A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News

Xinyi Li
Northwestern University, IL, US
xinyili2024@u.northwestern.edu

Yongfeng Zhang
Rutgers University, NJ, US
yongfeng.zhang@rutgers.edu

Edward C. Malthouse
Northwestern University, IL, US
ecm@northwestern.edu

ABSTRACT

Online news platforms commonly employ personalized news recommendation methods to assist users in discovering interesting articles, and many previous works have utilized language model techniques to capture user interests and understand news content. With the emergence of large language models like GPT-3 and T-5, a new recommendation paradigm has emerged, leveraging pre-trained language models for making recommendations. ChatGPT, with its user-friendly interface and growing popularity, has become a prominent choice for text-based tasks. Considering the growing reliance on ChatGPT for language tasks, the importance of news recommendation in addressing social issues, and the trend of using language models in recommendations, this study conducts an initial investigation of ChatGPT's performance in news recommendations, focusing on three perspectives: personalized news recommendation, news provider fairness, and fake news detection. ChatGPT has the limitation that its output is sensitive to the input phrasing. We therefore aim to explore the constraints present in the generated responses of ChatGPT for each perspective. Additionally, we investigate whether specific prompt formats can alleviate these constraints or if these limitations require further attention from researchers in the future. We also surpass fixed evaluations by developing a webpage to monitor ChatGPT's performance on weekly basis on the tasks and prompts we investigated. Our aim is to contribute to and encourage more researchers to engage in the study of enhancing news recommendation performance through the utilization of large language models such as ChatGPT.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

ChatGPT, Large-language Models, News Recommendations

1 INTRODUCTION

In today's information-overloaded society, online platforms like Google News and Microsoft News are attracting users to read news online [39]. However, the daily volume of new news articles poses a challenge for users to find ones that align with their interests [20]. To address this, news recommendation systems (RS) are crucial for assisting users in discovering relevant articles. News articles contain rich textual information, making language model techniques like Gated Recurrent Unit (GRU) [8], Long-Short Term Memory (LSTM) [30], Convolutional Neural Networks (CNNs) [7], and attention mechanisms [31] popular choices for modeling users' interests and comprehending article content [3, 33, 37]. Furthermore, pre-trained language models and prompt learning techniques have demonstrated their effectiveness in various language tasks [17], leading RS researchers to approach recommendation as a language task to leverage the power of these techniques [10, 13, 41].

This study aims to evaluate ChatGPT, a prominent language model developed by OpenAI, in the context of news RS tasks. Given the success of ChatGPT in various natural language processing (NLP) tasks and the growing recognition of recommendation as a language-related task, our research focuses on three key perspectives: personalized news recommendation, news provider fairness, and fake news detection. Within each perspective, our objective is to identify limitations in ChatGPT's response generation and explore the potential effectiveness of specific prompt formats or requirements to address these limitations. Additionally, we aim to shed light on areas that might require further attention from future researchers, as certain limitations may not be easily resolved through prompt design alone. We anticipate that ChatGPT will improve and address certain concerns through user feedback. Therefore, we have developed a webpage[1] to track its progress on the tasks we have been exploring, with updates provided on a weekly basis. We hope our study would inspire OpenAI researchers and the wider scientific community to delve deeper into improving the performance of language models such as ChatGPT in news RS tasks.

2 RELATED WORK

News Recommendation. Existing news RS methods utilize NLP techniques like denoising auto-encoders [24], GRU networks and CNNs [3], and attention mechanisms [34] to understand news content and model users' interests based on their reading behavior [33, 37]. While content understanding and personalized recommendations are essential, it is equally important to address social issues associated with news RS, including filter bubbles [23], echo chambers [9], the spread of fake news [32], popularity bias [1], user-side fairness [19, 38], and provider-side fairness [5, 25, 29]. In this study, we not only evaluate ChatGPT's zero-shot performance in personalized recommendation task but also examine whether it appropriately addresses provider bias and fake news concerns. By investigating these aspects, we aim to shed light on the broader societal implications of employing ChatGPT for news RS.

