Exploring Fine-tuning ChatGPT for News Recommendation

Anonymous ACL submission

Abstract

News recommendation systems (RS) play a pivotal role in the current digital age, shaping how individuals access and engage with information. The fusion of natural language processing (NLP) and RS, spurred by the rise of large language models such as the GPT and T5 series, blurs the boundaries between these domains, making a tendency to treat RS as a language task. ChatGPT, renowned for its user-friendly interface and increasing popularity, has become a prominent choice for a wide range of NLP tasks. While previous studies have explored ChatGPT on recommendation tasks, this study breaks new ground by investigating its fine-tuning capability, particularly within the news domain. In this study, we design two distinct prompts: one designed to treat news RS as the ranking task and another tailored for the rating task. We evaluate ChatGPT's performance in news recommendation by eliciting direct responses through the formulation of these two tasks. More importantly, we unravel the pivotal role of fine-tuning data quality in enhancing ChatGPT's personalized recommendation capabilities, and illustrates its potential in addressing the longstanding challenge of the "cold item" problem in RS. Our experiments, conducted using the Microsoft News dataset (MIND), reveal significant improvements achieved by ChatGPT after fine-tuning, especially in scenarios where a user's topic interests remain consistent, treating news RS as a ranking task. This study illuminates the transformative potential of fine-tuning ChatGPT as a means to advance news RS, offering more effective news consumption experiences.

1 Introduction

In today's information-rich society, the accessibility of online news platforms Google News and Microsoft News has surged, offering users a vast array of news articles for consumption (Wu et al., 2020). However, the sheer daily volume of new news articles presents a challenge for users seeking content aligned with their interests (Lian et al., 2018). To address this issue, news RS play a crucial role in helping users discover articles relevant to their preferences. By effectively tailoring news recommendations, these systems not only enhance the user experience but also play a pivotal role in ensuring that individuals remain well-informed and engaged in a world inundated with information.

In the realm of news RS, models designed to comprehend article content and user interests are vital for delivering relevant recommendations. Techniques like the Gated Recurrent Unit (GRU) (Cho et al., 2014), Long-Short Term Memory (LSTM) (Staudemeyer and Morris, 2019), Convolutional Neural Networks (CNNs) (Chen, 2015), and attention mechanisms (Vaswani et al., 2017) have been popular choices for modeling user interests and comprehending article content (An et al., 2019; Wu et al., 2022, 2019a). However, these existing models are trained from scratch and may necessitate architectural modification when additional information is introduced. In response to these challenges, recent studies have shifted their focus toward using pre-trained language models. To leverage the pre-trained language models, researchers have introduced the concept of prompt learning (Jin et al., 2021), where specific prompts guide the output generation. Prompt learning makes it possible to generate outputs that adapt to the input and has been an effective approach for various NLP tasks (Jin et al., 2021), prompting researchers in the RS domain to recognize the potential of treating recommendation as a language task, harnessing the power of these techniques (Cui et al., 2022; Geng et al., 2022; Xu et al., 2023).

ChatGPT, developed by OpenAI, has recently attracted substantial attention for its remarkable performance in various NLP tasks. While some preliminary studies have been conducted to explore its potential in recommendation tasks (Zhang et al., 2023; Li et al., 2023b; Liu et al., 2023; Bang et al., 2023), OpenAI's decision to allow fine-tuning of ChatGPT through their provided API represents an uncharted territory in research. This fine-tuning capability, offering the potential to enhance ChatGPT's performance, has yet to be examined.

To bridge the research gap, this study explores using ChatGPT to improve personalized news recommendations through fine-tuning, capitalizing on its linguistic capabilities. Specifically, our study entails the fine-tuning of ChatGPT by formulating the news recommendation as direct ranking and rating tasks. Furthermore, we delve into the critical role played by the quality of fine-tuning data in augmenting ChatGPT's capability in delivering better recommendations. Our experiments, conducted on the MIND dataset, reveal substantial improvements in ChatGPT's performance after fine-tuning, particularly when users maintain consistent topic interests. Additionally, our findings offer promising insights, indicating that fine-tuned performance surpasses certain established baselines when the proportion of "cold" items in the testing set falls below a certain threshold when treating news RS as a ranking task.

