IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 December 2023, accepted 11 February 2024, date of publication 20 February 2024, date of current version 29 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368027

RESEARCH ARTICLE

Empowering Few-Shot Recommender Systems With Large Language Models-Enhanced Representations

ZHOUMENG WANG
Marketing Programme, The Chinese University of Hong Kong Business School, Hong Kong
e-mail: johnnywang@link.cuhk.edu.hk

ABSTRACT Recommender systems utilizing explicit feedback have witnessed significant advancements and widespread applications over the past years. However, generating recommendations in few-shot scenarios remains a persistent challenge. Recently, large language models (LLMs) have emerged as a promising solution for addressing natural language processing (NLP) tasks, thereby offering novel insights into tackling the few-shot scenarios encountered by explicit feedback-based recommender systems. To bridge recommender systems and LLMs, we devise a prompting template that generates user and item representations based on explicit feedback. Subsequently, we integrate these LLM-processed representations into various recommendation models to evaluate their significance across diverse recommendation tasks. Our ablation experiments and case study analysis collectively demonstrate the effectiveness of LLMs in processing explicit feedback, highlighting that LLMs equipped with generative and logical reasoning capabilities can effectively serve as a component of recommender systems to enhance their performance in few-shot scenarios. Furthermore, the broad adaptability of LLMs augments the generalization potential of recommender models, despite certain inherent constraints. We anticipate that our study can inspire researchers to delve deeper into the multifaceted dimensions of LLMs' involvement in recommender systems and contribute to the advancement of the explicit feedback-based recommender systems field.

INDEX TERMS Large language models, recommender systems, ChatGPT, representations.

I. INTRODUCTION

Recommender systems are defined as techniques that utilize users' explicitly or implicitly expressed preferences to provide recommendations for items of interest, address the issue of information overload, and deliver novelty and surprise [1]. With the advancement of deep learning, the field of recommender systems has witnessed significant progress in recent years. Initially, collaborative filtering and ID-based methods are widely adopted across diverse recommendation scenarios [2], [3], [4]. Subsequently, there has been a growing research focus on incorporating textual side information into recommender systems to develop knowledge-based [5], [6], [7] and content-based [8], [9] approaches that effectively leverage explicit feedback.

However, the majority of recommendation methods continue to grapple with multiple long-standing challenges. The mobile nature of cyber users and the continuous emergence of new items have underscored the significance of few-shot scenarios, where recommender systems are required to provide recommendations based on limited user information. Simultaneously, recommender systems commonly possess a task-specific property that constrains their generalization capabilities across different data sources and application scenarios. Such property is currently being challenged in the dynamic cyberspace, where explicit feedback from users has become increasingly complex and overwhelming in volume. Moreover, as essential tools for consumer engagement, marketing, and business analysis [10], [11], recommender

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

systems necessitate interpretability and transparency; nevertheless, the integration of deep learning has hindered these aspects.

The recent advancements in large language models (LLMs) have offered promising prospects for addressing the aforementioned challenges. Emerging LLMs with generative and logical reasoning capabilities, such as ChatGPT, exhibit remarkable proficiency in text summarization and possess potential for association [12], [13], thereby endowing them with a natural aptitude for engagement in textual explicit feedback processing. Meanwhile, the integration of LLMs into diverse recommendation tasks from various perspectives has emerged as a pivotal area of investigation. Nevertheless, prior research [14] suggests that when employed directly and solely as a recommender system in few-shot scenarios, LLMs do not demonstrate superior performance across various tasks compared to traditional recommendation models. In contrast, recent studies highlight LLMs' effective participation in recommendations as a component of recommender systems [15], [16]. This motivates our novel research proposal: investigating the potential of utilizing LLMs to generate user and item representations using textual explicit feedback, thereby enhancing the performance of existing recommender models in few-shot scenarios.

To investigate this subject, we conduct an in-depth study by referencing previous research [14], [17]. We develop a template to process movie reviews from a deliberately selected public dataset using LLMs to generate user and item representations. These representations are then incorporated into selected recommendation models for evaluation on two tasks: interaction prediction and direct recommendation. To specifically investigate the extraction and association capabilities of the experimental LLMs, we manually adjusted the number of training samples to simulate a few-shot scenario.

Comprehensive experimental results indicate that utilizing LLMs for representation generation significantly enhances the performance of specific recommendation models in a few-shot scenario, demonstrating that LLMs can effectively serve as an explicit feedback processing method for multiple recommendation tasks. Our manual observations also suggest that certain LLMs with generative and logical reasoning capabilities possess a distinctive ability to generate supplementary information through association. LLMs' broad applicability across diverse scenarios and proficiency in processing textual information even in the absence of quantitative metrics can augment the generalization potential of recommender systems. It is worth noting that the observed enhancements are more pronounced in recommendation models that integrate neural networks. This phenomenon could be attributed to inherent constraints imposed by model structures and characteristics of the embeddings.

We hope the results of this experiment can inspire researchers to further explore the incorporation of LLMs into the recommendation process, while offering valuable insights in specific research fields, such as interpretability, cold-start challenges, and model enhancement within explicit feedback-involved recommender systems.

II. RELATED STUDY

A. EXPLICIT FEEDBACK FOR RECOMMENDATION

In contrast to implicit feedback derived primarily from user behavior observations, explicit feedback is openly and actively provided by users themselves to reflect their preferences and attitudes. The concept of explicit feedback mentioned in the book *Recommender Systems: An Introduction* encompasses ratings and annotations [18], while Konstan and Riedl [19] broaden its definition to include diverse forms of user-contributed content such as reviews, tags, blog posts, tweets, Facebook updates, among others.

