Exploring chatgpt's performance in news recommendation: a multi-faceted analysis

Hussein Ali Hussein Al Naffakh[1], *Ahmed* Dheyaa Radhi[2*], *Baqer* A Hakim [3], *AL-Ibraheemi* Fuqdan [3] and *Bourair* Al-Attar [4]

[1] College of Health and Medical Techniques, University of Alkafeel, Al-Najaf, Iraq
[2] College of Pharmacy, University of Al-Ameed, Karbala PO Box 198, Iraq
[3] College of Dentistry, University of Al-Ameed, Karbala PO Box 198, Iraq
[4] College of Medicine, University of Al-Ameed, Karbala PO Box 198, Iraq

Abstract. Using complex language models has been a common strategy as personalized news recommendation systems are adopted by online news sites more and more. Before the epoch of GPT-3, news recommendation systems underwent a progression from rule-based and collaborative filtering approaches in the pre-2010s, through the integration of neural networks in the 2010s, to the emergence of earlier iterations of large language models like GPT-2 in 2019. Pre-trained language models have ushered in a new era of recommendation paradigms, thanks to the emergence of huge language models like GPT-3 and T-5. With its easy-to-use interface, ChatGPT is becoming more and more popular for text-based jobs. Focusing on news provider fairness, individualized news recommendations, and fake news identification, this study starts an inquiry into ChatGPT's efficacy in news recommendations. We acknowledge that ChatGPT's output sensitivity to input phrasing is a limitation, and our goal is to investigate these limitations from each angle. We also investigate whether certain prompt formats can help to alleviate these constraints or if more research is needed. To go beyond set assessments, we create a webpage where ChatGPT's performance on the examined activities and prompts is tracked once a week. Through the use of big language models, this work seeks to improve news recommendation performance and stimulate more research in this area.

1 Introduction

Online news sources like Google News and Microsoft News have drawn individuals looking to consume news digitally in our information-rich, modern culture [1]. However, finding information that speaks to their particular interests might be extremely difficult for consumers due to the daily flood of fresh news stories [2]. To overcome this difficulty, news recommendation systems (RS) become essential instruments that assist readers in finding relevant articles to read. Language model techniques such as Gated Recurrent Unit (GRU) [3], Long-Short Term Memory (LSTM) [4], Convolutional Neural Networks (CNNs) [5], and attention mechanisms [6] have become popular due to the text-rich nature of news articles. These techniques are useful for understanding article content and modeling users' interests [7].

In addition, it has been demonstrated that pre-trained language models and prompt learning strategies work well in a variety of language tasks [8], which has led recommendation systems (RS) researchers to approach recommendation as a language task to fully utilize these strategies [9]. This research aims to evaluate ChatGPT, a well-known language model created by OpenAI, in the context of news RS tasks. Our research focuses on three main areas: news provider fairness, false news detection, and personalized news recommendation, in light of ChatGPT's achievements in various natural language processing (NLP) tasks and the growing recognition of recommendation as a language-related problem. Our objective is to identify constraints in ChatGPT's response generation from each viewpoint and investigate the possible efficacy of particular prompt forms or specifications to alleviate these constraints. Furthermore, we hope to draw attention to areas that may require additional investigation by future researchers, acknowledging that some restrictions would be difficult to overcome with prompt design alone.

* Corresponding author: ahmosawi@alameed.edu.iq

Additionally, we aim to investigate strategies for enhancing the performance of language models, such as ChatGPT, in news recommendation system (RS) tasks.

The research begins with a related work to lay the groundwork and then moves on to an experiment section that details the research design and data collection. The Evaluation Performance section offers the findings, and the commentary explains them in the context of the larger research. The conclusion summarizes major findings and suggests future study directions.

2 Related Work

News Recommendation Many Natural Language Processing (NLP) approaches are used in news recommendation systems (RS) nowadays. These techniques include CNNs [5], GRU networks, attention mechanisms [10], and denoising auto-encoders [11]. By understanding news content, these methods seek to predict individuals' interests based on their reading habits [12, 13]. Personalized recommendations and content comprehension are important, but addressing societal concerns related to news RS is just as important. These concerns include user-side fairness [14], provider-side fairness [5], filter bubbles, echo chambers [9], the spread of fake news [15], popularity prejudice [1], and popularity bias. This study examines ChatGPT's usefulness in mitigating provider bias and fake news issues, in addition to assessing its zero-shot performance in personalized recommendation tasks. With the help of this inquiry, we hope to shed light on the wider societal ramifications of incorporating ChatGPT into news reporting.