2 Related Work

Sequential News Recommendation. Sequential news recommendation methods are centered around predicting a user's preference for a candidate article based on their prior reading behavior. They play a critical role in delivering timely and relevant content to users in dynamic news environments. The wealth of textual information within news articles has prompted the application of language techniques to extract valuable insights and understand user interests (An et al., 2019; Wu et al., 2022, 2019a). For instance, Okura *et al.* (Okura et al., 2017) introduced the use of a denoising autoencoder to analyze news representations and utilized a GRU network to model users' interests. An *et al.* (An et al., 2019) adopted CNN and attention mechanisms to learn news representations from attributes such as title, topic, and subtopic. The NRMS model proposed by Wu *et al.* (Wu et al., 2019b) explored news representation from titles using a word-level, multi-head, self-attention mechanism and an additive word-attention network. In this work, instead of constructing models from scratch for news recommendation, we focus on leveraging pre-trained large language models (LLMs), specifically ChatGPT, to enhance news RS.

Large Language Models and RS. Pre-trained language models like BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), trained on extensive datasets, have demonstrated remarkable adaptability to various downstream tasks, and the integration of prompt learning techniques (Cho et al., 2014) has further enhanced their performance. This transformation has not been confined to NLP alone, it has also extended its reach to the realm of RS. Increasingly, recommendation tasks are being approached as language tasks. Researchers have proposed a multitude of innovative approaches in this context, including the conversion of item-based recommendation into text-based tasks (Geng et al., 2022), the utilization of textual descriptions for understanding user behavior (Cui et al., 2022), personalized prompt learning for explainable recommendation (Li et al., 2022), the learning of LLM-compatible IDs for precise generation, and the adoption of flexible multi-modality modeling methodologies for RS (Geng et al., 2023). LLMs and prompt learning techniques have also found their way into the field of news recommendation. For instance, Zhang *et al.* (Zhang and Wang, 2023) employed prompt learning to address news recommendation by framing it as a slot filling task for [MASK] prediction, while Li *et al.* (Li et al., 2023a) formulated news recommendation as a direct generative recommendation task using a pre-trained T5 (Raffel et al., 2020) as the backbone.

ChatGPT has rapidly gained widespread popularity, prompting numerous studies to explore its capabilities and constraints. Qin *et al.* (Qin et al., 2023) conducted an evaluation of ChatGPT's performance across a spectrum of NLP tasks, while Bang *et al.* (Bang et al., 2023) comprehensively assessed its abilities in multitasking, multimodal applications, and multilingual contexts. On a parallel front, Liu *et al.* (Liu et al., 2023) constructed a benchmark to evaluate ChatGPT's proficiency in various RS tasks, including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Dai *et al.* (Dai et al., 2023) conducted experiments to enhance ChatGPT's recommendation capabilities by aligning it with traditional information retrieval ranking capabilities, including point-wise, pair-wise, and list-wise methods. While previous studies have emphasized ChatGPT's zero-shot or few-shot capabilities for RS, in this paper, we aim to conduct a preliminary evaluation of ChatGPT's potential in news recommendation, uniquely positioned after fine-tuning, which involves customizing ChatGPT for news recommendation using the MIND dataset. Furthermore, we seek to uncover how the quality of fine-tuning data samples impact ChatGPT's efficacy for news recommendations.

3 Recommendation Prompts

A distinguishing feature of ChatGPT is its ability to yield impressive results when using the released model and subsequently fine-tuning it, particularly in cases where data is limited. In this section, we delve into the assessment of ChatGPT's recommendation capabilities, focusing on its performance after fine-tuning. To explore fine-tuned ChatGPT's suitability for news RS, we meticulously designed prompts tailored to two common and critical tasks in the RS domain: ranking and rating tasks.

