Is ChatGPT a Good Recommender? A Preliminary Study

Junling Liu*
william.liuj@gmail.com
Alibaba Group
China

Chao Liu*
chize.lc@antgroup.com
Ant Group
China

Peilin Zhou*
zhoupalin@gmail.com
Hong Kong University of Science and Technology(Guangzhou)
China

Renjie Lv
lvrenjie.lrj@antgroup.com
Ant Group
China

Kang Zhou
kangbeyond89@163.com
Alibaba Group
China

Yan Zhang
yanbest0117@163.com
Alibaba Group
China

ABSTRACT

Recommendation systems have witnessed significant advancements and have been widely used over the past decades. However, most traditional recommendation methods are task-specific and therefore lack efficient generalization ability. Recently, the emergence of ChatGPT has significantly advanced NLP tasks by enhancing the capabilities of conversational models. Nonetheless, the application of ChatGPT in the recommendation domain has not been thoroughly investigated. In this paper, we employ ChatGPT as a general-purpose recommendation model to explore its potential for transferring extensive linguistic and world knowledge acquired from large-scale corpora to recommendation scenarios. Specifically, we design a set of prompts and evaluate ChatGPT's performance on five recommendation scenarios, including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Unlike traditional recommendation methods, we do not fine-tune ChatGPT during the entire evaluation process, relying only on the prompts themselves to convert recommendation tasks into natural language tasks. Further, we explore the use of few-shot prompting to inject interaction information that contains user potential interest to help ChatGPT better understand user needs and interests. Comprehensive experimental results on Amazon Beauty dataset show that ChatGPT has achieved promising results in certain tasks and is capable of reaching the baseline level in others. We conduct human evaluations on two explainability-oriented tasks to more accurately evaluate the quality of contents generated by different models. And the human evaluations show ChatGPT can truly understand the provided information and generate clearer and more reasonable results. We hope that our study can inspire researchers to further explore the potential of language models like ChatGPT to improve recommendation performance and contribute to the advancement of the recommendation systems field.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Large-Language Model, ChatGPT, Recommendation System

ACM Reference Format:
Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As a crucial technique for addressing information overload and enhancing user experience, recommendation systems have witnessed significant advancements over the past decade and have been widely used in various web applications such as product recommendation [32, 49, 51, 59], video recommendation [39, 54, 66], news recommendation [55–57], music recommendation [27, 47] and so on. In the meanwhile, with the development of deep learning, recommendation systems have gone through several stages. In early ages, collaborative filtering-based methods [5, 6, 44, 62] are primarily used to model the user's behavior patterns from the user-item interactions. Later on, with the introduction of user and item side information into recommendation systems, content-based recommendation [36, 37, 40, 53, 58] and knowledge-based recommendation [2, 8, 16, 18] have gained attention due to their ability to provide personalized recommendations.

However, most traditional recommendation methods are task-specific. Therefore, specific data is required to train specific models for different tasks or application scenarios, which lack efficient generalization ability. To address this issue, researchers have shifted their focus towards implementing Pretrained Language Models (PLMs) in recommendation scenarios since PLMs have demonstrated impressive adaptability to improve the performance of downstream NLP tasks significantly. To effectively convert user interaction data into text sequences, a variety of prompts [64] is designed to convert user interaction data into text sequences. Furthermore, P5 [19] and M6-Rec [11] focus on building a foundation model to support a wide range of recommendation tasks.

Recently, the emergence of ChatGPT has significantly advanced NLP tasks by enhancing the capabilities of conversational models, making it a valuable tool for businesses and organizations. Chataug et al. [12] leverages ChatGPT to rephrase sentences for text data augmentation. Jiao et al. [23] finds the translation ability of ChatGPT performs competitively with commercial translation products

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
2023. ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

on high-resource and low-resource languages. Bang et al. [3] finds ChatGPT outperforms the previous state-of-the-art zero-shot model by a large margin in the sentiment analysis task. Nonetheless, the application of ChatGPT in the recommendation domain has not been thoroughly investigated, and whether ChatGPT can perform well on classical recommendation tasks remains an open question. Therefore, it is necessary to establish a benchmark to preliminarily evaluate and compare ChatGPT with traditional recommendation models, thereby providing valuable insights and facilitating further exploration of the potential of large-scale language models in recommendation systems.

To bridge this research gap, in this paper, we directly employ ChatGPT as a general-purpose recommendation model that can handle various recommendation tasks, and attempt to explore whether the extensive linguistic and world knowledge acquired from large-scale corpora can be effectively transferred to recommendation scenarios. Our main contribution is the construction of a benchmark to track ChatGPT's performance in recommendation scenarios, and a comprehensive analysis and discussion of its strengths and limitations. Specifically, we design a set of prompts and evaluate ChatGPT's performance on five recommendation tasks, including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Unlike traditional recommendation methods, we do not fine-tune ChatGPT during the entire evaluation process, relying only on the prompts themselves to convert recommendation tasks into natural language tasks. Furthermore, we explore the use of few-shot prompting to inject interaction information that contains user potential interests to help ChatGPT better understand user needs and preferences.

Comprehensive experimental results on Amazon Beauty dataset reveal that, from the perspective of accuracy, ChatGPT performs well in rating prediction but poorly in sequential and direct recommendation tasks, achieving only similar performance levels to early baseline methods on certain metrics. On the other hand, while ChatGPT demonstrates poor performance in terms of objective evaluation metrics for explainable recommendation tasks such as explanation generation and review summarization, our additional human evaluations show that ChatGPT outperforms state-of-the-art methods. This highlights the limitations of using an objective evaluation approach to accurately reflect ChatGPT's true explainable recommendation capabilities. Furthermore, despite ChatGPT's unsatisfactory performance in accuracy-based recommendation tasks, it is worth noting that ChatGPT has not been specifically trained on any recommendation data. Thus, there is still significant potential for improvement in future research by incorporating more relevant training data and techniques. We believe that our benchmark not only sheds light on ChatGPT's recommendation capabilities but also provides a valuable starting point for researchers to better understand the advantages and shortcomings of ChatGPT in recommendation tasks. Moreover, we hope that our study can inspire researchers to design new methods that leverage the strengths of language models like ChatGPT to improve recommendation performance, and contribute to the advancement of the recommendation systems field.