In previous studies, ratings have been regarded as a crucial form of explicit feedback that enhances the performance of recommender systems [20], [21] and can be combined with implicit feedback to cater to diverse recommendation tasks [22], [23]. Text, serving as another manifestation of explicit feedback, can also be leveraged by recommender systems. Textual explicit feedback is commonly manifested as user reviews and comments [24] that are generated in various languages [25]. Other forms of textual explicit feedback include but are not limited to Tweets [26], web chats [27], messages accompanied by geographic information [28], and Tags [29]. Therefore, natural language processing (NLP) plays a crucial role in constructing recommender systems that rely on textual explicit feedback. Text mining has long been considered as an essential prerequisite in various recommendation models [24], encompassing techniques such as Latent Dirichlet Allocation (LDA) [30], TF-IDF [29], word segmentation [25], rule-based classifiers [31], and more. The processed text can be leveraged to support recommender systems built through approaches such as collaborative filtering [25], [30], content-based filtering [27] or knowledge-based [28].

In recent years, the embedding process has emerged as a prominent focus in recommendation studies due to advancements in related research. The utilization of LLMs in recommendation has been increasingly prevalent owing to their proficiency in comprehending and processing human natural language [32]. Transformer architecture models (*e.g.*, BERT, GPT, and T5 [33]) have been extensively employed in aspects including Pre-training, Fine-tuning, and Prompting [32]. Attention mechanism has also been integrated in the development of recommender system models. For instance, NARRE [34], a neural attention recommendation framework utilizing user reviews, is introduced to simultaneously predict users' ratings towards items and generate review-level explanations for the prediction. Other attention models such as TARMF [35] and MPCN [36] that leverage textual explicit feedback also exhibit superior performance across diverse recommendation tasks compared to existing deep learning-based recommendation models (*e.g.*, ConvMF [37], DeepCoNN [38]).

B. CHATGPT FOR RECOMMENDATION

Released by OpenAI in 2022, ChatGPT [39] is an advanced LLM and dialogue system that has demonstrated exceptional performance across various vertical domains. It showcases remarkable capabilities in context-based comprehension, summarization, and text generation [12]. The investigation into the methodology of transferring and employing ChatGPT's extensive knowledge and paradigm acquired from large-scale corpora to recommendation scenarios has emerged as a cutting-edge pursuit in the academic domain.

ChatGPT can independently serve as a versatile recommendation model capable of handling various recommendation tasks. Liu et al. [14] consider ChatGPT as a self-contained recommender system and construct a benchmark to track its performance in specific recommendation tasks, such as rating prediction and direct recommendation. ChatGPT can also serve as a component of existing recommender systems. Gao et al. [16] introduce Chat-REC, which employs ChatGPT as an interface for conversational recommendations, thereby enhancing the performance of existing recommendation models and rendering the recommendation process more interactive and explainable. Dai et al. [13] propose ChatAug that utilizes ChatGPT to rephrase sentences for textual data augmentation, simultaneously demonstrating the effectiveness of ChatGPT as a text summarization tool when accompanied by pretrained language models (BERT).

In terms of natural language generation tasks, ChatGPT demonstrates remarkable proficiency in generating persuasive recommendation interpretations and advertisements under specific conditions [14], [40]. Related research also suggests that the engagement of ChatGPT could be a innovative solution to address few-shot learning challenges [41]. However, recent research [14] reveals that when employed independently in few-shot scenarios as a recommender system, ChatGPT's performance falls short compared to a series of classical recommendation models across diverse recommendation tasks, such as top-N direct recommendations. The aforementioned studies inspire us to explore the utilization of ChatGPT as an explicit feedback processing method indirectly participating in few-shot recommendation scenarios.

III. REPRESENTATIONS GENERATION

A. TASK FORMULATION

ChatGPT is designed to excel in user-oriented tasks, enabling us to adopt prompting paradigms [42] to target specific tasks without the need for fine-tuning. Drawing partly from relevant studies [14], our experiment initially utilizes the well-established ChatGPT model, *gpt-3.5-turbo*, to generate textual user and item representations by providing ChatGPT with tailored prompts. Each prompt consists of three components: review injection, task description, and format indicator. The review injection is designed to provide ChatGPT with a sequence of reviews from the same subject (a specific user or item). The task description aims to elucidate the input materials and establish clear task requirements. The format indicator serves to standardize response formats and constrain content scope. Additionally, we set a limiter when generating prompts to prevent them from exceeding the maximum token limit in ChatGPT.

Given that the API interface of ChatGPT necessitates its invocation in a conversational format, we assume the template τ, which denotes the procedure for employing ChatGPT to generate a textual representation of a specific subject by utilizing its review collections. Formally, this can be expressed as

$$\tau = [X]suffix[Y] \tag{1}$$

where Y represents a slot subsequently filled by ChatGPT's response, *suffix* (i.e., task description and format indicator) identifies certain text specifically designed to guide ChatGPT in accomplishing representation generation, and input X is a sequence of reviews r pertaining to the specific subject, formally:

$$X = \{r_1, r_2, r_3, \ldots, r_n\}. \tag{2}$$

B. GENERATE TEXTUAL REPRESENTATIONS BY USING CHATGPT

The example in Fig.1 illustrates the generation of a user representation through template τ. It is noteworthy that the generation of the item representations also adheres to template τ, albeit with a slightly different *suffix*; we modify the task description context for item representations generation to "(...Based on your understanding of these movie reviews, summarize the movie's tag and scenes, associate and infer what type of audience and fans may be attracted by this movie." In response to this description, ChatGPT would provide associations and inferences such as "Audiences who prefer heartwarming scenes and happy endings" for item representations.

ChatGPT incorporates a certain degree of randomness to ensure the diversity of generated response, which may pose challenges in terms of reproduction and evaluation. The implementation of the format indicator component has been observed to effectively standardize the responses and mitigate irrelevant variations. During preliminary training with small sample sizes, ChatGPT exhibits exceptional association and inference capabilities that surpass our initial expectations. In certain instances, ChatGPT accurately "guesses" a specific movie and subsequently retrieves comprehensive information from its own database, even when the movie title is not explicitly mentioned in the original reviews. To ensure controlled variables, we explicitly instruct ChatGPT to exclusively focus on materials provided by us when generating representations.