Ranking. The ranking task in RS involves generating an ordered list of items for a user based on their preferences, historical interactions, or contextual information. The primary goal is to present the most relevant items at the top of the list to enhance the user's experience. In the context of our study, the ranking task is exemplified by the prompt shown in Figure 1. For a user denoted as $u \in \mathcal{U}$, we provide the articles that the user most recently interacted with $\{h_1, h_2, \dots\} \in \mathcal{I}$. Simultaneously, we also supply a list of candidate articles, denoted as $\{c_1, c_2, \dots\} \in \mathcal{I}$. The system is asked to directly sort these candidate articles based on the user's preference, which are analyzed from the user's past interactions with articles.

Rating. The rating task in RS is centered around the prediction of a rating score to a specific item for a user and this task is prevalent in scenarios where users explicitly rate items, providing feedback on their preferences. In the standard rating task prompt we designed, shown in Figure 1, a user denoted as u is presented with the articles he/she most recently read $\{h_1, h_2, \dots\} \in \mathcal{I}$, along with a list of candidate articles, denoted as $\{c_1, c_2, \dots\} \in \mathcal{I}$. We then instruct the system to directly predict the rating scores for the candidate articles. The rating scale employed ranges from 1 to 5, where 5 denotes the highest score and 1 represents the lowest score. The system is encouraged to provide rating scores by making comparisons among the candidate articles.

4 Experiments

In this section, we conduct experiments to assess the effectiveness of fine-tuning ChatGPT. Through the performance comparison, we aim to answer the following research questions:

- **RQ1:** How does the performance of fine-tuned ChatGPT compare to that of ChatGPT in a zero-shot setting and other baseline models?
- **RQ2:** How does fine-tuned ChatGPT perform by prompting news RS as different tasks – ranking and rating?
- **RQ3:** How does the sample size used for fine-tuning affect the performance of fine-tuned ChatGPT?
- **RQ4:** What properties of data samples used for fine-tuning affect the performance of fine-tuned ChatGPT?

4.1 Experimental Settings

4.1.1 Dataset

For our experimental studies, we utilize the MIND dataset (Wu et al., 2020), which is a benchmark dataset in English for news recommendations. News recommendation presents unique challenges compared to other domains, as it may not always be highly personalized, and the nature of news is characterized by rapid changes (Dai et al., 2023). To comprehensively assess whether fine-tuning can enhance news recommendation performance, we conduct evaluations across two distinct groups of customers:

- **Group 1:** This group consists of 100 randomly selected customers whose clicked article in the impression aligns with the topics they have previously read, i.e., the clicked article is from a topic that they have previously read.
- **Group 2:** This group consist of 100 randomly selected customers whose clicked article in the impression is from a different topic than those they have previously read.

The division of customers into these two groups allows us to capture the dual nature of news recommendation: personalized recommendation that align with user interests and the challenges of offering recommendations outside a user's established

Ranking

A user read the following 5 articles:
1. A white restaurant owner admitted to enslaving a black man. He got 10 years in prison.
2. Election Day 2019: Most Important Races and What to Watch For
...
5. Four ways Trump's tax returns could come to light
Based on the user's interest from the articles he read before, rank the 5 candidates delimited by triple backticks.
```A. Maurkice Pouncey suspended three games, B. Johnson opens door to subpoenaing whistleblower, Schiff, Bidens, ...```
Do not explain reasons in the response and only return a python list of the first letter for each article.

Rating

A user read the following 5 articles:
1. A white restaurant owner admitted to enslaving a black man. He got 10 years in prison.
2. Election Day 2019: Most Important Races and What to Watch For
...
5. Four ways Trump's tax returns could come to light
Based on the user's interest from the articles he read before, please predict the user's ratings for the 5 candidates delimited by triple backticks.
```A. Maurkice Pouncey suspended three games, B. Johnson opens door to subpoenaing whistleblower, Schiff, Bidens, ...```
(1 being the lowest and 5 being the highest).