2 RELATED WORK

2.1 Large Language Models and ChatGPT

Language Models (LMs) are a fundamental component of natural language processing (NLP) and have been the focus of research for several decades. Recently, the emergence of large-scale LMs has led to significant progress in NLP tasks such as machine translation[1, 9, 61], summarization[33, 46], and dialogue generation[14, 28].

Large Language Models (LLMs) are a subclass of LMs that leverage massive amounts of data and computational resources to achieve state-of-the-art performance on a wide range of NLP tasks. The history of LLMs can be traced back to the early work on neural networks and language modeling. [4] introduced neural language models that learned to predict the next word in a sentence given the previous words. Later, the development of recurrent neural networks (RNNs) and long short-term memory (LSTM) networks further improved the ability of models to capture long-term dependencies in language[22]. However, traditional neural language models still struggled with capturing the rich semantic and contextual relationships present in natural language. The introduction of the Transformer architecture by [52] was a major breakthrough in this area. The Transformer model utilizes self-attention mechanisms to capture the relationships between all elements in a sequence simultaneously, allowing for more comprehensive contextual understanding. This architecture has been used as the backbone of many successful LLMs, including BERT[13], GPT-2[41], and XLNet[60].

ChatGPT[38] is a state-of-the-art dialogue system developed by OpenAI in 2022. It is a state-of-the-art natural language processing (NLP) model that has been widely used in various vertical domains, such as text generation and dialogue systems. In text generation, ChatGPT has shown impressive results in generating coherent and diverse text, surpassing the performance of previous models [7]. In dialogue systems, it has been used for task-oriented and open-domain conversations, achieving state-of-the-art performance in both settings [65]. Although the value of ChatGPT has been validated in various fields, whether it can still be effective in the recommendation domain remains an under-explored topic, which motivates us to construct such a benchmark to gain insights into the potential of large language models for recommendation systems.

2.2 Language Model for Recommendation

Language Models (LMs), such as BERT [13] and GPT [38], have demonstrated impressive adaptability to improve the performance of downstream NLP tasks significantly, thanks to extensive linguistic and world knowledge learned from large-scale corpora. Inspired by these achievements, an increasing amount of attention is being paid for the application of LMs in recommender scenarios, yielding several recent breakthroughs in this field. For instance, LMRecSys [64] utilizes prompts to reconstitute some recommendation tasks as multi-token cloze tasks, aiming to address zero-shot and data efficiency issues. P5 [19] is the first attempt to integrate different recommendation tasks within a shared conditional language generation framework (i.e., T5 [42]). To effectively convert user interaction data into text sequences, a variety of prompts are desgined to accomodate the specific characteristics of each recommendation task. Similarly, M6-Rec [11] focuses on building a foundation

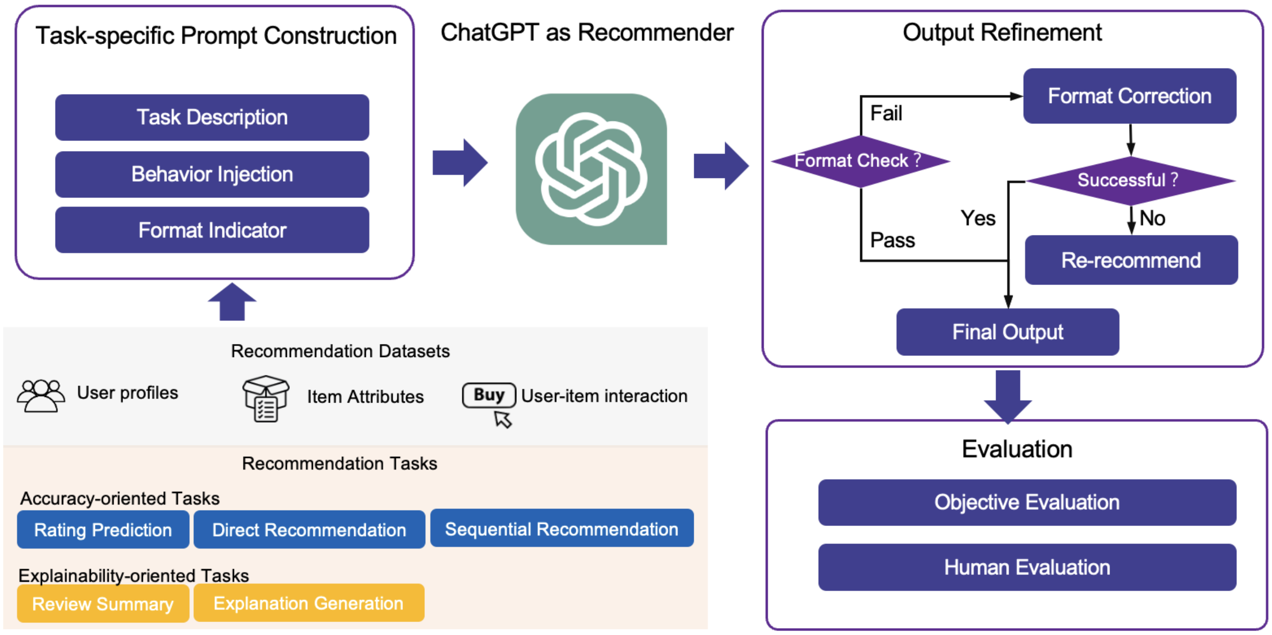

Figure 1: Workflow of utilizing ChatGPT to perform five recommendation tasks and evaluating its recommendation performance.

model to support a wide range of recommendation tasks, including retrieval, ranking, and explanation generation, etc. Notably, the authors also provide practical solutions for model deployment in real-world settings. Chat-REC [17], a concurrent work closely related to our study, leverages ChatGPT as an interface for conversational recommendations, thereby augmenting the performance of existing recommender models and rendering the recommendation process more interactive and explainable.

Different from Chat-REC, our work is inspired by P5 and treats ChatGPT as a self-contained recommendation system that does not rely on any external systems. Based on this, we conduct a thorough evaluation and comparison of its performance on classic recommendation tasks including sequential recommendation, rating prediction, etc. By doing so, we hope our analysis can offer valuable insights for researchers to delve deeper into the potential of large-scale language models in the domain of recommendation.