C. EMBED TEXTUAL REPRESENTATIONS BY USING LANGUAGE MODELS

After generating textual user and item representations, we employ MacBERT [43], a pre-trained LLM for Chinese,

Prompt	Review Injection	1. This documentary seems more like an animated work with beautiful sceneries and eye-catching plots. The shot of the bear cub standing up when attacked by a cheetah and the mother bear not far behind it started to roar is super super COOOOL! 2. Instead of being a conformist pilot who fights for a fleeting faith, why not become a free-flying Mediterranean pirate in your broken airplane? It's never enough to watch a movie of Miyazaki just once ~ 3. The terminal of love is the reason why I depart, no matter how far. (Note: This is a lyric from a famous Chinese song sung by Eason Chan.) 4. (…)
	Task Description	The aforementioned movie reviews, authored by the same user, are arranged in chronological order from earliest to latest. Based on your understanding of these movie reviews, make reasonable associations and inferences about the movie tags, styles, and elements this user currently prefers.
	Format Indicator	Concise answer, using keywords and short sentences, in enumeration form, answer within 300 words, starting with "Movie Tags:"
ChatGPT Response		Tags: love, adventure story, warm-heart, freedom, psychology, religion, animal, nature Movie styles: children, animation, documentary, travel film Elements: flying, religious faith, philosophical thinking on human nature, warm and touching scenarios, attractive close-ups and detailed scenes, furry lovely animals, nature adventures experience, stories in vacations

FIGURE 1. Example of using ChatGPT to generate a textual user representation. Notably, the original reviews, prompts, and ChatGPT responses are all in Chinese; we employ ChatGPT to translate them into english for improved readability.

to embed them to become our experimental dataset. Simultaneously, we construct a control dataset by concatenating reviews that belong to the same subject (item or user), embedding them with MacBERT, and merging the outputs. Additionally, we use a pre-trained Chinese Word2vec model [44] that does not employ attention mechanism to generate embeddings as an extra reference in some cases.

The length of each embedding generated using MacBERT is 1,024, while the length of those generated using Word2vec is 200. Considering the superior efficiency of MacBERT in natural language embedding tasks, we primarily utilize MacBERT-processed embeddings as our main control datasets and only refer to experimental results obtained from using Word2vec-processed embeddings under specific conditions. The model selection as well as the embedding process partially drew upon a relevant study [13].

IV. EVALUATION

To assess the effectiveness of LLMs as a textual explicit feedback processing method for recommender systems, we conduct ablation studies on diverse tasks with the aim of answer the following research questions:

- RQ1: Do the LLM-processed user and item representations exhibit disparities compared to the original reviews?
- RQ2: How effectively do these representation function across different recommendation models and tasks, in a few-shot scenario?
- RQ3: Do the textual representations generated by ChatGPT in our experiment possess additional observable attributes and features, beyond those demonstrated in the aforementioned experiment results?

A. EXPERIMENTAL SETUP

1) WORKFLOW

Building upon previous studies [13], [14], we design our experimental workflow as follows: Firstly, we construct eligible datasets that include explicit user feedback and relevant information (Section IV.B). Secondly, the select user profiles and reviews are transformed into prompts for ChatGPT to generate textual user and item representations (elaborated in Section III). Thirdly, the textual representations generated by ChatGPT undergo manual observation for case study purposes (Section IV.E) while concurrently being embedded by using MacBERT to construct an experimental dataset. Finally, the experimental dataset is incorporated into selected recommendation models for various recommendation tasks(Section IV.C, 4.D), along with control datasets. The complete workflow of our experimental process is illustrated in Fig.2.

2) BASELINES AND METRICS

In Section IV.C, we examine the disparities between the embeddings in the experimental dataset (ChatGPT-processed and MacBERT-embedded) and the embeddings in the control

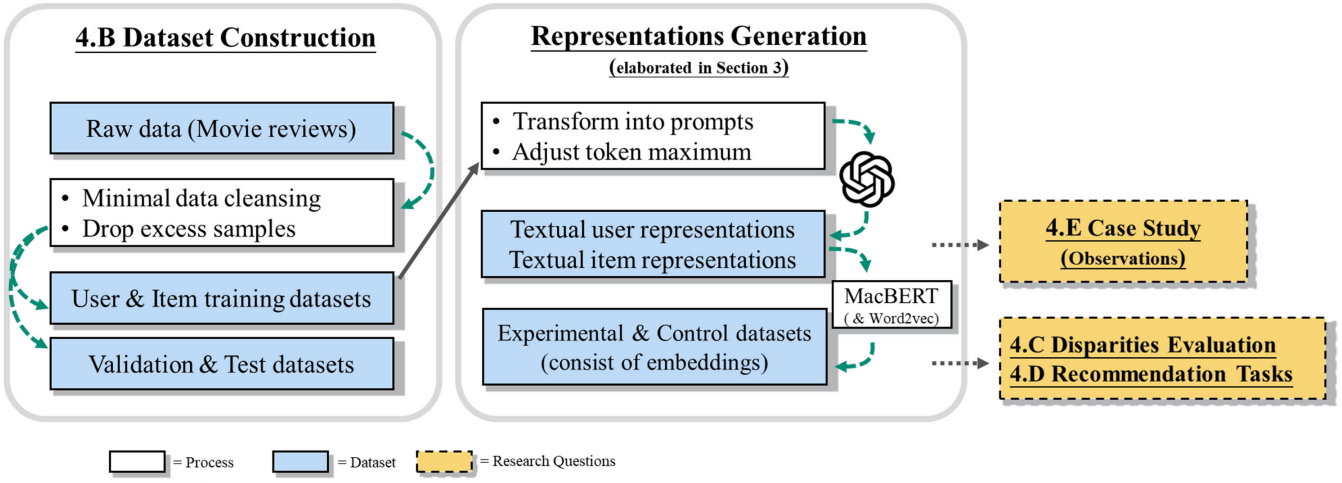

FIGURE 2. Schematic representation of the complete experimental workflow.