Figure 1: Example prompts of both ranking and rating tasks, with the system content 'You are a news recommender now'.

preferences. This division also enables evaluation to determine if fine-tuning ChatGPT could enhance news recommendation across diverse scenarios.

4.1.2 Baselines

We compare the performance of fine-tuned ChatGPT with the following baseline models:

- **NAML** (Wu et al., 2022): models users' and articles' representations via multi-view self-attention.
- **LSTUR** (An et al., 2019): captures a user's interests by modeling both his long- and short-term preferences.
- **NRMS** (Wu et al., 2019b): models users' and articles' representations via multi-head self-attention network.
- **Popularity** (Ji et al., 2020): recommends the top-k popular articles.
- **Zero-shot**: recommends the top-k articles from the candidate pool, using ChatGPT's zero-shot capabilities.

4.1.3 Metrics

In numerical evaluations, we adopt metrics top-k Normalized Discounted Cumulative Gain (NDCG@k) and Mean Reciprocal Rank (MRR@k) to assess the news recommendation performance.

4.1.4 Implementation Details.

We evaluate using ChatGPT for news recommendation using *gpt-3.5-turbo* for fine-tuning and zero-shot experiments.

It is noteworthy that when utilizing zero-shot performance, the output generated by ChatGPT may not always adhere to the desired format requirements. To ensure compliance with format requirements and to meet the criteria for evaluation, a regeneration approach is employed, iteratively generating responses until the required format is met. Furthermore, in the context of the rating task within the zero-shot setting, where diverse rating values are anticipated to reflect comparative preferences, an additional format requirement is introduced. This requirement instructs ChatGPT to predict distinct rating scores for various candidates.

Additionally, it's important to mention that the data samples used for fine-tuning remain consistent for ranking and rating tasks, separately for Group 1 and Group 2 customers. This consistency is crucial in ensuring fair and meaningful comparisons. The individuals in the training data do not overlap with those in the test data, although articles in the training set may appear in the test data. The fine-tuning epoch and other hyper-parameters are automatically selected by OpenAI based on the size of fine-tuning dataset. During the fine-tuning process, with a fixed prompt, fixed group, and a fixed fine-tuning sample size, we conduct five independent experiments using five independent training datasets. This approach evaluates the reliability of our findings.

5 Performance Evaluations

5.1 RQ1&2: Performance Comparison

Table 1 presents the performance results for various models, including baseline models, zero-shot ChatGPT using the news RS ranking and rating task formulations, and fine-tuned ChatGPT with these same formulated tasks. We conduct separate evaluations for Group 1 and Group 2 customers, and here are our observations:

The first 4 baselines exhibit no variance, as they are intentionally trained with a significantly larger number of data samples than the fine-tuning sample sizes, aiming at establishing them as performance upper bounds for a more rigorous and superior baseline comparison. The popularity baseline stands out as a strong baseline for news recommendation, which is in line with the findings of many other research works (Dai et al., 2023; Qi et al., 2021). It consistently outperforms the zero-shot ChatGPT and other deep neural-based models for both Group 1 and Group 2 users. This is particularly evident when readers have engaged with articles from diverse topics. These findings underscore the distinct nature of news recommendation, where user behavior may not always align closely with personalized recommendations, as seen in other domains like e-commerce (Jonnalagedda et al., 2016; Yang, 2016).

In the zero-shot setup, ChatGPT's performances lag behind that of popularity-based models. When users' topic interests change, as observed in Group 2, ChatGPT's zero-shot performance using ranking task formulation falls short of all baseline models. This suggests that ChatGPT's strength lies in semantic understanding and its tendency to recommend articles similar to those previously read by users. However, an intriguing finding is that for the rating task the zero-shot performance on Group 1 and Group 2 customers are similar to each other. One possible explanation is that, within the zero-shot setup, ChatGPT interprets the rating task in a manner akin to sentiment classification, where a rating of 1 represents strongly negative and 5 indicates strongly positive. To substantiate this hypothesis, we introduce a similar prompt as the rating task, instructing ChatGPT to generate relative sentiment scores for candidate articles directly. The resulting zero-shot performance, treated as a sentiment classification task, is illustrated in Figure 2. Our findings provide empirical support for our hypothesis that within the zero-shot context, ChatGPT perceives the rating task as a form of sentiment classification for candidates, a perspective that exhibits notable zero-shot performance as demonstrated in previous work (Wang et al., 2023). This interpretation results in the noteworthy performance observed with Group 2 customers, and the observation also demonstrates the importance of proper prompt-based task formulation when fine-tuning ChatGPT for downstream tasks.