3 RECOMMENDATION WITH CHATGPT

The workflow of using ChatGPT to complete recommendation tasks is illustrated in Fig.1, which consists of three steps. First, different prompts are constructed based on the specific characteristics of the recommendation tasks (Section 3.1). Second, these prompts are used as inputs for ChatGPT, which generates the recommendation results according to the requirements specified in the prompts. Finally, the output from ChatGPT is checked and refined by the refinement module, and the refined results are returned to the user as the final recommendation results (Section 3.2).

3.1 Task-specific Prompt Construction

In this section, we investigate the recommendation capability of ChatGPT by designing prompts tailored to different tasks. Each prompt comprises three parts: task description, behavior injection, and format indicator. The task description is utilized to adapt recommendation tasks to natural language processing tasks. The behavior injection is designed to assess the impact of few-shot prompting, which incorporates user-item interaction to aid ChatGPT in capturing user preferences and needs more effectively. The format indicator serves to constrain the output format, making the recommendation results more comprehensible and assessable.

3.1.1 Rating Prediction. Rating prediction is a crucial task in recommendation systems that aims to predict the ratings that a user would give to a particular item. This task is essential in personalizing recommendations for users and improving the overall user experience. Some recent advancements in this field include the use of deep learning models[20], and the use of matrix factorization techniques[26], which are effective in dealing with the sparsity problem in recommendation systems. In line with the innovative recommendation paradigm of the LLM, we conducted experiments on a rating task that involved formulating two unique prompt types to elicit the results. We provide some sample prompts in Fig.2.

3.1.2 Sequential Recommendation. Sequential recommendation is a subfield of recommender systems that aims to predict a user's next item or action based on their past sequential behavior. It has received increasing attention in recent years due to its potential applications in various domains, such as e-commerce, online advertising, and music recommendation. In sequential recommendation,

Rating Prediction

zero-shot

How will user rate this product_title: "SHANY Nail Art Set (24 Famous Colors Nail Art Polish, Nail Art Decoration)" , and product_category: Beauty? (1 being lowest and 5 being highest) Attention! Just give me back the exact number a result , and you don't need a lot of text.

few-shot

Here is user rating history:
1. Bundle Monster 100 PC 3D Designs Nail Art Nailart Manicure Fimo Canes Sticks Rods Stickers Gel Tips, 5.0;
2. Winstonia's Double Ended Nail Art Marbling Dotting Tool Pen Set w/ 10 Different Sizes 5 Colors - Manicure Pedicure, 5.0;
3. Nail Art Jumbo Stamp Stamping Manicure Image Plate 2 Tropical Holiday by Cheeky&reg, 5.0 ;
4.Nail Art Jumbo Stamp Stamping Manicure Image Plate 6 Happy Holidays by Cheeky&reg, 5.0;
Based on above rating history, please predict user's rating for the product: "SHANY Nail Art Set (24 Famouse Colors Nail Art Polish, Nail Art Decoration)", (1 being lowest and5 being highest,The output should be like: (x stars, xx%), do not explain the reason.)

Sequential Recommendation

zero-shot

Requirements: you must choose 10 items for recommendation and sort them in order of priority, from highest to lowest. Output format: a python list. Do not explain the reason or include any other words.
The user has interacted with the following items in chronological order: ['Better Living Classic Two Chamber Dispenser, White', 'Andre Silhouettes Shampoo Cape, Metallic Black', ……, 'John Frieda JFHA5 Hot Air Brush, 1.5 inch'].Please recommend the next item that the user might interact with.

few-shot

Requirements: you must choose 10 items for recommendation and sort them in order of priority, from highest to lowest. Output format: a python list. Do not explain the reason or include any other words.
Given the user's interaction history in chronological order: ['Avalon Biotin B-Complex Thickening Conditioner, 14 Ounce', 'Conair 1600 Watt Folding Handle Hair Dryer', ……, 'RoC Multi-Correxion 4-Zone Daily Moisturizer, SPF 30, 1.7 Ounce'], the next interacted item is ['Le Edge Full Body Exfoliator - Pink']. Now, if the interaction history is updated to ['Avalon Biotin B-Complex Thickening Conditioner, 14 Ounce', 'Conair 1600 Watt Folding Handle Hair Dryer',……, 'RoC Multi-Correxion 4-Zone Daily Moisturizer, SPF 30, 1.7 Ounce', 'Le Edge Full Body Exfoliator - Pink'] and the user is likely to interact again, recommend the next item.

Direct Recommendation

zero-shot

Requirements: you must choose 10 items for recommendation and sort them in order of priority, from highest to lowest. Output format: a python list. Do not explain the reason or include any other words.
The user has interacted with the following items (in no particular order): [""Skin Obsession Jessner's Chemical Peel Kit Anti-aging and Anti-acne Skin Care Treatment"", 'Xtreme Brite Brightening Gel 1oz.',……, 'Reviva - Light Skin Peel, 1.5 oz cream']. From the candidates listed below, choose the top 10 items to recommend to the user and rank them in order of priority from highest to lowest. Candidates: ['Rogaine for Women Hair Regrowth Treatment 3- 2 ounce bottles', 'Best Age Spot Remover', ……""L'Oreal Kids Extra Gentle 2-in-1 Shampoo With a Burst of Cherry Almond, 9.0 Fluid Ounce""].

few-shot

Requirements: you must choose 10 items for recommendation and sort them in order of priority, from highest to lowest. Output format: a python list. Do not explain the reason or include any other words.
The user has interacted with the following items (in no particular order): ['Maybelline New York Eye Studio Lasting Drama Gel Eyeliner, Eggplant 956, 0.106 Ounce', ""L'Oreal Paris Healthy Look Hair Color, 8.5 Blonde/White Chocolate"", ……, 'Duo Lash Adhesive, Clear, 0.25 Ounce']. Given that the user has interacted with 'WAWO 15 Color Professionl Makeup Eyeshadow Camouflage Facial Concealer Neutral Palette' from a pool of candidates: ['MASH Bamboo Reusable Cuticle Pushers Remover / Manicure Pedicure Stick', 'Urban Decay All Nighter Long-Lasting Makeup Setting Spray 4 oz', ...,'Classic Cotton Balls Jumbo Size, 100 Count'], please recommend the best item from a new candidate pool, ['Neutrogena Ultra Sheer Sunscreen SPF 45 Twin Pack 6.0 Ounce', 'Blinc Eyeliner Pencil - Black', ...,'Skin MD Natural + SPF15 combines the benefits of a shielding lotion and a sunscreen lotion']. Note that the candidates in the new pool are not ordered in any particular way.