dataset (non-ChatGPT-processed and MacBERT-embedded). We employ three statistical methods [45], namely cosine similarity, Manhattan distance, and Euclidean distance, to quantify the semantic relationships between embeddings of each subject (user/item) across the two datasets, namely $embX$ from the experimental datasets and $embX'$ from the control datasets. We computed the mean cosine similarity, mean Manhattan distance, and mean Euclidean distance by averaging the results across all the subjects. The formula is presented below, where n represents the size of the dataset and d is the length of an individual embedding (1,024 for MacBERT embeddings):

$$\textit{mean Manhattan distance} = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{d}|embX_{ki} - embX'_{ki}| \tag{3}$$

$$\textit{mean Euclidean distance} = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\sum_{i=1}^{d}(embX_{ki} - embX'_{ki})^2} \tag{4}$$

$$\textit{mean cosine similarity} = \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i=1}^{d} embX_{ki}embX'_{ki}}{\sqrt{\sum_{i=1}^{d} embX_{ki}^2}\sqrt{\sum_{i=1}^{d} embX_{ki}'^2}}. \tag{5}$$

In Section IV.D, we evaluate the effectiveness of incorporating the LLM-processed embeddings into classical recommendation models for two recommendation tasks: interaction prediction and direct recommendation. The former constitutes a pivotal component in some neural network-based recommender systems [46], [47], while the latter represents a prevalent recommendation task.

For interaction prediction (i.e., predicting whether a user will engage in interaction with a specific item), we employ Linear, MLP [47], and CNN [48] models as our baselines. We consider user-item interactions as labels; specifically, ground truth interactions will be labeled as 1, while negative samples (labeled as 0) are generated by randomly assigning each user an item that they have not interacted with in reality. Given the binary classification nature of the task, we utilize Accuracy, Precision, and F1 Score as evaluation metrics to assess performance. For direct recommendation (i.e., recommending items that are most likely to align with a user's preferences), we employ BPR-MF [49], NCF-Linear, NCF-MLP, and NCF-CNN [50] as baselines.

We evaluate their performance using widely adopted metrics in recommender system studies, namely top-k Hit Ratio (HR@k) and top-k Mean Reciprocal Rank (MRR@k). Considering the few-shot scenario, we report results on either HR@10,100 and either MRR@10,100. It is worth noting that despite varying in structural configurations, the aforementioned baselines integrating MLP and CNN neural networks have a comparable number of layers and are not fine-tuned respectively.

B. DATASET CONSTRUCTION

The dataset employed in our experiment is the publicly available Douban Chinese Moviedata-10M [51], which shares similarities with the benchmark MovieLens dataset [52] in terms of content and format. The Douban dataset encompasses a substantial amount of explicit feedback provided by platform users, each sample presented as a user-item interaction comprising a user ID, an item ID, a piece of movie review, and other pertinent information such as a rating and a timestamp. In contrast to the MovieLens dataset, the Douban dataset primarily comprises Chinese text, and encompasses a substantial number of colloquial expressions, internet memes, emojis, and other intricate linguistic corpora. We intentionally perform minimal data cleansing to thoroughly evaluate ChatGPT's proficiency in handling real-world explicit feedback.

To construct our experimental dataset, a cohort of 1,000 users is randomly selected. We extract the historical user-item

TABLE 1. User training dataset details.

Dataset	Statistical information	
User representation training set	Interaction sample count	7,270
	Proportion of users with samples <= 5	49.30%
	Proportion of reviews with words <= 100	81.03%
	Proportion of reviews with words <= 150	97.66%

TABLE 2. Semantic distances between experimental and control datasets.

Dataset	Statistical Methods	Result
User Embeddings	Mean Cosine similarity	0.94
	Mean Euclidean distance	7.62
	Mean Manhattan distance	194.09
Item Embeddings	Mean Cosine distance	0.95
	Mean Euclidean distance	6.84
	Mean Manhattan distance	174.87

interaction samples of these users and sort them in chronological order. The item IDs corresponding to the two most recent interactions are extracted as test and validation samples, respectively, and are concatenated into their respective sets. The remaining interaction samples of these users constitute the training dataset for inputting into ChatGPT to generate textual user representations. To simulate a few-shot scenario, we artificially control the number of interaction samples per user by randomly discarding excess samples while ensuring at least one sample per user remains. Detailed statistical information about the user training dataset is provided in Tab.1.

After extracting all the samples corresponding to the aforementioned 1,000 users, we construct the remaining samples as the item training dataset. Each item in the dataset is designed to have at least one corresponding sample. We apply filtering rules to ensure that all items present in the user training dataset, validation dataset, and test dataset exist simultaneously in the item training dataset. Moreover, in order to maintain control variables when constructing the control datasets, we restrict the number of samples per item to a maximum of 10. Following filtration, approximately 98.05% of the reviews in these samples are less than 150 words long (with 79.84% being under 100 words).

Eventually, during the stage of constructing the training dataset, we obtain a user training dataset consisting of 1,000 users, encompassing a total of 7,270 interaction samples; additionally, we create validation and test datasets with each comprising 1,000 samples (one per user); finally, an item training dataset is compiled containing over 300,000 interaction samples from 38,750 items. By providing ChatGPT with tailored prompts derived from the two training datasets (elaborated in Section III), we generate a corresponding number of textual user and item representations, which are then combined to form textual user and item representation datasets and subsequently embedded by language models to form experimental datasets. The workflow for constructing the datasets is illustrated in Fig.3.

As outlined in Section III, we employ MacBERT and Word2vec to embed the textual item and user representations for generating embedding datasets. Additionally, we build control datasets in accordance with the methodology detailed in the section.