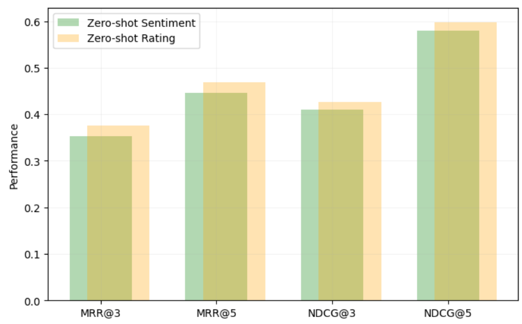

Figure 2: Comparison of zero-shot ChatGPT performance between sentiment classification for candidate article and the rating task among Group 2 customers. Five independent experiments are conducted, and the figure shows the average performances.

Under the fine-tuning setup, there is a notable improvement in performance compared to zero-shot with ranking task, particularly in Group 1. This improvement may be attributed to the fact that fine-tuning not only enhances ChatGPT's semantic understanding but also makes more effective use of position information. During the fine-tuning process, the clicked articles are consistently placed at the first position in the generated ranking list response, regardless of their original position in the provided candidate list. This allows the model to better exploit the positional information. In contrast, for the rating prompt, the five scores may manifest at various positions within the generated responses. For customers in Group 2, fine-tuning also leads to improvements compared to zero-shot, albeit not as substantial as observed in Group 1, even when using the same fine-tuning sample sizes. One possible explanation is that ChatGPT tends to recommend articles from diverse topics after fine-tuning. However, within the provided candidate articles, multiple options may satisfy this diversity

Method	Group 1				Group 2			
	MRR@3	MRR@5	NDCG@3	NDCG@5	MRR@3	MRR@5	NDCG@3	NDCG@5
NAML	0.4086	0.4882	0.4689	0.6136	0.3614	0.4599	0.4123	0.5917
LSTUR	0.4129	0.5041	0.4614	0.6256	0.4128	0.4978	0.4596	0.6213
NRMS	0.4263	0.4984	0.4913	0.6219	0.3972	0.4911	0.4438	0.6151
Popularity	0.4764	0.5423	0.5355	0.6551	**0.5264**	**0.5826**	**0.5829**	**0.6855**
Zero-shot (Ranking)	0.4446±0.0023	0.5258±0.0021	0.4936±0.0023	0.6420±0.0016	0.2935±0.0023	0.3930±0.0021	0.3604±0.0024	0.5415±0.0015
Zero-shot (Rating)	0.3735±0.0029	0.4540±0.0024	0.4428±0.0029	0.5886±0.0018	0.3754±0.0112	0.4691±0.0079	0.4266±0.0133	0.5984±0.0063
Fine-tuned (Ranking)	**0.5278±0.0719**	**0.5928±0.0598**	**0.5755±0.0682**	**0.6930±0.0454**	0.3802±0.0279	0.4690±0.0221	0.4372±0.0298	0.5989±0.0169
Fine-tuned (Rating)	0.3794±0.0249	0.4659±0.0223	0.4494±0.0234	0.5969±0.0168	0.3637±0.0209	0.4538±0.0199	0.4261±0.0245	0.5865±0.0145

Table 1: The news recommendation performance on customers. Bold numbers indicate the best performance. 5 independent experiments are conducted for zero-shot ranking and rating, while 25 independent experiments are conducted for fine-tuning setting. The statistical significance was assessed using the Student's t-test, with a significance level of $p < 0.05$.

requirement. Without knowledge of the popularity of these articles, ChatGPT might randomly select one to fulfill the diversity requirement.