Figure 2: Example prompts of accuracy-based tasks on *Beauty* dataset. The black texts represent the description of the task, the red texts indicate the format requirements, the blue texts represent user historical information or few-shot information, and the gray texts indicate the current input.

researchers have proposed various methods, including recurrent neural networks[31], contrastive learning[68], and attention-based models[52], for capturing the temporal dependencies and patterns in user-item interactions. We have devised three distinct prompt formats for the sequential recommendation task family. These include: 1) direct prediction of the user's next item based on their interaction history, 2) selection of a possible next item from a list of candidates, where only one item is positive and based on the user's interaction history, and 3) prediction of whether a specific item will be the next one interacted with by the user, using their previous interaction history as a basis. These prompt formats have been designed to enhance the accuracy and effectiveness of sequential recommendations, and are grounded in rigorous academic principles. Examples of these prompts can be seen in Fig.2.

3.1.3 *Direct Recommendation.*

Direct Recommendation, also known as explicit feedback recommendation or rating-based recommendation, is a type of recommendation system that relies on explicit feedback from users in the form of ratings or reviews. Unlike other recommendation systems that rely on implicit feedback, such as

Explanation Generation

zero-shot

Here is user's interaction history:
1. Prolab Caffeine- Maximum Potency 200 mg 100 Tablets
2. DevaCurl Mist-er Right Lavender Curl Revitalizer 12.0 oz
3. TIGI Catwalk Curl Collection Curlesque Leave-In Conditioner, 7.27 Ounce
4. e.l.f. Pigment Eyeshadow, Naturally Nude, 0.05 Ounce
Help user generate a 5.0-star explanation about this product "SHANY Nail Art Set (24 Famouse Colors Nail Art Polish, Nail Art Decoration)" with around 10 words.

few-shot

Here are some recommended products and their corresponding explanations for user:
1. product "TIGI Catwalk Curl Collection Curlesque Curls Rock Amplifier, 5.07 Ounce, Packaging May Vary" and its explanation "One of the few things I have found that work for white people with curly hair"
2. product "DevaCurl Mist-er Right Lavender Curl Revitalizer 12.0 oz" and its explanation "it makes my hair greasy and gross"
Help user generate a 5.0-star explanation about this product "SHANY Nail Art Set (24 Famouse Colors Nail Art Polish, Nail Art Decoration)" with around 10 words?

Review Sumarization

zero-shot

Here are some summaries of user:
1. "Loving this sooo muchh"
2. "Amazing color"
3. "..Hong Kong Collection <3 OPI"
Write a short sentence to summarize the following product review from user: "So i was pretty excited that i got this in the mail, but seriously.....i think its just the color of mine, i don't know, not good cover stick... Received it sticking to the top...so basically it was broken when i opened it because of the air mail looks very pasty...very white i shall say...im never buying this product ever..". The sentence should be around 4 words.

few-shot

Here are some reviews and their corresponding summaries of user:
1. Review:"After watching kardashian episode back in 2009 kim mentioned OPI my private jet.. and i was like what is that? i looked it up online and LOVED it and i just got it 1 week shipping.. awesomee just love this color its sparkley brown and its turns black sometimes cool!! well for mee looolll LOVE this color <3 on my toes and fingers lol". Summary:"Loving this sooo muchh"
2. Review:"I love this and im glad im adding this to my collection! (: a nice top coat or alone very shimmery and very pretty, especially the top brush cap thing its silver than the original black! (: i recommend this". Summary:"Amazing color"
Write a short sentence to summarize the following product review from user: "So i was pretty excited that i got this in the mail, but seriously.....i think its just the color of mine, i don't know, not good cover stick... Received it sticking to the top...so basically it was broken when i opened it because of the air mail looks very pasty...very white i shall say...im never buying this product ever..". The sentence should be around 4 words.

Figure 3: Example prompts of explainability-oriented tasks on *Beauty* dataset. The black texts represent the description of the task, the red texts indicate the format requirements, the blue texts represent user historical information or few-shot information, and the gray texts indicate the current input.

user behavior or purchase history, direct recommendation systems are able to provide more personalized and accurate recommendations by taking into account the explicit preferences of users. For this task, we develop the item selection prompt that selects the most appropriate item from a list of potential candidates. These prompt formats are based on rigorous academic principles and aim to optimize the accuracy and relevance of recommendations. Examples of these prompts can be seen in Fig.2.

3.1.4 Explanation Generation. Explanation generation refers to providing users or system designers with explanations to clarify why such items are recommended. In this way, it enhances the transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction of recommendation systems. Furthermore, it facilitates system designers in diagnosing, debugging, and refining the recommendation algorithm. Large language models such as ChatGPT can use the vast amount of knowledge they contain to learn the user's interests through their historical interaction records and provide reasonable explanations for their behavior. Specifically, We ask ChatGPT model to generate a textual explanation to justify a user's preference towards a selected item as shown in Fig.3. For each category, additional auxiliary information such as the hint word and the star rating could be included.

3.1.5 Review Summarization. Automatic generation of summaries is becoming increasingly important in Natural Language Processing, as the demand for concise and easily comprehensible content continues to grow. Similar to the explanation generation task, we create two types of prompts: zero/few-shot prompts, and provide some example prompts in Fig.3.