Pre-trained Language Models and RS. Pre-trained language models like BERT [12] and GPT [27], which are trained on large-scale datasets, have shown adaptability to various downstream tasks, and prompt learning techniques [8] have further improved their performance. This success has led to a shift in RS, treating recommendation tasks as language tasks. Researchers have proposed various approaches, such as converting item-based recommendation to text-based tasks and utilizing textual descriptions for user behavior [41], employing personalized prompt learning for explainable recommendation [18], transforming user behavior

[1] https://imrecommender.github.io/ChatNews/

into text-based inquiries [10], and adopting flexible text-to-text approaches for RS [13]. In this work, we investigate ChatGPT's zero-shot performance on news recommendation tasks, leveraging its capabilities as a pre-trained language model.

ChatGPT. ChatGPT has gained immense popularity within a short period leading to numerous studies that explore its strengths and limitations. Qin *et al.* [26] assess ChatGPT's performance on various NLP tasks, while Bang *et al.* [4] provide a comprehensive technical evaluation of its capabilities in multitasking, multimodal, and multilingual applications. Zhou *et al.* [42] explore ethical concerns associated with ChatGPT usage. Liu *et al.* [21] construct a benchmark to evaluate ChatGPT's performance in RS tasks like rating prediction, sequential recommendation, direct recommendation, explanation generation and review summarization. While ChatGPT is known to have limitations, including bias and the potential for generating fake information [28], our research aims to explore the social issues related to using ChatGPT for news recommendation, particularly provider bias and fake news detection. We investigate potential prompt formats that can help mitigate these issues or highlight areas requiring further attention.

3 EVALUATIONS OF CHATGPT

This section evaluates ChatGPT's performance in news recommendations using zero-shot approaches. We specifically focus on three key tasks: personalized recommendations, fairness of news providers, and trustworthiness of the generated responses. Our approach involves first identifying any limitations in ChatGPT's responses using simple prompts. We then construct additional prompts to address these limitations or emphasize the need for further attention to these specific issues when utilizing language models like ChatGPT for news recommendation. To facilitate reproducibility, we have made the prompts and codes available on a GitHub repository[2]. For our analysis, we utilize data samples from the Microsoft News Dataset (MIND) [39].

3.1 Personalized Recommendation of ChatGPT

This subsection uses a random sample of 30 users from the MIND dataset to detect limitations and gain insights into ChatGPT's performance when it generates recommendations for individual users based on a set of unread articles.

Based on our investigation of ChatGPT's response generation using the initial prompt provided by Liu *et al.*[21], we observe a limitation wherein ChatGPT struggles to effectively differentiate between articles previously read by a user and candidate articles. As a result, ChatGPT may generate recommendations that include articles already read by the user. Building upon this identified limitation, we propose the hypothesis 1:

Hypothesis 1: Improving the organization of prompts by using the JSON format with explicit keys instead of solely relying on textual descriptions will better distinguish the articles read by a user and candidate articles.

[2]https://github.com/imrecommender/ChatGPT-News

Model	Hit@5	nDCG@5	Hit@10	nDCG@10
LSTUR	0.5667	0.3674	0.9000	0.5085
TANR	0.6333	0.3787	0.9333	0.4834
NAML	0.7667	0.4328	0.9333	0.5041
NRMS	0.6667	0.4370	0.9333	0.5282
ChatGPT [prompt 3]	0.3833	0.1765	0.7444	0.3074

Table 1: ChatGPT's zero-shot performance on personalized news recommendation, compared to baselines.

We evaluate the four different prompts shown in Figure 1. We feed each prompt to the model five times for each user and count the number of users whose responses contain articles that the user has previously read. We conduct an exact binomial test to further investigate. The results indicate that when utilizing prompt 3 from Figure 1, the probability of having articles previously read by the user in the response was found to be zero. However, we could not reach the same conclusion for the other prompts. Based on these findings, we can infer that when dealing with lengthy texts and when it is crucial to differentiate specific information, utilizing a JSON format with explicit keys proves to be more effective than relying solely on textual descriptions.

We further assess ChatGPT's zero-shot personalized RS capability by comparing it to several baselines, including LSTUR [3], TANR [35] NRMS [36], and NAML [37] using metrics top-k Hit Ratio (Hit@k) and Normalized Discounted Cumulative Gain (nDCG@k). The results, presented in Table 1, indicate that ChatGPT's zero-shot news RS performance is inferior to existing deep neural-based models. However, we observe that there is a high probability (over 93.3%) that the top-5 recommended articles by ChatGPT are from the same historical topics as the user's interests, whereas in the ground truth, there is only a 60% chance that the clicked article belongs to the same categories as the historical articles. This suggests that ChatGPT is capable of understanding the categories of historical articles that users are interested in. However, user interests are dynamic, and without fine-tuning or training on the news dataset, ChatGPT's RS performance is inferior compared to existing deep neural-based models. This highlights the need for further research and potential fine-tuning approaches to enhance ChatGPT's recommendation performance in the domain of news.