In total, we acquire the following datasets:

- A pair of textual representation datasets (user & item).
- A pair of experimental datasets (user & item): ChatGPT-processed + MacBERT-embedded
- Three pairs of control datasets (user & item): Only MacBERT-embedded; Only Word2vec-embedded; ChatGPT-processed + Word2vec-embedded.

C. DISPARITIES EVALUATION (RQ1)

In this section, we quantify the semantic relationships between embeddings of each subject (user/item) across the experimental dataset (ChatGPT-processed + MacBERT-embedded) and the control dataset (MacBERT-embedded). The evaluation method proposed in Section IV.A is employed to obtain the statistical measurement results presented in Tab.2.

The cosine similarity metric primarily focuses on the angular relationship between two vectors in a multi-dimensional space. When comparing two semantically similar sentences, regardless of their length, the angle between their vectors becomes smaller, resulting in a higher value for cosine similarity. Euclidean distance and Manhattan distance calculations encompass both direction and magnitude, which can serve as a complementary measure to cosine similarity.

When comparing the experimental and control datasets, both in terms of items and users, we observe that the result of Mean Cosine distance approaches 1, indicating a significant semantic similarity between the representations generated by ChatGPT and the original reviews. We also note that the Mean Euclidean and Manhattan distances deviate significantly from zero. Based on these results, we suggest that while the ChatGPT-generated representations demonstrate comparable semantics to the original reviews, they do exhibit significant disparities in terms of information content and quantity. This discrepancy may be attributed to their truncated length and refined content. In general, the aforementioned findings partially substantiate the effectiveness of ChatGPT in extracting a substantial portion of salient features and crucial information from the original reviews, albeit with a reconfigured textual composition and altered content. The reconfiguration and alteration will be examined in Section IV.E through a detailed case study.

D. PERFORMANCE COMPARISON ON RECOMMENDATION TASKS (RQ2)

Fig.4 depicts the workflow for conducting ablation experiments on two recommendation tasks using user-item interactions and the user and item embeddings from both experimental and control datasets. Notably, we conduct 10 independent repetitions to train each model in the two

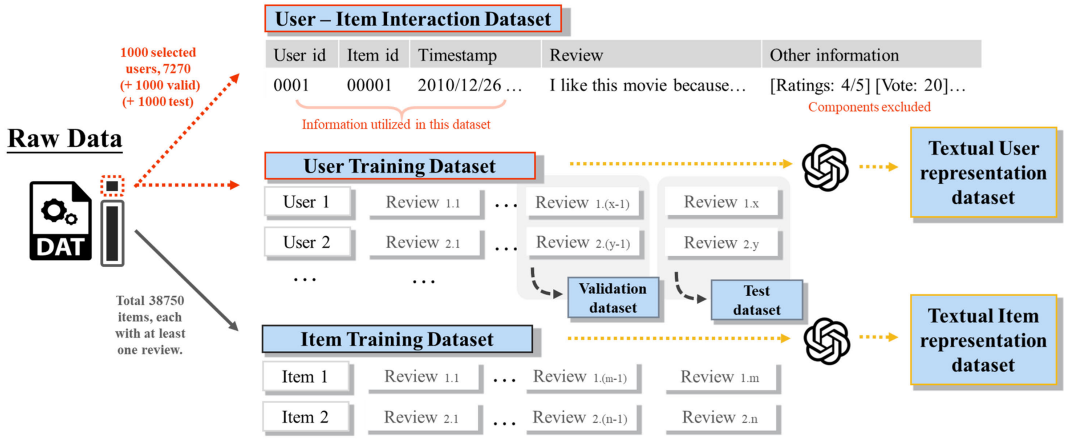

FIGURE 3. Schematic representation of the datasets construction workflow.

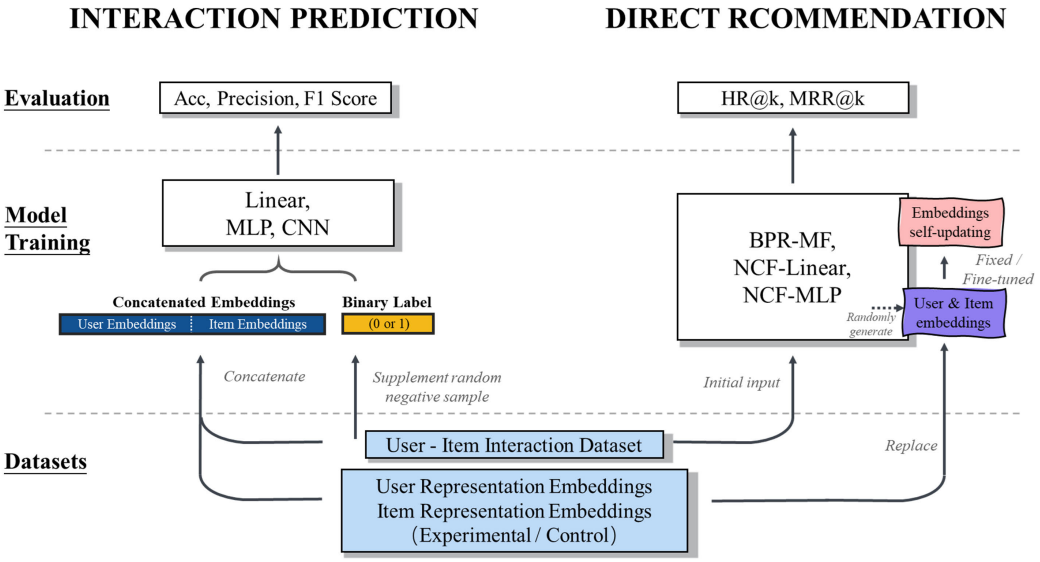

FIGURE 4. Schematic representation of the experimental workflow for two recommendation tasks.

recommendation tasks and report the average results, aiming to comprehensively investigate their overall performance.