Pre-trained Language Models Trained on large datasets, pre-trained language models such as BERT [16] and GPT [17] have proven adaptable to various downstream tasks. The incorporation of fast learning strategies [3] has improved their output even further. This accomplishment has led to a paradigm change in recommendation systems (RS), where tasks linked to recommendations are now considered language-related activities. Scholars have proposed several tactics, including using textual descriptions to comprehend user behavior and transforming item-based recommendations into text-based activities [18]. They have also adopted flexible text-to-text techniques for RS [19], converted user behavior into text-based inquiries [20], and investigated tailored prompt learning for transparent suggestions [21]. Examining ChatGPT's zero-shot performance in news recommendation tasks, this study leverages its advantages as a pre-trained language model.

ChatGPT With ChatGPT's quick rise to popularity, many studies have examined its features and limitations. While Bang et al. [22] undertake a comprehensive technical study, examining ChatGPT's multitasking, multimodal, and multilingual applications, Qin et al. [23] test the platform's performance across a range of Natural Language Processing (NLP) tasks. Zhou et al. [24] examine the moral implications of ChatGPT use, and Liu et al. [25] provide a standard by which to measure ChatGPT's performance in recommendation system (RS) tasks such as rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Even though its drawbacks, like bias and the capacity to spread false information, are acknowledged [26], With a particular focus on provider bias and fake news identification, our research delves into the societal ramifications of using ChatGPT for news recommendation. Our goal is to find possible prompt formats that could allay these worries or highlight areas that need more investigation.

3 ChatGPT Evaluation

In this section, we will be using zero-shot techniques to evaluate the effectiveness of ChatGPT in providing news suggestions. Our focus will be on three crucial tasks: offering personalized recommendations, ensuring impartiality among news sources, and verifying the accuracy of the answers provided. To begin with, we will test simple queries to identify any limitations in ChatGPT's responses. Based on the results, we will provide further prompts to address any shortcomings or highlight the need for additional research when using language models like ChatGPT for news recommendations. We have also made our programs and prompts available on a GitHub repository to improve reproducibility. For our investigation, we will be using data samples from the Microsoft News Dataset (MIND) [1].

4 Experiment

Experiments are conducted in this part to assess the effectiveness of ChatGPT fine-tuning. The main research question that this performance comparison aims to answer is: How much does the performance of ChatGPT under zero-shot settings and different baseline models differ from that of ChatGPT that has been fine-tuned?

4.1 Experimental Settings

Dataset We use the MIND dataset [1] in the context of our experimental investigations. When considering news recommendations in comparison to other domains, there are unique obstacles because news material is dynamic and subject to frequent changes [27]. To fully assess the effectiveness of fine-tuning in enhancing the efficiency of news recommendations, assessments are carried out methodically over two distinct consumer segments:

Group 1: This cohort consists of 100 randomly selected customers whose clicked article in the impression corresponds to a topic they have previously interacted with, i.e., the clicked article is related to a topic they have previously looked at.

Group 2: This set consists of 100 randomly chosen consumers whose clicked article in the impression is related to a different topic than the ones they have already read.

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).

Fig.1. shows example prompts with the system content for both rating and ranking.

Customers are divided into these two categories to enable a comprehensive investigation of the two aspects of news recommendation: tailored recommendations that correspond to specific user preferences and the difficulties of providing recommendations that go beyond a user's predetermined preferences. This division strengthens the evaluation procedure, which seeks to determine whether using ChatGPT's fine-tuning capabilities may improve news recommendation performance in a variety of contexts.

Baselines The following baseline models are used to compare the performance of ChatGPT after it has been adjusted:

NAML [13]: This model uses multi-view self-attention to represent the features of people and articles.

LSTUR [7]: By modeling a user's short- and long-term preferences, LSTUR can capture their interests.

NRMS [28]: Using a multi-head self-attention network, NRMS mimics the features of users and articles.

Popularity [29]: The most well-liked items are suggested by this baseline.

Zero-shot: Using ChatGPT's zero-shot features, this baseline suggests the top-k articles from the candidate pool.

Measures In quantitative evaluations, we use Mean Reciprocal Rank (MRR@k) and top-k Normalized Discounted Cumulative Gain (NDCG@k) as measures to assess news recommendation performance.

Implementation Specifics We utilize ChatGPT, get-3.5-turbo, for both fine-tuning and zero-shot tests to evaluate news recommendations. It is important to remember that ChatGPT's output might not always follow the formatting specifications while using zero-shot performance. A regeneration strategy is used to guarantee adherence to format requirements and meet evaluation criteria. This is producing responses iteratively until the desired format is obtained. Furthermore, an extra format requirement is introduced for the rating task in the zero-shot setting, where a variety of rating values are anticipated to indicate comparative preferences. This clause tells ChatGPT to forecast different rating scores for different applicants.