3.2 Output Refinement

To ensure the diversity of generated results, ChatGPT incorporates a degree of randomness into its response generation process, which may result in different responses for the same input. However, when using ChatGPT for recommendation, this randomness can sometimes cause difficulties in evaluating the recommended items. While the format indicator in the prompt construction can partially alleviate this issue, in practical usage, it still cannot guarantee the anticipated output format. Therefore, we devise output refinement module to check the format of ChatGPT's output. If the output passes the format check, it is directly used as the final output. If not, it is modified based on pre-defined rules. If the format correction is successful, the corrected result is used as the final output. If not, the corresponding prompt is fed into ChatGPT for a re-recommendation until the format requirements are met. It is worth noting that different tasks have different output format requirements when evaluating ChatGPT. For example, for rating prediction, only a specific score is needed, whereas for sequential or direct recommendation, a list of recommended items is required. Particularly for sequence recommendation, it is challenging to feed all the items in the dataset to ChatGPT at once. As a result, ChatGPT's output may not correctly match the item set in the dataset. To address this issue, we introduce a text matching method based on similarity in the correction process to map ChatGPT's predictions back to the original dataset. Although this method may not perfectly reflect ChatGPT's ability, it can still indirectly demonstrate its potential in sequential recommendation.

4 EVALUATION

To evaluate ChatGPT, we conduct extensive experiments on the real-world Amazon dataset. Through the performance comparison with various representative methods and ablation studies on different tasks, we aim to answer the following research questions:

- **RQ1**: How does ChatGPT perform as compared with the state-of-the-art baseline models?
- **RQ2**: What is the impact of few-shot prompting on performance?
- **RQ3**: How do we design the human evaluation to assess explanation generation and summarization tasks?

4.1 Experimental Setup

4.1.1 Datasets. We conduct numerical and human evaluations on the real-world Amazon recommendation dataset. The Amazon dataset contains the customer review text with accompanying metadata on 29 categories of products. This paper focuses on evaluating the *Beauty* category.

4.1.2 Metrics. In numerical evaluations, we employ Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for rating prediction. And we adopt top-k Hit Ratio (HR@k), top-k Normalized Discounted Cumulative Gain (NDCG@k) for sequential recommendation and direct recommendation which are widely used in related works [19, 67]. Specifically, we report results on HR@{1,5,10}, NCGG@{5,10} for evaluation. Besides, n-gram Bilingual Evaluation Understudy (BLEU-n) and n-gram Recall-Roiented Understudy for Gising Evaluation (ROUGE-n) are used to evaluate the explanation generation and review summarization tasks. In human evaluations, we have designed and deployed a crowdsourcing task to assess the qualities of the generated explanations and review summaries. Through this task, we aim to accurately evaluate the effectiveness of the content by gathering feedback from a diverse range of human evaluators.

4.1.3 Implementation Details. In order to verify that we can directly apply the knowledge learned by ChatGPT to recommendation scenarios without the need for a large amount of task-specific data for training, we apply *gpt-3.5-turbo* to conduct few-shot and zero-shot experiments for the five tasks mentioned above. We collect n items that users have interacted with and k shots of historical records to enable ChatGPT to learn users' interests implicitly. In this experiment, we use the titles of the items as meta information, and set $n = 10$ and $k = 3$ due to the limitation of a maximum context length of 4096 tokens in ChatGPT. We ramdomly sample 100 records from the test set proposed by P5 [19] for evaluation. For direct recommendation, we set the number of negative samples to 99, thus forming a candidate list of length 100 with one positive item. Also, due to the addition of the candidate pool in the request, we set the number of shots to 1. For sequential recommendation, we input the user's historical interacted items in order and let ChatGPT predict the title of the next item that the user might interact with, and use BERT[13] to calculate the vector of the predicted title and compute the similarity between the predicted title vector and the title vectors of all items, and select the item with the highest similarity as the predicted item. For human evaluation on explanation generation and review summarization, we sample some results of different methods for each task, and each result will be scored and ranked by three human evaluators. After obtaining the manually annotated results, we will calculate the average top1 ratio and average ranking position of different methods to measure their generation performance.

Table 1: Performance comparison on rating prediction.

Methods	Beauty	
	RMSE	MAE
MF	1.1973	0.9461
MLP	1.3078	0.9597
ChatGPT(zero-shot)	1.4059	1.1861
ChatGPT(few-shot)	**1.0751**	**0.6977**

4.2 Baselines for multiple tasks

Following P5 [19], we gather a range of approaches that are representative of various tasks. For rating prediction, we employ MF [25] and MLP [10] as our baselines, both evaluated using mean square root loss. For direct recommendation, we use BPR-MF [43], BPR-MLP [10] and SimpleX [35] as baselines. For sequential recommendation, we adopt Caser [50], HGN [34], GRU4Rec [21], BERT4Rec [48], FDSA [63], SASRec [24] and S^3-Rec [67] as baselines for comparison. For explanation generation, we utilize Attn2Seq [15], NRT [30] and PETER [29] as baselines. For review summarization, we adopt pretrained T0 [45] and GPT-2 [41] as baselines. For more details, you can refer to P5 [19] or relevant articles.

4.3 Performance Comparison on 5 Tasks (RQ1&2)

4.3.1 Rating prediction. To evaluate the rating prediction performance of ChatGPT, zero-shot and few-shot prompts were employed, and the results obtained from the Beauty dataset were summarized in Tab.1. The results indicate that, for the seen category on the Beauty dataset, few-shot prompts outperform MF and MLP in terms of both MAE and RMSE. These results provide evidence supporting

Table 2: Performance comparison on sequential recommendation.

Methods	Beauty			
	HR@5	NDCG@5	HR@10	NDCG@10
Caser	0.0205	0.0131	0.0347	0.0176
HGN	0.0325	0.0206	0.0512	0.0266
GRU4Rec	0.0164	0.0099	0.0283	0.0137
BERT4Rec	0.0203	0.0124	0.0347	0.0170
FDSA	0.0267	0.0163	0.0407	0.0208
SASRec	0.0387	0.0249	0.0605	0.0318
S^3-Rec	0.0387	0.0244	0.0647	0.0327
P5-B	**0.0493**	**0.0367**	**0.0645**	**0.0416**
ChatGPT(zero-shot)	0.0000	0.0000	0.0000	0.0000
ChatGPT(few-shot)	0.0135	0.0135	0.0135	0.0135

the feasibility of utilizing a conditional text generation framework for rating prediction.