For the interaction prediction task, we concatenate the user embedding and item embedding from the same interaction samples (including randomly generated negative samples) and input them into binary classification models along with their labels for training. Subsequently, we assess the model's Accuracy, Precision, and F1 Score on the ground truth test dataset.

For the direct recommendation task, we initialize the recommendation model with user-item interaction dataset. Upon model initialization, the BPR and NCF models automatically generate random embeddings for users and items, which are subsequently updated during training based on the models' learning from the user-item interaction dataset (with ratings). After model training, these fine-tuned embeddings serve as a foundation for recommending items to selected users. In our study, we eliminate ratings by substituting them with a uniform constant to prevent the recommendation model from relying on ratings. As compensation, we replace the model-automatically-generated embeddings with the user and item embeddings in our experimental and control datasets. We assess the performance of the models using HR (Hit Rate) and MRR (Mean Reciprocal Rank), while additionally considering scenarios where these embeddings continue to undergo fine-tuning or remain fixed during model training.

1) INTERACTION PREDICTION

For the interaction prediction task, we conduct ablation experiments on experimental and control datasets using classical Linear, MLP, and CNN models respectively. The

TABLE 3. Performance comparison on interaction prediction tasks.

Method	Dataset	Statistical Measurements		
		Accuracy	Precision	F1 Score
MLP	**ChatGPT + MacBERT**	**0.592**	**0.601**	**0.632**
	Only MacBERT	0.552	0.570	0.523
	ChatGPT + Word2vec	0.501	0.500	0.500
	Only Word2vec	0.500	0.500	0.500
Linear*	All datasets	0.500	0.250	0.335
CNN*	All datasets			

* The Linear and CNN models exhibited unsuccessful convergence, with their Precision rate oscillating between either 0 or 0.5 and while F1 Score oscillating between either 0 or 0.67. We calculated the average of all experimental results and hereby provide an explanation.

statistical measurements obtained from these experiments are reported in Tab.3.

Based on our observations, under the same MLP model, the experimental dataset demonstrate superiority over the control datasets. The results suggest that the incorporating ChatGPT-processed representation embeddings holds the potential to enhance certain recommender models that employ neural networks in a few-shot scenario.

Notably, among all experimental models that integrated neural networks, the MLP model stands out as the only one to exhibit statistically significant results in both experimental and control datasets. In contrast, we observe that the CNN model exhibited a significantly high training loss and failed to successfully converge during training. We speculate that this phenomenon can be attributed to the length of the concatenated embedding and the limited number of the training samples, as certain neural networks may encounter detrimental effects on learning and convergence with a few-shot scenario characterized by an abundance of training features. This partially elucidates the unsatisfactory model performance observed in our experimental findings.

2) DIRECT RECOMMENDATION

For the direct recommendation task, we conduct ablation experiments using experimental and control datasets on the BPR and NCF recommendation models, and investigate the impact of enabling or disabling automatic model updating during training. The specific experimental results are presented in Tab.4 and Tab.5, with all outputs appropriately rounded to ensure a reader-friendly presentation.. Due to significant variations in performance among different recommendation models, we adopt HR and MRR @10 for NCF models and @100 for BPR models, respectively, to effectively showcase their performance. Furthermore, we present the percentage improvement of experimental models in comparison to the baseline model (which employs randomly generated embeddings) across diverse datasets, with a primary focus on results demonstrating an increase of 200% or more for emphasis.

The ablation experiments demonstrate the significance of utilizing ChatGPT-processed embeddings to enhance a series of recommended models in few-shot scenarios. This enhancement is particularly evident in recommendation

TABLE 4. Performance comparison on BPR-MF model.

Method		Dataset	Statistical Measurements			
			HR@100		MRR@100	
BPR-MF	Fine-tuned	ChatGPT + MacBERT	0.003		0.003	
		Only MacBERT	0.003		0.004	200%
		ChatGPT + Word2vec	**0.011**	**550%**	**0.008**	**400%**
		Only Word2vec	0.008	400%	0.006	300%
	Fixed	ChatGPT + MacBERT	0.006	300%	0.001	
		Only MacBERT	0.006	300%	0.001	
		ChatGPT + Word2vec	0.005	250%	0.001	
		Only Word2vec	0.003		0.001	
	Random		0.002	100%*	0.002	100%*

* The table presents the significant results of the experimental models in comparison to the baseline model across diverse datasets, denoted as %.

TABLE 5. Performance comparison on NCF models.

Method		Dataset	Statistical Measurements			
			HR@10		MRR@10	
NCF-Linear	Fine-tuned	ChatGPT + MacBERT	0.041		0.003	
		Only MacBERT	0.033		0.003	
	Fixed	ChatGPT + MacBERT	0.080	267%	0.004	200%
		Only MacBERT	0.071	237%	0.006	300%
	Random		0.030	100%*	0.002	100%*
NCF-MLP	Fine-tuned	ChatGPT + MacBERT	0.092		0.006	
		Only MacBERT	0.081		0.004	
	Fixed	**ChatGPT + MacBERT**	**0.210**	**412%**	**0.012**	**300%**
		Only MacBERT	0.162	318%	0.009	225%
	Random		0.051	100%	0.004	100%
NCF-CNN	Fine-tuned	ChatGPT + MacBERT	0.080		0.006	
		Only MacBERT	0.054		0.005	
	Fixed	ChatGPT + MacBERT	0.104	248%	0.013	260%
		Only MacBERT	0.080		0.007	
	Random		0.042	100%	0.005	100%

* The table presents the significant results of the experimental models in comparison to the baseline models across diverse datasets and model structures, denoted as %.

models that incorporate neural networks. Specifically, NCF-MLP outperforms NCF-CNN in terms of both HR and MRR metrics; models that fixed embeddings during training exhibit comparatively superior performance compared to those fine-tuned.