In addition, it is important to note that the data samples used for fine-tuning are the same for both ranking and rating jobs, but for consumers in Group 1 and Group 2, respectively. Ensuring homogeneity is crucial for ensuring fair and significant comparisons. Interestingly, whereas articles from the training set may repeat in the test data set, the persons in the training data set do not overlap with those in the test data set. OpenAI uses the size of the fine-tuning dataset to automatically compute the fine-tuning epoch and other hyperparameters. Using five different training

datasets, we conduct five independent experiments during the fine-tuning process, each with a fixed prompt, fixed group assignment, and constant fine-tuning sample size. This methodological approach helps evaluate the robustness and dependability of our results.

5 Evaluations of Performance

The performance results for several models are shown in Table 1. These results include baseline models, zero-shot ChatGPT using news RS ranking and rating tasks, and optimized ChatGPT using the same designed tasks. Different assessments are carried out for clients in Groups 1 and 2, resulting in the following findings:

The first four baselines, which were purposefully trained with a substantially higher number of data samples than the fine-tuning sample sizes and placed as performance upper bounds for a more rigorous baseline comparison, show no fluctuation. Notably, the popularity baseline turns out to be a strong standard for news recommendation, which is consistent with findings from other studies [27, 30]. It consistently performs better than other deep neural-based models and zero-shot ChatGPT for both Group 1 and Group 2 users. This is especially noticeable when readers have engaged with articles on a variety of subjects. These results highlight how unique news recommendations are, as user behavior might not always match personalized recommendations, as seen in other industries like e-commerce [31, 32].

	Group 1				Group 2			
Method	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. Customer performance for news recommendations.

ChatGPT performs worse in the zero-shot setting than popularity-based models, which is especially noticeable when users' topic interests fluctuate, as Group 2 shows. Using a ranking task formulation, ChatGPT's zero-shot performance in this case is not as good as any baseline model. This suggests that ChatGPT is particularly good at grasping semantics, mostly by suggesting articles that are similar to ones that users have already read. Nonetheless, an interesting finding is that Group 1 and Group 2 consumers' zero-shot performance in the rating job is similar. A plausible explanation is that ChatGPT perceives the rating task similarly to sentiment classification in the zero-shot configuration, where a rating of 1 denotes strongly negative emotion and a rating of 5 denotes very positive sentiment. To verify this theory, we present a prompt that is comparable to the rating task and tells ChatGPT to immediately produce relative sentiment scores for candidate articles. Figure 2 shows the resulting zero-shot performance, which is viewed as a sentiment classification task. Our results provide empirical support for the theory that ChatGPT views the rating job as a type of sentiment classification for candidates in the zero-shot setting. As earlier work has shown, this perspective produces substantial zero-shot performance [33]. This view highlights the significance of appropriate prompt-based task formulation for fine-tuning ChatGPT for downstream activities, and it also contributes to the noteworthy performance shown by Group 2 consumers.