However, under the fine-tuning setup, the performances are similar when using rating task, whether applied to Group 1 or Group 2, as compared to zero-shot approach. This might be attributed to the fact that, even during fine-tuning, while semantic understanding can be improved, the model's capacity for handling numerical comparison remains relatively unchanged. Additionally, the rating task lacks the advantage of utilizing positional information from the generated responses during fine-tuning, unlike the ranking task. Furthermore, the rating prompt necessitates the assignment of scores for all candidate simultaneously, making the rating task more challenging. Lastly, it's worth noting that different customers may use the rating scale differently (i.e., the system must learn user biases). This finding that ranking outperforms the rating task aligns with prior research, particularly comparing point-wise and list-wise ranking (Dai et al., 2023).

5.2 RQ3: Performance Under Different Fine-tuning Sample Sizes

Our experiments reveal a notable performance enhancement in ChatGPT when using ranking tasks after fine-tuning. In this subsection, we investigate how the quantity of fine-tuning samples affects fine-tuned ChatGPT's recommendation performance using ranking task. We conduct experiments varying the sample size within the range {50, 80, 100, 120}. For each sample size, fine-tuning is performed independently five times, utilizing distinct fine-tuning data samples. The results are presented in Figure 3, demonstrating the performance differences across sample sizes, as measured by NDCG@3.

Our observations indicate that the average performance (i.e., NDCG@k) remains consistent across different fine-tuning sample sizes, suggesting that the quantity of fine-tuned data samples does not significantly affect the performance of fine-tuned ChatGPT (the p-value from a one-way ANOVA, testing the equality of means, exceeds 0.1). Additionally, when the same number of samples was used for fine-tuning, there was variability in performance on the same test set. This not only reaffirms that performance is not solely contingent on the fine-tuning sample sizes but also emphasizes our interest in identifying the quality of data samples that enhance performance.

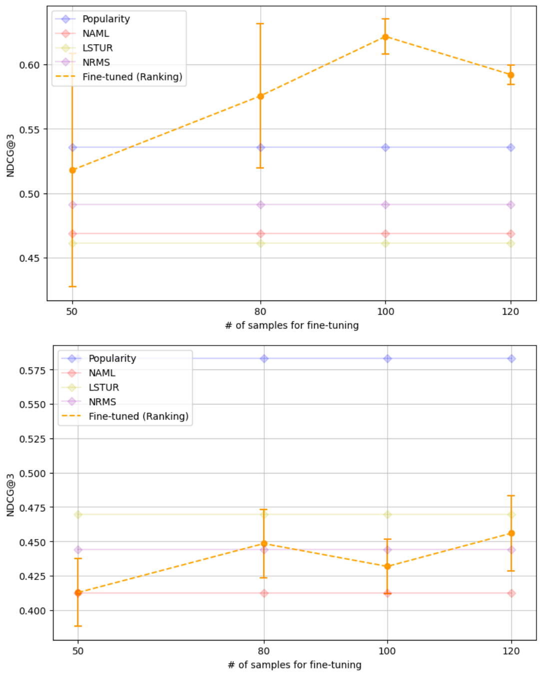

Figure 3: Recommendation performances with different quantities of fine-tuning samples. The first subfigure is for Group 1 while the second is for Group 2 readers.

5.3 RQ4: Quality of Fine-tuning Samples

ChatGPT, when using ranking tasks after fine-tuning, even outperforms the popularity-based model for Group 1 users. In this subsection, we delve deeper into the realm of ranking tasks and aim to detect specific factors that boost fine-tuned ChatGPT's performance in news recommendation.