Based on the experimental results, we suggest that the integration of neural networks enhances the recommendation models' capacity to process LLM-generated embeddings, which implies a substantial number of training features.

We speculate that the limited sample size poses challenges for all neural networks, thereby compromising the validity of LLM-generated embeddings when automatically fine-tuned, whereas MLP is the sole network demonstrating superior adaptability in few-shot scenarios in our experiments (as evidenced by the results presented in the interaction prediction recommendation task). Meanwhile, recommendation models that do not incorporate neural networks encounter significant difficulties when dealing with lengthy embeddings. This could partially account for the superior experimental results obtained by utilizing Word2vec-embedded embeddings (which have shorter lengths compared

to MacBERT-embedded embeddings) in BPR-MF models as opposed to other datasets.

E. CASE STUDY (RQ3)

In addition to conducting ablation experiments, we perform a comprehensive case study on the textual user and item representations to complement our findings and uncover potentially overlooked information within the embedding process. Our manual observations suggest that ChatGPT demonstrates exceptional proficiency in processing explicit textual feedback.

Specifically, it consistently demonstrates precise recognition and comprehension of contextual information with varying sentiment tendencies, even in the absence of quantitative metrics such as ratings. Notably, ChatGPT effectively handles reviews that contain positive, neutral, and negative snippets simultaneously by either disregarding the negative portion or considering an opposing viewpoint for recommendations. Additionally, ChatGPT adeptly identifies quotations within the reviews (e.g., movie lines, plots, extra materials) and utilizes them appropriately. The aforementioned observations collectively suggest that ChatGPT holds the potential to enhance the generalization capability of recommendation models by providing adaptability for diverse recommendation scenarios, such as social media platforms that exclusively comprising textual content.

Meanwhile, in contrast to conventional language models, ChatGPT demonstrates a unique ability to generate expansion context even when provided with limited information. While traditional NLP approaches primarily focus on keyword identification and extraction, ChatGPT goes beyond by introducing new content that may deviate from the original corpus. For instance, as depicted in Fig.1, ChatGPT suggests the keyword ''furry lovely animals,'' possibly due to the user's preference for documentaries featuring bears and animations. Essentially, ChatGPT ''refines and reinforces'' initial representations by augmenting them with supplementary information through association and inference. This could partially account for the observed semantic similarity yet content disparity between the experimental and control datasets, as evidenced by the findings in Section IV.C. Furthermore, the effectiveness of the refined and reinforced representations is demonstrated with support from the experimental results presented in Section IV.D. This partially indicates that the additional information contained within these representations, generated through ChatGPT's association and inference, carries significant implications. In other words, these supplementary pieces of information reasonably reflect users' underlying thoughts to a certain extent. To summarize, ChatGPT demonstrates its effectiveness in handling few-shot recommendation scenarios compared to conventional language models, owing to its distinctive capabilities in associative thinking and logical reasoning.

It is noteworthy that in this experiment, ChatGPT functions as a symbolic representation of emerging LLMs endowed with generative and logical reasoning capabilities. Considering the continuous advancements in technology, forthcoming LLMs equipped with enhanced proficiencies in association and inference may ultimately supplant ChatGPT within our experimental framework. Nevertheless, the insights derived from this investigation retain significant reference value for future studies.

V. CONCLUSION

In this study, we conduct ablation experiments to assess the effectiveness of harnessing LLMs to enhance few-shot recommender systems in various recommendation tasks. Despite the limitations imposed by model structures, the inclusion of LLM-processed representations significantly enhances the performance of specific neural network-based recommendation models in our experimental few-shot scenario. Based on the experimental results, we suggest that LLMs equipped with generative and logical reasoning capabilities can serve as an effective NLP method for recommender systems, proficiently handling textual explicit feedback through their distinctive capabilities and enhancing the generalization potential of recommendation models. Moving forward, we envision integrating additional recommendation models based on neural networks into our study. Furthermore, we are intrigued by the potential business applications (e.g., marketing analytics, advertisement generation) of the ChatGPT-generated textual user and item representations.

ACKNOWLEDGMENT

The author wishes to extend his appreciation to Prof. Jimbo, Prof. Howard, Li Zhi, and Lei for their support throughout his academic journey.

REFERENCES

[1] L. Mocean and C. M. Pop, ''Marketing recommender systems: A new approach in digital economy,'' *Inf. Economica*, vol. 16, no. 4, p. 142, 2012.
[2] J. Bobadilla, S. Alonso, and A. Hernando, ''Deep learning architecture for collaborative filtering recommender systems,'' *Appl. Sci.*, vol. 10, no. 7, p. 2441, Apr. 2020.
[3] F. Rezaimehr and C. Dadkhah, ''A survey of attack detection approaches in collaborative filtering recommender systems,'' *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2011–2066, Mar. 2021.
[4] S. Zhang, L. Yao, A. Sun, and Y. Tay, ''Deep learning based recommender system: A survey and new perspectives,'' *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, Jan. 2020.
[5] F. Cena, L. Console, and F. Vernero, ''Logical foundations of knowledge-based recommender systems: A unifying spectrum of alternatives,'' *Inf. Sci.*, vol. 546, pp. 60–73, Feb. 2021.
[6] M. Dong, X. Zeng, L. Koehl, and J. Zhang, ''An interactive knowledge-based recommender system for fashion product design in the big data environment,'' *Inf. Sci.*, vol. 540, pp. 469–488, Nov. 2020.
[7] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, and A. Darwesh, ''A systematic study on the recommender systems in the E-commerce,'' *IEEE Access*, vol. 8, pp. 115694–115716, 2020.
[8] D. Mittal, S. Shandilya, D. Khirwar, and A. Bhise, ''Smart billing using content-based recommender systems based on fingerprint,'' in *ICT Analysis and Applications*, vol. 2. Cham, Switzerland: Springer, 2020, pp. 85–93.
[9] Y. Pérez-Almaguer, R. Yera, A. A. Alzahrani, and L. Martínez, ''Content-based group recommender systems: A general taxonomy and further improvements,'' *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115444.
[10] A. Ansari, S. Essegaier, and R. Kohli, ''Internet recommendation systems,'' *J. Marketing Res.*, vol. 37, no. 3, pp. 363–375, Aug. 2000.