4.3.2 Sequential recommendation. To assess the sequential recommendation capability of ChatGPT, we conducted both zero-shot and few-shot experiments, the results of which are shown in Tab.2. We found that, compared to the baselines, ChatGPT's performance in the zero-shot prompting setup is considerably inferior, with all metrics being significantly lower than the baselines. However, under the few-shot prompting setup, while there is a relative improvement in performance, such as NDCG@5 surpassing GRU4Rec, ChatGPT is still generally outperformed by classical sequential recommendation methods in most cases. There are possibly two main reasons contributing to this outcome: First, during the prompting design process, all items are represented by their titles. Although this approach can alleviate the cold-start problem to some extent, it may cause ChatGPT to focus more on semantic similarity rather than the transition relationships between items, which are crucial for effective recommendations. Second, due to the length constraint of the prompts, it is not possible to input all items from the item set into ChatGPT. This leads to ChatGPT lacking constraints in predicting the title of the next item, resulting in generating item titles that do not exist in the dataset. Although it is possible to map these predicted titles to existing titles in the dataset through semantic similarity matching, our experiments show that this mapping does not result in significant gains. Therefore, for sequential recommendation tasks, merely employing ChatGPT is not a suitable choice. Further exploration is needed to introduce more guidance and constraints to help ChatGPT accurately capture historical interests and make reasonable recommendations within a limited scope.

4.3.3 Direct recommendation. Tab.3 illustrates the performance of ChatGPT on the direct recommendation task. Unlike the sequential recommendation setup, direct recommendation requires the recommendation model to select the most relevant item for the user from a limited-sized item pool. We observed that, when using zero-shot prompting, the recommendation performance is significantly inferior to supervised recommendation models. This can be attributed to the insufficient information provided to ChatGPT, resulting in an inability to capture user interests and generating more random recommendations. While few-shot prompting can improve ChatGPT's recommendation performance by providing some of the user's historical preferences, it still fails to surpass the baseline performance.

Table 3: Performance comparison on direct recommendation.

Methods	Beauty			
	HR@5	NDCG@5	HR@10	NDCG@10
BPR-MF	0.1426	0.0857	0.2573	0.1224
BPR-MLP	0.1392	0.0848	0.2542	0.1215
SimpleX	**0.2247**	**0.1441**	**0.3090**	**0.1711**
P5-B	0.1564	0.1096	0.2300	0.1332
ChatGPT(zero-shot)	0.0217	0.0111	0.0652	0.0252
ChatGPT(few-shot)	0.0349	0.0216	0.0930	0.0398

It is worth noting that we discovered during the experiments that the construction of the item pool, specifically whether the item pool is shuffled or not, has a considerable impact on the direct recommendation performance. In an extreme scenario where the ground truth item is placed at the first position in the item pool, we found that the evaluation metrics were approximately ten times higher than when the item pool was shuffled. This finding suggests that ChatGPT exhibits a positional bias for the input item pool within the prompt, tending to consider items towards the beginning of the pool as more important, and thus more likely to be recommended. This additional bias introduced by the language model renders using ChatGPT for direct recommendation a challenging endeavor.

4.3.4 Explanation Generation. In Tab.4, both zero-shot and few-shot prompts are used to evaluate ChatGPT's performance on explanation generation. From the metrics perspective, the P5 model has a better performance. As language models, P5 and ChatGPT have different design goals and application scenarios. P5 aims to generate explanatory language similar to known texts. Therefore, P5 focuses on learning text structure and grammar rules during training, making the generated results more standardized, as shown in Fig.4. In contrast, ChatGPT focuses more on language interaction and diversity. Its application scenario is usually to simulate human conversation, so it needs to consider multiple factors such as context, emotion, and logic when generating text to better express human thinking and language habits. This design is bound to make the text generated by ChatGPT more diverse and creative. Besides, P5 is fine-tuned on *Beauty* dataset while ChatGPT is utilized in a zero-shot or few-shot experimental seting. Therefore, it is understandable that ChatGPT may not perform as well as P5 in metrics. Hence, we introduce human evaluation to better measure the performance of different models in generating content.

4.3.5 Review summarization. We conduct zero-shot and few-shot experiments to evaluate ChatGPT's ability on review summarization, as shown in Tab.5. Similar to the explanation generation task, ChatGPT does not have an advantage in metrics evaluation. However, although the summary result of P5 has extracted some keywords, it has ignored relevant information from the entire review. In contrast, ChatGPT can generate more effective and meaningful summaries by deeply understanding and summarizing the reviews.

Explanation Generation Results

Reviews	Results
Absolutely great product. I bought this for my fourteen year old niece for Christmas and of course I had to try it out, then I tried another one, and another one and another one. So much fun! I even contemplated keeping a few for myself!	**Ground truth:** "Absolutely great product" **P5's output:** "great colors and great price for the price" **ChatGPT's output:** "Love this nail art set - perfect colors and variety!"
Love the colors. Didn't get any doubles. 1 bottle was not fully closed and the bottle chipped on the neck of the bottle. But being where the break was I just closed it and it is still usable. I wouldn't recommend this for painting your full nail (It is for art), but I would for stamping and nail art. Small brushes great for that. Not all work for stamping though, like the metallic ones.	**Ground truth:** "I wouldn't recommend this for painting your full nail (It is for art)" **P5's output:** "great price and great price and great price" **ChatGPT's output:** "SHANY's Nail Art Set is a must-have for creative nails."
Wow, this is the best deal I've seen on nail polish in a long time. You get so many vibrant beautiful colors to choose from. These are nail art brushes for fine detail. I love that you can get a whole kit for this price!	**Ground truth:** "this is the best deal I've seen on nail polish in a long time" **P5's output:** "great price and great quality and great price" **ChatGPT's output:** "SHANY's Nail Art Set is a must-have for stunning manicures."

Figure 4: Example explanation results of different models on *Beauty* dataset.

Table 4: Performance comparison on explanation generation (%).