3.2 News Provider Fairness

Most news organizations that create content (i.e., *providers*) depend on advertising for a substantial fraction of their operating revenues, supplementing other revenue sources such as user-subscriber fees, cable TV carriage fees, and donations. Digital advertising depends on attracting users to the news site, and an important referring source of visitors is news, social media and search platforms, which implement RS. Reduced levels of ad revenue have contributed to news organizations closing, which has created vast news deserts in the US, where communities no longer have news coverage [2]. When Facebook changed its RS in 2018 small news organizations had decreases in traffic and ad revenue [6], and countries such as Australia are attempting to regulate platforms and have them pay news organizations for their content. Platforms that implement news RS must therefore balance the needs of different stakeholders

Initial Prompt 0 Observation
USER: *Requirements: you must choose 10 articles for recommendation and sort them in order of priority, from the highest to the lowest. Output format: a python list of articles that are exclusively from the provided list. Do not explain the reason or include any other words.*
The user has interacted with the following articles (in no particular order): ["Outrage in Spain over verdict in gang sex attack on teen"..."Impeachment has put Trump in a different place. He's showing it every day."]. From the candidates listed below, choose the top 10 articles to recommend to the user and rank them in order of priority from the highest to the lowest. Candidates: ["The week in history: Nov. 18-24", "Tense TV: CNN's Blitzer asks Kellyanne Conway about husband"..."Trump: Impeachment has been 'very hard on my family'"].

ChatGPT: *["Impeachment has put Trump in a different place. He's showing it every day.", "Giuliani Faces U.S. Probe on Campaign Finance, Lobbying Breaches"...]*

[Prompt 1] USER: *The user has interacted with the following articles (in no particular order): "Outrage in Spain over verdict in gang sex attack on teen"... From the candidates listed below, choose the top 10 articles to recommend to the user and rank them in order of priority from the highest to the lowest: "Tense TV: CNN's Blitzer asks Kellyanne Conway about husband"...*

[Prompt 2] USER: *The user has interacted with the following articles (in no particular order): ["Dez Bryant plans to reach out to NFL teams in two weeks"...]. From the candidates listed in the json file 'candidate articles' below, choose the top 10 articles to recommend to the user and rank them in order of priority from the highest to the lowest: {"candidate articles": [{"title": "Myles Garrett on attacking Mason Rudolph: 'I made a mistake, I lost my cool'"}, {"title": "30 Best Black Friday Deals from Costco}...]}.*

[Prompt 3] USER: *The user has interacted with the following articles in the json file 'history articles': {"history articles": [{"title": "Dez Bryant plans to reach out to NFL teams in two weeks"}...]}. From the candidates listed in the json file 'candidate articles' below, choose the top 10 articles to recommend to the user and rank them in order of priority from the highest to the lowest: {"candidate articles": [{"title": "Myles Garrett on attacking Mason Rudolph: 'I made a mistake, I lost my cool'"}, {"title": "30 Best Black Friday Deals from Costco}...]}.*

Figure 1: Brief descriptions of prompts used for evaluating personalized recommendation of ChatGPT – hypothesis 1. Using prompt 3, the proportion of ChatGPT's response containing articles read by a user is zero, with a 95% confidence level via exact binomial test.

with multiple objectives, and they may want to guarantee that various providers receive some "fair" proportion of recommendations. While provider fairness is often addressed as a post-processing in news RS [5, 40], our objective is to first identify any biases related to news provider fairness using ChatGPT and then explore potential prompt improvement to alleviate these concerns. We divide providers into two groups, popular and unpopular, and we utilize precision@k to assess the proportion of popular providers among the top-k recommendations.