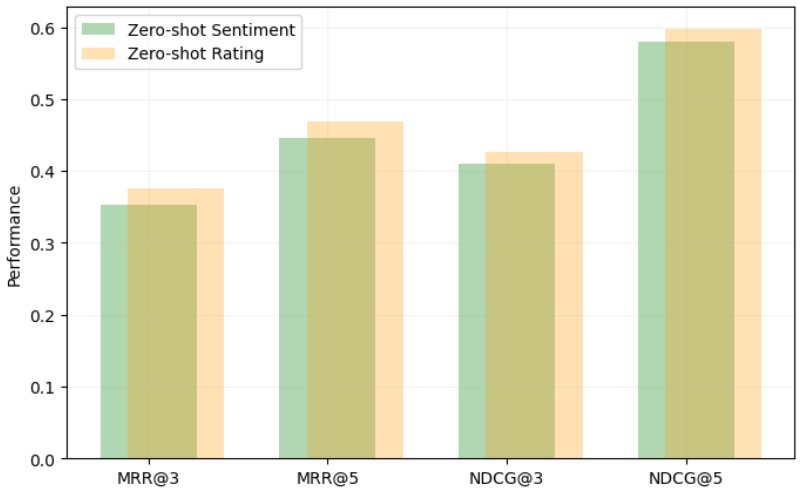

Fig. 2. zero-shot performance

Performance in the fine-tuning setup is significantly better than zero-shot, particularly in Group 1. This improvement can be attributed to the fact that fine-tuning allows for more efficient use of location information in addition to improving ChatGPT's semantic understanding. No matter where they started in the given candidate list, the clicked articles always end up at the top of the created ranking list response once it has been fine-tuned. The model can more effectively utilize positional information thanks to this clever positioning. In contrast, the five scores for the rating question could show up in different places within the generated responses. Even with the same fine-tuning sample sizes, customers in Group 2 see gains from fine-tuning when compared to zero-shot, though not to the same extent as in Group 1. A tenable rationale is that, after fine-tuning, ChatGPT frequently suggests articles from a variety of subjects. Nonetheless, a number of the suggested candidate articles could satisfy this diversity requirement. Without knowing how well-liked these articles are, ChatGPT may choose a random one to satisfy the diversity criteria. On the other hand, when applying for the rating job within the fine-tuning configuration—whether it is applied to Group 1 or Group 2—performances seem comparable when compared to the zero-shot technique. This resemblance can be explained by the fact that the model's capacity to handle numerical comparisons stays largely constant even after fine-tuning, even though semantic understanding can be improved. Furthermore, unlike the ranking task, the rating task does not benefit from the incorporation of positional information from the generated responses during fine-tuning. Moreover, the rating prompt requires you to rate each candidate at the same time, which makes the rating task more difficult. It's important to remember that various consumers might use the rating scale in different ways, requiring the system to pick up on user biases. The finding that ranking performs better than the rating task is consistent with previous studies, particularly when contrasting point-wise and list-wise ranking techniques [27].

6 Conclusion

Finally, our thorough analysis of ChatGPT's performance across a variety of activities and customer segments, compared to baseline and zero-shot methods, has allowed us to draw important conclusions about how well the model performs when fine-tuning news recommendations. By using two separate evaluation groups, Groups 1 and 2, we were able to capture the subtle difficulties that come with making recommendations for news—finding a balance between making recommendations that are tailored to the interests of the user and figuring out how to suggest content that goes beyond their pre-established preferences.

Our results highlight ChatGPT's advantages and disadvantages in various settings. The performance of ChatGPT in the zero-shot configuration showed significant differences when users' topic interests changed, especially in the rating job. This highlights the model's natural ability to understand semantics but also highlights the difficulties it has in adjusting to changing user preferences. After ChatGPT was fine-tuned, significant gains were observed, particularly in Group 1, where the model demonstrated better semantic comprehension and better use of positional information. Nonetheless, the evaluation task conducted in the fine-tuning configuration demonstrated fewer significant enhancements, suggesting that the model's ability to perform numerical comparisons stays mostly unaltered during this procedure.

Moreover, the observed differences in performance between the rating and ranking tasks are consistent with previous studies, emphasizing the difficulties in simultaneously rating a wide pool of applicants and the inherent difficulty in using positional information for the rating prompt. In conclusion, our work sheds insight into the nuances of the dynamics involved in fine-tuning ChatGPT for news recommendation, highlighting its strengths and weaknesses as well as the significance of task formulation and prompt design. These insights highlight the necessity for a thorough assessment of task-specific needs and the changing nature of user preferences, paving the way for well-informed judgments when employing ChatGPT in real-world news recommendation systems.

References

1. 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. 336–345 (2019).
2. 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).
3. Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis. University of Waterloo (2015).
4. 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 (2014).
5. 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 (2022).
6. 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 (2023).
7. 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 (2018).
8. 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 (2022).
9. Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. EMNLP (2023).
10. 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. 1749– 1752 (2020).
11. 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 (2021).
12. Nirmal Jonnalagedda, Susan Gauch, Kevin Labille, and Sultan Alfarhood. 2016. Incorporating popularity in a personalized news recommender system. PeerJ Computer Science 2 (2016).
13. Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for an explainable recommendation. arXiv preprint arXiv:2202.07371 (2022).
14. Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. PBNR: Prompt-based News Recommender System. arXiv preprint arXiv:2304.07862 (2023).
15. Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News. arXiv preprint arXiv:2306.10702 (2023).
16. 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. 3805–3811 (2018).
17. 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 (2023).

18. 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. 1933–1942 (2017).
19. 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 (2021).
20. 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 (2023).
21. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018). [22] 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 (2020), 5485–5551.
22. 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 (2019).
23. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is allyou need. Advances in neural information processing systems 30 (2017).
24. ZengzhiWang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339 (2023).
25. Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. arXiv preprint arXiv:1907.05576 (2019).
26. ChuhanWu, FangzhaoWu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. 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). 6389–6394 (2019).
27. 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. 2382–2386.
28. Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, WinnieWu, et al. 2020. Mind: A large-scale dataset for news recommendation. In Proceedings of the 58th Annual Meeting of the Association for mputational Linguistics. 3597–3606.
29. Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. 2023. OpenP5: Benchmarking oundation Models for Recommendation. arXiv:2306.11134 (2023).
30. JungAe Yang. 2016. Effects of popularity-based news recommendations (“mostviewed”)
31. on users’ exposure to online news. Media Psychology 19, 2, 243–271 (2016).
32. 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 (2023).
33. Zizhuo Zhang and BangWang. 2023. Prompt learning for news recommendation. arXiv preprint arXiv:2304.05263 (2023).