Methods	Beauty			
	BLUE4	ROUGE1	ROUGE2	ROUGEL
Attn2Seq	0.7889	12.6590	1.6820	9.7481
NRT	0.8295	12.7815	1.8543	9.9477
PETER	1.1541	14.8497	2.1413	11.4143
P5-B	0.9742	16.4530	1.8858	11.8765
PETER+	**3.2606**	**25.5541**	**5.9668**	**19.7168**
ChatGPT(zero-shot)	0.0000	8.5992	0.6995	4.7564
ChatGPT(few-shot)	1.1967	11.4103	2.5675	5.9119

Table 5: Performance comparison on review summarization (%).

Methods	Beauty			
	BLUE4	ROUGE1	ROUGE2	ROUGEL
T0	1.2871	1.2750	0.3904	0.9592
GPT-2	0.5879	3.3844	0.6756	1.3956
P5-B	**2.1225**	**8.4205**	**1.6676**	**7.5476**
ChatGPT(zero-shot)	0.0000	3.8246	0.2857	3.1344
ChatGPT(few-shot)	0.0000	2.7822	0.0000	2.4328

As shown in Fig.5. Hence, we also conduct human evaluation in this task.

4.4 Human Evaluation (RQ3)

As shown in the experiments above, we conducted numerical evaluations on the explanation generation and review summarization tasks using the test set constructed by P5. However, the ground-truth explanations generated by P5 are not truly accurate because P5 extracts sentences from views commenting on one or more item feature words as users' explanations about their preferences. In that case, we designed human evaluations to better assess the performance of ChatGPT. Specifically, we randomly sample 20 prompts for explanation generation and 97 prompts for review summarization from the *Beauty* dataset and display every generated explanation or summary to several evaluators. The evaluators rank the results generated by ChatGPT, baseline, and ground truth for assessment. avg_top1_ration represents the proportion in which the prompt ranked first among the prompts. avg_position denotes the average position of sorting for each prompt.

For explanation generation task, as shown in Tab.6, the results of the four manual annotators have a certain degree of subjectivity, but the score distribution is relatively consistent, with a general consensus that the explanations generated by ChatGPT are clearer and more reasonable, even better than the ground truth. Meanwhile, P5's performance is the worst, with explanations tending towards a generic style and sentences that are not fluent. We can also draw the same conclusion from the examples in Tab.4. For review summarization task, we can find in Fig.5 that the contents summarized in P5 are too general and do not extract useful information. However, ChatGPT can truly understand the reviews and provide accurate summaries, rather than simply extracting a few keywords from the reviews. As shown in Tab.7, all annotators unanimously agree that

Table 6: Human evaluation for explanation generation on *Beauty* dataset.

Methods	Evaluators				avg_top1_ration	avg_position
	Eva_1	Eva_2	Eva_3	Eva_4		
Ground truth	25.0%	45.0%	45.0%	50.0%	38.0%	1.83
P5	0.0%	0.0%	0.0%	0.0%	0.0%	2.71
ChatGPT(zero-shot)	75.0%	55.0%	55.0%	50.0%	**62.0%**	**1.46**

Table 7: Human evaluation for review summarization on *Beauty* dataset.

Methods	Evaluators					avg_top1_ration	avg_position
	Eva_1	Eva_2	Eva_3	Eva_4	Eva_5		
Ground truth	12.5%	10.6%	8.7%	17.3%	22.1%	14.2%	2.91
P5	5.8%	0.0%	5.7%	11.5%	19.2%	8.5%	3.16
ChatGPT(zero-shot)	46.2%	37.5%	36.5%	45.2%	23.1%	37.7%	**1.90**
ChatGPT(few-shot)	35.6%	51.9%	49.0%	26.0%	35.6%	**39.6%**	2.01

Review Summarization Results

Reviews	Results
"I truly love this soap. I have very sensitive skin and this is one of the few soaps that doesn't dry out or break out my skin. Would recommend to others. It smells soft as well"	**Ground truth:** "Soap" **P5's output:** "Great soap" **ChatGPT's output:** "Gentle, effective soap recommended."
"These brushes are okay. I don't think they're anything special, but for the price their quality is okay. I don't know that I would buy them again though.".	**Ground truth:** "It's okay" **P5's output:** "Okay" **ChatGPT's output:** "Average brushes for price."
"Great to use after the microdermabrasion roller needle process. Skin absorbs it quickly. Face looks more rejuvenated in the morning."	**Ground truth:** "Works Well" **P5's output:** "Great" **ChatGPT's output:** "Quickly absorbed rejuvenating serum."
"you can see and feel that it's working. easy to use too. after a few times you'll get the hang of it."	**Ground truth:** "works!" **P5's output:** "works" **ChatGPT's output:** "Effective and user-friendly."

Figure 5: Example summarization results of different models on *Beauty* dataset.

ChatGPT has the best performance, surpassing ground truth and P5 by a large margin.

5 CONCLUSION AND FUTURE WORK

In this paper, we construct a benchmark to evaluate ChatGPT's performance in recommendation tasks and compare it with traditional recommendation models. The experimental results show that ChatGPT performs well in rating prediction but poorly in sequential and direct recommendation tasks, indicating the need for further exploration and improvement. Despite its limitations, ChatGPT outperforms state-of-the-art methods in terms of human evaluation for explainable recommendation tasks, highlighting its potential in generating explanations and summaries. We believe that our study provides valuable insights into the strengths and limitations of ChatGPT in recommendation systems, and we hope that it can inspire future research to explore the use of large language models to enhance recommendation performance. Moving forward, we plan to investigate better ways to incorporate user interaction data into large language models and bridge the semantic gap between language and user interests.

REFERENCES

[1] Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089* (2019).

[2] Pegah Malekpour Alamdari, Nima Jafari Navimipour, Mehdi Hosseinzadeh, Ali Asghar Safaei, and Aso Darwesh. 2020. A systematic study on the recommender systems in the E-commerce. *Ieee Access* 8 (2020), 115694–115716.

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[4] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, and J. Shawe-Taylor. 2003. Journal of Machine Learning Research 3 (2003) 1137–1155 Submitted 4/02; Published 2/03 A Neural Probabilistic Language Model. *JMLR.org* 6 (2003).

[5] Jesus Bobadilla, Santiago Alonso, and Antonio Hernando. 2020. Deep learning architecture for collaborative filtering recommender systems. *Applied Sciences* 10, 7 (2020), 2441.