[11] A. V. Bodapati, "Recommendation systems with purchase data," *J. Marketing Res.*, vol. 45, no. 1, pp. 77–93, Feb. 2008.
[12] T. B. Brown, "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.
[13] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, "AugGPT: Leveraging ChatGPT for text data augmentation," 2023, *arXiv:2302.13007*.
[14] 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*.
[15] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. Di Noia, and E. Di Sciascio, "Evaluating ChatGPT as a recommender system: A rigorous approach," 2023, *arXiv:2309.03613*.
[16] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-REC: Towards interactive and explainable LLMs-augmented recommender system," 2023, *arXiv:2303.14524*.
[17] Z. Kefato, S. Girdzijauskas, N. Sheikh, and A. Montresor, "Dynamic embeddings for interaction prediction," in *Proc. Web Conf.*, Apr. 2021, pp. 1609–1618.
[18] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[19] J. A. Konstan and J. Riedl, "Recommender systems: From algorithms to user experience," *User Model. User-Adapted Interact.*, vol. 22, nos. 1–2, pp. 101–123, Apr. 2012.
[20] Q. Zhao, F. M. Harper, G. Adomavicius, and J. A. Konstan, "Explicit or implicit feedback? Engagement or satisfaction: A field experiment on machine-learning-based recommender systems," in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, Apr. 2018, pp. 1331–1340.
[21] S.-Y. Liu, H. H. Chen, C.-M. Chen, M.-F. Tsai, and C.-J. Wang, "IPR: Interaction-level preference ranking for explicit feedback," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 1912–1916.
[22] N. N. Liu, E. W. Xiang, M. Zhao, and Q. Yang, "Unifying explicit and implicit feedback for collaborative filtering," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2010, pp. 1445–1448.
[23] G. Jawaheer, M. Szomszor, and P. Kostkova, "Comparison of implicit and explicit feedback from an online music recommendation service," in *Proc. 1st Int. Workshop Inf. Heterogeneity Fusion Recommender Syst.*, Sep. 2010, pp. 47–51.
[24] Y. Betancourt and S. Ilarri, "Use of text mining techniques for recommender systems," in *Proc. 22nd Int. Conf. Enterprise Inf. Syst.*, 2020, pp. 780–787.
[25] D. Miao and F. Lang, "A recommendation system based on text mining," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2017, pp. 318–321.
[26] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, "Collaborative personalized tweet recommendation," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2012, pp. 661–670.
[27] S. Loh, F. Lorenzi, R. Saldaña, and D. Licthnow, "A tourism recommender system based on collaboration and text analysis," *Inf. Technol. Tourism*, vol. 6, no. 3, pp. 157–165, Jan. 2003.
[28] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D. Z. Rodríguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2124–2135, Apr. 2019.
[29] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent Dirichlet allocation for tag recommendation," in *Proc. 3rd ACM Conf. Recommender Syst.*, Oct. 2009, pp. 61–68.
[30] N. Jakob, S. H. Weber, M. C. Müller, and I. Gurevych, "Beyond the stars: Exploiting free-text user reviews to improve the accuracy of movie recommendations," in *Proc. 1st Int. CIKM workshop Topic-Sentiment Anal. Mass Opinion*, Nov. 2009, pp. 57–64.
[31] Y. Li, J. Nie, Y. Zhang, B. Wang, B. Yan, and F. Weng, "Contextual recommendation based on text mining," in *Proc. Coling, Posters*, 2010, pp. 692–700.
[32] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, and Q. Li, "Recommender systems in the era of large language models (LLMs)," 2023, *arXiv:2307.02046*.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[34] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1583–1592.
[35] Y. Lu, R. Dong, and B. Smyth, "Coevolutionary recommendation model: Mutual learning between ratings and reviews," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 773–782.
[36] Y. Tay, A. T. Luu, and S. C. Hui, "Multi-pointer co-attention networks for recommendation," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2309–2318.
[37] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 233–240.
[38] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, Feb. 2017, pp. 425–434.
[39] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[40] M. Remountakis, K. Kotis, B. Kourtzis, and G. E. Tsekouras, "Using ChatGPT and persuasive technology for personalized recommendation messages in hotel upselling," *Information*, vol. 14, no. 9, p. 504, Sep. 2023.
[41] D. Di Palma, "Retrieval-augmented recommender system: Enhancing recommender systems with large language models," in *Proc. 17th ACM Conf. Recommender Syst.*, Sep. 2023, pp. 1369–1373.
[42] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.
[43] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," 2020, *arXiv:2004.13922*.
[44] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (Short Papers)*, vol. 2, 2018, pp. 175–180.
[45] D. Verma and S. N. Muralikrishna, "Semantic similarity between short paragraphs using deep learning," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2020, pp. 1–5.
[46] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 191–198.
[47] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, and M. Ispir, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.
[48] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.
[49] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," 2012, *arXiv:1205.2618*.
[50] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.
[51] D. Liu, Y. Gao, and Y. Xu. (2019). *Douban Moviedata*. [Online]. Available: http://moviedata.csuldw.com/ and https://github.com/csuldw/AntSpider
[52] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, Jan. 2016.

ZHOUMENG WANG received the bachelor's and Master of Science degrees in marketing from The Chinese University of Hong Kong, Shenzhen, in 2022 and 2023, respectively.

Since his graduation, he has been a research assistant. He received the prestigious Bowen Scholarship and consistently made the Dean's List twice during his study. His research interests include large language models, recommender systems, and digital marketing. In 2022, he has participated in a Kaggle Recommender System Competition and won the Silver Medal.