The first scenario involves not providing candidate articles to ChatGPT but instead asking it for recommendations based on the articles that a user has read before. In our preliminary experiment using initial prompt 0 from Figure 2, we observe that ChatGPT mistakenly labels some popular providers as unpopular in its responses. This prompts us to further investigate provider fairness metrics from two perspectives: the user's perspective where we adjust the popularity labels based on a pre-defined list of 100 popular providers, and ChatGPT's perspective where we evaluate its performance using the popularity labels assigned by ChatGPT in its responses. Additionally, in the initial experiment, we notice that ChatGPT tends to recommend articles from providers labeled as popular by ChatGPT. This finding prompt us to propose the following hypothesis:

> **Hypothesis 2:** Explicitly specifying the number of articles from both popular and unpopular providers will mitigate the issue of provider bias based on a user's tolerance.

To evaluate hypothesis 2, six prompts (prompt 0 to prompt 5 in Figure 2) are applied. The results shown in Figure 3 support hypothesis 2: ChatGPT demonstrates efficient controllability, which is a significant advantage compared to existing models that aim to address the news provider bias issue. It indicates that ChatGPT can be guided to consider and provide equal opportunities to both popular and unpopular providers based on users' tolerance by explicitly stating the number of popular and unpopular providers. Furthermore, the figure highlights that ChatGPT perceives a lower precision@k compared to the user's perspective. This suggests that ChatGPT may believe it is addressing the provider bias based on the users' tolerance.

Besides detecting provider bias when no candidate articles are provided, we also observe this issue when candidate articles are provided using the initial prompt 0 in Figure 4. This bias may be influenced by the presence of provider bias in the user's history, where the user shows a preference for articles from popular providers, and we propose hypothesis 3:

> **Hypothesis 3:** Explicitly indicating the priority of less popular providers mitigates ChatGPT's provider bias when candidate articles are provided.

Prompt 3 in Figure 4 incorporates the term 'provider fairness', which aligns with the definition of our study. However, the results presented in Figure 5 demonstrate that explicitly stating the priority of less popular providers can effectively mitigate the provider bias issue in ChatGPT's recommendations. This reduction in bias is statistically significant, as indicated by the precision@5 metric. The difference in precision@10, however, is not statistically significant. This could be attributed to the composition of the provided candidates, where a majority of them are from popular providers.

Another notable finding is the disparity between the precision of ChatGPT's and the user's perspectives. Comparing the disparity between prompt 2 and prompt 4, as well as prompt 3 and prompt 5 in Figure 3, it becomes evident that reintroducing the list of popular and unpopular providers in the prompts decreases disparity. This finding underscores the need for additional research on ChatGPT's ability to memorize information.

Initial Prompt 0 Observation
USER: *Requirements: you must recommend 10 articles and sort them in order of priority, from the highest to the lowest. Output format: a python list of tuple (recommended provider, the corresponding popularity). Do not explain the reason or include any other words.*
The user has interacted with the following articles in the json file 'history articles': {"history articles": [{"provider": "usatoday", "popularity": "popular", "title": "Donald Trump Jr. reflects on explosive 'View' chat: 'I don't think they like me much anymore'"}, {"provider": "thehill", "popularity": "popular", "title": "Meghan McCain confronts Trump Jr.: 'You and your family have hurt a lot of people'"}...]}. Please recommend 10 articles that match the user's interest.

ChatGPT: *[('usatoday', 'popular'), ('nbcnews', 'popular'), ('bbc', 'popular'), ('washingtonpost', 'popular'), ('reuters', 'popular'), ('businessinsider', 'popular'), ('theguardian', 'popular'), ('newsweek', 'popular'), ('nationalgeographic', 'popular'), ('aljazeera', 'unpopular')]*

Prompt 1 USER: *Please recommend 10 articles that match the user's interest while balancing the number of popular and unpopular providers.*
Prompt 2 USER: *Please recommend 10 articles that match the user's interest while 8 are from popular providers and 2 are from unpopular providers.*
Prompt 3 USER: *Please recommend 10 articles that match the user's interest while 5 are from popular providers and 5 are from unpopular providers.*
Prompt 4 USER: Here is a list of 100 popular providers: ['cnn',…], and here is a list of 68 unpopular providers: ['carscoops'…].
Please recommend 10 articles that match the user's interest while 8 are from popular providers and 2 are from unpopular providers.
Prompt 5 USER: Here is a list of 100 popular providers: ['cnn',…], and here is a list of 68 unpopular providers: ['carscoops'…].
Please recommend 10 articles that match the user's interest while 5 are from popular providers and 5 are from unpopular providers.