[6] Jesús Bobadilla, Fernando Ortega, Abraham Gutiérrez, and Santiago Alonso. 2020. Classification-based deep neural network architecture for collaborative filtering recommender systems. (2020).

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, and D. Amodei. 2020. Language Models are Few-Shot Learners. (2020).

[8] Federica Cena, Luca Console, and Fabiana Vernero. 2021. Logical foundations of knowledge-based recommender systems: A unifying spectrum of alternatives. *Information Sciences* 546 (2021), 60–73.

[9] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849* (2018).

[10] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[11] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. *CoRR* abs/2205.08084 (2022).

[12] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007* (2023).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[14] B. Dhingra, L. Li, X. Li, J. Gao, and D. Li. 2016. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. (2016).

[15] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 623–632.

[16] Min Dong, Xianyi Zeng, Ludovic Koehl, and Junjie Zhang. 2020. An interactive knowledge-based recommender system for fashion product design in the big data environment. *Information Sciences* 540 (2020), 469–488.

[17] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *arXiv preprint arXiv:2303.14524* (2023).

[18] Achraf Gazdar and Lotfi Hidri. 2020. A new similarity measure for collaborative filtering based recommender systems. *Knowledge-Based Systems* 188 (2020), 105058.

[19] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[20] X. He, L. Liao, H. Zhang, L. Nie, and T. S. Chua. 2017. Neural Collaborative Filtering. *International World Wide Web Conferences Steering Committee* (2017).

[21] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[22] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[23] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745* (2023).

[24] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[25] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[26] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. IEEE, Computer Journal, 42(8), 30-37. *Computer* 42, 8 (2009), 30–37.

[27] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 35–42.

[28] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan. 2016. A Persona-Based Neural Conversation Model. *arXiv e-prints* (2016).

[29] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601* (2021).

[30] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 345–354.

[31] Z. C. Lipton, J. Berkowitz, and C. Elkan. 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. *Computer Science* (2015).

[32] Guoguang Liu. 2022. An ecommerce recommendation algorithm based on link prediction. *Alexandria Engineering Journal* 61, 1 (2022), 905–910.

[33] Y. Liu. 2019. Fine-tune BERT for Extractive Summarization. (2019).

[34] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.

[35] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1243–1252.

[36] Darshita Mittal, Sanyukta Shandilya, Dhruv Khirwar, and Archana Bhise. 2020. Smart billing using content-based recommender systems based on fingerprint. In *ICT Analysis and Applications: Proceedings of ICT4SD 2019, Volume 2*. Springer, 85–93.

[37] Cataldo Musto, Giovanni Semeraro, Marco De Gemmis, and Pasquale Lops. 2016. Learning word embeddings from wikipedia for content-based recommender systems. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*. Springer, 729–734.

[38] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023).

[39] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2022. "It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. In *Proceedings of the international AAAI conference on web and social media*, Vol. 16. 723–734.

[40] Yilena Pérez-Almaguer, Raciel Yera, Ahmad A Alzahrani, and Luis Martínez. 2021. Content-based group recommender systems: A general taxonomy and further improvements. *Expert Systems with Applications* 184 (2021), 115444.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.

[43] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[44] Fatemeh Rezaimehr and Chitra Dadkhah. 2021. A survey of attack detection approaches in collaborative filtering recommender systems. *Artificial Intelligence Review* 54 (2021), 2011–2066.

[45] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).

[46] A. See, P. J. Liu, and C. D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. (2017).

[47] Jagendra Singh, Mohammad Sajid, Chandra Shekhar Yadav, Shashank Sheshar Singh, and Manthan Saini. 2022. A Novel Deep Neural-based Music Recommendation Method considering User and Song Data. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 1–7.

[48] 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*. 1441–1450.

[49] Zhu Sun, Jie Yang, Kaidong Feng, Hui Fang, Xinghua Qu, and Yew Soon Ong. 2022. Revisiting Bundle Recommendation: Datasets, Tasks, Challenges and Opportunities for Intent-aware Product Bundling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2900–2911.

[50] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.

[51] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2021. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, Vol. 54. ACM New York, NY, USA, 1–23.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. *arXiv* (2017).

[53] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. 2017. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*. 1–6.

[54] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[55] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[56] Chuhan Wu, Fangzhao Wu, Tao Qi, Qi Liu, Xuan Tian, Jie Li, Wei He, Yongfeng Huang, and Xing Xie. 2022. Feedrec: News feed recommendation with various user feedbacks. In *Proceedings of the ACM Web Conference 2022*. 2088–2097.

[57] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.

[58] Yueqi Xie, Jingqi Gao, Peilin Zhou, Qichen Ye, Yining Hua, Jaeboum Kim, Fangzhao Wu, and Sunghun Kim. 2023. Rethinking Multi-Interest Learning for Candidate Matching in Recommender Systems. *arXiv preprint arXiv:2302.14532* (2023).

[59] Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1611–1621.

[60] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. (2019).

[61] Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, and Jie Yang. 2022. GreenPLM: Cross-lingual pre-trained language models conversion with (almost) no cost. *arXiv preprint arXiv:2211.06993* (2022).

[62] Feng Zhang, Victor E Lee, Ruoming Jin, Saurabh Garg, Kim-Kwang Raymond Choo, Michele Maasberg, Lijun Dong, and Chi Cheng. 2019. Privacy-aware smart city: A case study in collaborative filtering recommender systems. *J. Parallel and Distrib. Comput.* 127 (2019), 145–159.

[63] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, Xiaofang Zhou, et al. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation.. In *IJCAI*. 4320–4326.

[64] Yuhui Zhang, HAO DING, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language Models as Recommender Systems: Evaluations and Limitations. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*. https://openreview.net/forum?id=hFx3fY7-m9b

[65] Y. Zhang, S. Sun, M. Galley, Y. C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. (2019).

[66] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 43–51.

[67] 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*. 1893–1902.

[68] Peilin Zhou, Jingqi Gao, Yueqi Xie, Qichen Ye, Yining Hua, and Sunghun Kim. 2022. Equivariant Contrastive Learning for Sequential Recommendation. *arXiv preprint arXiv:2211.05290* (2022).