The intriguing observation that fine-tuned ChatGPT, using ranking tasks, can even outperform the popularity-based model for Group 1 users motivates us to analyze the impact of the proportion of top-ranked articles in the test set that were also present in the training set. A higher proportion indicates more overlap between the articles users engaged with during training and testing. The first subfigure in Figure 4 illustrates that fine-tuned ChatGPT's performance for Group 1 customers shows improvement as the overlap ratio increases toward a certain threshold with statistical significance (p-value < 0.05). This finding may offer a possible explanation for the model's superior performance compared to the popularity-based model for Group 1 users. When ChatGPT encounters articles during testing that it has previously interacted with during the fine-tuning process, it might discern implicit popularity signals from these articles, utilizing the positional information derived from the ranking task. Group 1 users, with their consistent interests and ChatGPT's proficiency in textual understanding, benefit from this approach, leveraging both positional information and semantic understanding. Notably, this factor does not yield statistically significant effects for Group 2 users (p-value > 0.1).

We also investigate the impact of the presence of "cold" articles in the candidates during testing. A candidate article is labeled as "cold" if it is not part of the fine-tuning samples. As observed in the last two subfigures of Figure 4, we find that the proportion of cold articles significantly influences fine-tuned ChatGPT's performance (p-values below 0.05 for Group 1 and below 0.1 for Group 2). In general, we notice that as the ratio of "cold" articles in the test set increases, the fine-tuned ChatGPT's performance decreases. The observation that fine-tuned ChatGPT can surpass specific baselines, which are trained with more data samples and fewer "cold" items during evaluation, underscores ChatGPT's potential in addressing the "cold" item challenge in RS, as long as the ratio of "cold" articles remains within a particular threshold, as shown in Figure 4.

5.4 Computational Cost

In our experiments, fine-tuning ChatGPT with a maximum of 120 samples typically took around 30 minutes to complete. This is done with an average of approximately 310 input tokens and using the default number of epochs once the fine-tuning process began.

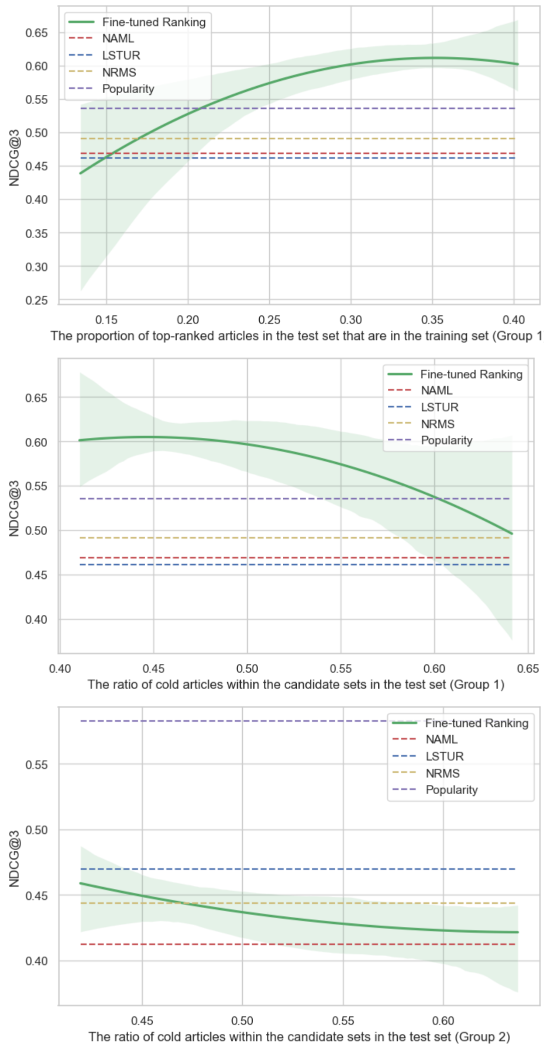

Figure 4: Influence of overlap ratio and "cold" articles on fine-tuned ChatGPT's news recommendation performance using the ranking prompt. Fine-tuned ChatGPT outperforms all baselines for Group 1 readers when the ratio of "cold" articles < 0.6, and surpasses the NAML baseline for Group 2 readers when the ratio < 0.45.

