.ipm.2022.103067

Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023a. Is ChatGPT a good recommender? A preliminary study. *arXiv preprint arXiv:2304.10149v2*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35. https://doi.org/10.1145/3560815

Qijiong Liu, Jieming Zhu, Quanyu Dai, and Xiaoming Wu. 2022. Boosting deep CTR prediction with a plug-and-play pre-trainer for news recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2823–2833. International Committee on Computational Linguistics.

Siwei Liu, Zaiqiao Meng, Craig Macdonald, and Iadh Ounis. 2023c. Graph neural pre-training for recommendation with side information. *ACM Transactions on Information Systems*, 41(3):1–28. https://doi.org/10.1145/3568953

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692v1*.

Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip Yu. 2023d. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5879–5900.

Yong Liu, Susen Yang, Chenyi Lei, Guoxin Wang, Haihong Tang, Juyong Zhang, Aixin Sun, and Chunyan Miao. 2021. Pre-training graph transformer with multimodal side information for recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2853–2861. Association for

Computing Machinery. `https://doi.org/10.1145/3474085.3475709`

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537. International Joint Conferences on Artificial Intelligence Organization. `https://doi.org/10.24963/ijcai.2022/773`

Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li, and Xiaojie Wang. 2023. Multimodal recommendation dialog with subjective preference: A new challenge and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3515–3533, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.findings-acl.217`

Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Razin, Ori Katz, and Noam Koenigstein. 2020. RecoBERT: A catalog language model for text-based recommendations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1704–1714. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.findings-emnlp.154`

Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. 2023. Language-guided music recommendation for video via prompt analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14784–14793. IEEE Computer Society. `https://doi.org/10.1109/CVPR52729.2023.01420`

Gustavo Penha and Claudia Hauff. 2020. What does BERT know about books, movies and music? Probing BERT for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 388–397. Association for Computing Machinery. `https://doi.org/10.1145/3383313.3412249`

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.naacl-main.410`

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897. `https://doi.org/10.1007/s11431-020-1647-3`

Zhaopeng Qiu, Xian Wu, Jingyue Gao, and Wei Fan. 2021. U-BERT: Pre-training user representations for improved recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4320–4327. `https://doi.org/10.1609/aaai.v35i5.16557`

Aravind Sankar, Junting Wang, Adit Krishnan, and Hari Sundaram. 2021. ProtoCF: Prototypical collaborative filtering for few-shot recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 166–175. Association for Computing Machinery. `https://doi.org/10.1145/3460231.3474268`

Rohan Sarkar, Navaneeth Bodla, Mariya Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. 2022. OutfitTransformer: Outfit representations for fashion recommendation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2262–2266. `https://doi.org/10.1109/CVPRW56347.2022.00249`

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5953–5959. International Joint Conferences on Artificial Intelligence. `https://doi.org/10.24963/ijcai.2019/825`

Kyuyong Shin, Hanock Kwak, Su Young Kim, Max Nihlén Ramström, Jisu Jeong, Jung-Woo Ha, and Kyung-Min Kim. 2023. Scaling law for recommendation models: Towards

1568

general-purpose user representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4596–4604. https://doi.org/10.1609/aaai.v37i4.25582

Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-shot recommendation as language modeling. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 223–230. Springer International Publishing. https://doi.org/10.1007/978-3-030-99739-7_26

Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4Rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 143–153. Association for Computing Machinery. https://doi.org/10.1145/3460231.3474255

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450. Association for Computing Machinery. https://doi.org/10.1145/3357384.3357895

Hui Wang, Kun Zhou, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023a. Curriculum pre-training heterogeneous subgraph transformer for top-n recommendation. *ACM Transactions on Information Systems*, 41(1):1–28. https://doi.org/10.1145/3528667

Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. 2022a. RecInDial: A unified framework for conversational recommendation with pretrained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 489–500. Association for Computational Linguistics.

Lingzhi Wang, Shafiq Joty, Wei Gao, Xingshan Zeng, and Kam-Fai Wong. 2022b. Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge. *arXiv preprint arXiv:2209.11386v1*.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2022c. Learning when and what to quote: A quotation recommender system with mutual promotion of recommendation and generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3094–3105. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.225

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2023b. Quotation recommendation for multi-party online conversations based on semantic and topic fusion. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3594633

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022d. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937. Association for Computing Machinery. https://doi.org/10.1145/3534678.3539382

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1652–1656. Association for Computing Machinery. https://doi.org/10.1145/3404835.3463069

Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022a. MM-Rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2560–2564. Association for Computing Machinery.

Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Xu Zhang, Leyu Lin, and Qing He. 2022b. Personalized prompts for

sequential recommendation. *arXiv preprint arXiv:2205.09666v2*.

Chaojun Xiao, Ruobing Xie, Yuan Yao, Zhiyuan Liu, Maosong Sun, Xu Zhang, and Leyu Lin. 2021. UPRec: User-aware pre-training for recommender systems. *arXiv preprint arXiv:2102.10989v1*.

Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, Bhuvan Middha, Fangzhao Wu, and Xing Xie. 2022. Training large-scale news recommenders with pretrained language models in the loop. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4215–4225. Association for Computing Machinery. `https://doi.org/10.1145/3534678.3539120`

Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and informative review generation for explainable recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13816–13824. `https://doi.org/10.1609/aaai.v37i11.26618`

Xin Xin, Tiago Pimentel, Alexandros Karatzoglou, Pengjie Ren, Konstantina Christakopoulou, and Zhaochun Ren. 2022. Rethinking reinforcement learning for recommendation: A prompt perspective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1347–1357. Association for Computing Machinery. `https://doi.org/10.1145/3477495.3531714`

Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022a. Improving conversational recommendation systems' quality with context-aware item meta-information. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 38–48. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.findings-naacl.4`

Yoonseok Yang, Kyu Seok Kim, Minsam Kim, and Juneyoung Park. 2022b. GRAM: Fast fine-tuning of pre-trained language models for content-based collaborative filtering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–851. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.61`

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Curran Associates Inc.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. 2023. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20. `https://doi.org/10.1109/TKDE.2023.3282907`

Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, and Qi Liu. 2022. Tiny-NewsRec: Effective and efficient PLM-based news recommendation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5478–5489. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.368`

Fajie Yuan, Xiangnan He, Haochuan Jiang, Guibing Guo, Jian Xiong, Zhezhao Xu, and Yilin Xiong. 2020a. Future data helps training: Modeling future contexts for session-based recommendation. In *Proceedings of The Web Conference 2020*, pages 303–313. Association for Computing Machinery. `https://doi.org/10.1145/3366423.3380116`

Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020b. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1469–1478. Association for Computing Machinery. `https://doi.org/10.1145/3397271.3401156`

Zheni Zeng, Chaojun Xiao, Yuan Yao, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2021. Knowledge transfer via pre-training for recommendation: A review and prospect. *Frontiers in big Data*, 4. `https://doi.org/10.3389/fdata.2021.602071`, PubMed: 33817631

Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021a. UNBERT: User-news matching bert for news recommendation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3356–3362. International Joint Conferences on Artificial Intelligence Organization. Main Track. `https://doi.org/10.24963/ijcai.2021/462`

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607. Association for Computing Machinery. `https://doi.org/10.1145/3580305.3599921`

Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021b. Language models as recommender systems: Evaluations and limitations. In *NeurIPS 2021 Workshop on I (Still) Can't Believe It's Not Better*.

Yujing Zhang, Zhangming Chan, Shuhao Xu, Weijie Bian, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. KEEP: An industrial pre-training framework for online recommendation via knowledge extraction and plugging. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3684–3693. Association for Computing Machinery. `https://doi.org/10.1145/3511808.3557106`

Zizhuo Zhang and Bang Wang. 2023. Prompt learning for news recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 227–237. Association for Computing Machinery. `https://doi.org/10.1145/3539618.3591752`

Qihang Zhao. 2022. RESETBERT4Rec: A pre-training model integrating time and user historical behavior for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1812–1816. Association for Computing Machinery. `https://doi.org/10.1145/3477495.3532054`

Jiayi Zheng, Ling Yang, Heyuan Wang, Cheng Yang, Yinghong Li, Xiaowei Hu, and Shenda Hong. 2022. Spatial autoregressive coding for graph neural recommendation. *arXiv preprint arXiv:2205.09489v2*.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1893–1902. Association for Computing Machinery. `https://doi.org/10.1145/3340531.3411954`