6 Conclusion

In this study, we conduct experiments that showcase the substantial benefits of fine-tuning ChatGPT for news recommendation. This may seem like a trivial task. However, as we have shown in the research, the performance of fine-tuning depends on several factors such as the topic alignment, prompt formulation, sample sizes, and the quality of fine-tuning samples. More specifically, we compare the effectiveness of ranking and rating tasks for fine-tuning ChatGPT, and our results indicate that ranking consistently outperforms rating by leveraging both positional cues from the generated responses during fine-tuning and semantic understanding. The challenges of rating tasks become evident as ChatGPT struggles with making numerical comparisons when tasked with generating ratings for all candidates simultaneously. Additionally, ChatGPT sometimes interprets the rating task as a sentiment classification task in the zero-shot mode, particularly for Group 2 customers. Moreover, we delve into the factors influencing ChatGPT's ranking performance after fine-tuning. Our investigation unveils the degree of overlap between the articles users interacted with during both training and testing is a significant factor when user interests remain consistent. One of the most promising findings in our study is ChatGPT's potential to address the "cold" item issue in RS. Despite competing with baselines trained on larger datasets with fewer "cold" items during evaluation, fine-tuned ChatGPT consistently outperforms them within a specific threshold of ratio of "cold" items. This observation underscores ChatGPT's capacity to mitigate the "cold" item issue to enhance RS.

For future studies, we envision several promising research directions. Given the fundamental role of popularity in news recommendation, a notable avenue for future exploration is the effective incorporation of popularity-related information into prompts. Additionally, enhancing news recommendation for users when their interests undergo shifts, particularly via fine-tuning ChatGPT, holds significant potential for further advancement.

7 Limitations

In this study, our primary focus revolves around the recommendation performance of ChatGPT, particularly after fine-tuning with diverse recommendation prompts, including ranking and rating. We assess its efficacy across two distinct customer groups—those with consistent interests and those with varying topic preferences. However, it is crucial to acknowledge certain limitations within the scope of our investigation.

One notable limitation is that our designed prompts involve presenting candidate articles to ChatGPT. While this approach allows us to evaluate recommendation performance, it does not directly address the potential ethical considerations or the risk of hallucination issues that might arise if ChatGPT were tasked with generating recommendations without specific article candidates. This avenue remains unexplored within the confines of our current study and presents an opportunity for future research to delve into the intricacies of using ChatGPT in a more unconstrained recommendation setting. An additional limitation is the observation that, even after fine-tuning, ChatGPT may exhibit suboptimal performance for users with diverse interests. However, no specific potential solution was put forth in this study.

References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345.

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*.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084*.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. *arXiv preprint arXiv:2305.02182*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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). *RecSys*.

Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. Vip5: Towards multimodal foundation models for recommendation. *EMNLP*.

Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2020. A re-visit of the popularity baseline in recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1749–1752.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2021. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Nirmal Jonnalagedda, Susan Gauch, Kevin Labille, and Sultan Alfarhood. 2016. Incorporating popularity in a personalized news recommender system. *PeerJ Computer Science*, 2:e63.

Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *arXiv preprint arXiv:2202.07371*.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023a. Pbnr: Prompt-based news recommender system. *arXiv preprint arXiv:2304.07862*.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023b. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv preprint arXiv:2306.10702*.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

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

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1933–1942.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Pp-rec: News recommendation with personalized user interest and time-aware news popularity. *arXiv preprint arXiv:2106.01300*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ralf C Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019b. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394.

Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. 2022. Is news recommendation a sequential recommendation task? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386.

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*, pages 3597–3606.

Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. 2023. Openp5: Benchmarking foundation models for recommendation. *arXiv:2306.11134*.

JungAe Yang. 2016. Effects of popularity-based news recommendations (“most-viewed”) on users’ exposure to online news. *Media Psychology*, 19(2):243–271.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*.

Zizhuo Zhang and Bang Wang. 2023. Prompt learning for news recommendation. *arXiv preprint arXiv:2304.05263*.