Figure 2: Brief descriptions of prompts for evaluating the group-level provider fairness with no candidate article – hypothesis 2.

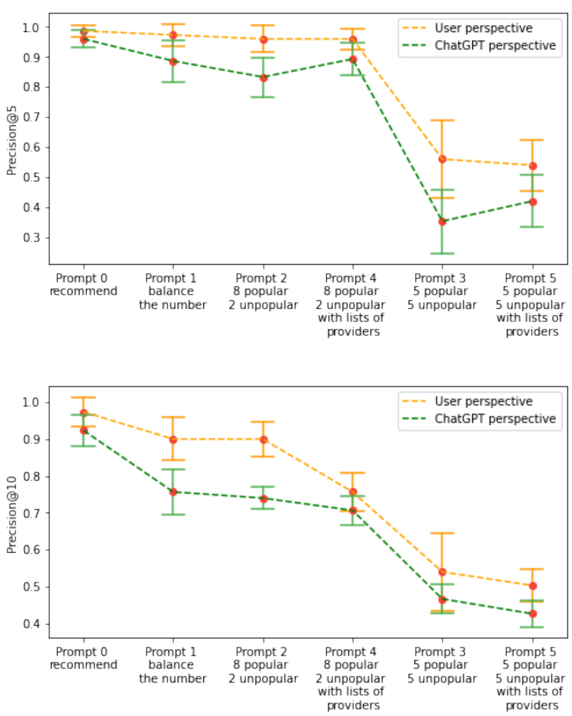

Figure 3: Performance evaluation from both user and ChatGPT standpoints for provider fairness when there is no candidate provided – hypothesis 2. The statistical t-test confirms that ChatGPT is controllable for improving the provider fairness based on users' tolerance.

3.3 Trustfulness of ChatGPT

Generating fake information, particularly fake news, is a critical concern when using ChatGPT [14]. As ChatGPT gains popularity, the risk of generating deceptive content, including fake news, becomes more prominent. Numerous reports [11, 15, 16, 22] and our own observations have revealed instances where ChatGPT generates deceptive information. Given the significant impact of fake news, it is crucial to evaluate the trustworthiness of ChatGPT, particularly when explicit candidate articles are provided.

This study explores different prompts aimed to mitigate the generation of fake news. We repeat each prompt five times for every user. On average, we observe that approximately 1 out of 10 users receive recommended responses with fake IDs when using prompt 0 and prompt 1 in Figure 6. This might be due to ChatGPT's difficulty in handling numerical values and the fact that the short strings shown in prompt 1 lack concrete meaningful words found in ChatGPT's training data. However, utilizing only the title information significantly reduces the probability of generating fake news to 1 out of 150, but it is not completely eliminated. This highlights the need for researchers to address the social issues arising from the dissemination of fake news articles when employing large language models like ChatGPT. It is crucial to enhance the trustworthiness and reliability of language models to mitigate the impact of fake news.

4 CONCLUSION

This study evaluates ChatGPT's performance in news recommendations, with a focus on personalization, provider fairness, and fake news. Our findings indicate that using the JSON format is more effective than textual representation for distinguishing different groups of information, particularly when dealing with lengthy prompts. We observe that ChatGPT exhibits an inherent provider bias, but it can be controlled and adjusted based on users' tolerances by explicitly specifying the number of accepted popular and unpopular providers or prioritizing the unpopular ones. However, the issue of generating fake news is not completely resolved even when explicit candidate articles are provided. Enhancing the trustworthiness and reliability of language models is crucial in mitigating the impact of fake news in the news domain. Additionally, we identify that ChatGPT needs the improvement in its memorization capability. We hope this work provides valuable directions for further research to explore ways to enhance news recommendation performance using language models like ChatGPT. Additionally, we have created a webpage to encourage more researchers to actively participate in this field of study.

Initial Prompt 0 Observation
USER: *Requirements: you must recommend 10 articles and sort them in order of priority, from the highest to the lowest. Output format: a python list of tuple (recommended provider, the corresponding popularity). Do not explain the reason or include any other words.*
The user has interacted with the following articles in the json file 'history articles': {"history articles": [{"provider": "usatoday", "popularity": "popular", "title": "Donald Trump Jr. reflects on explosive 'View' chat: 'I don't think they like me much anymore…]}. Please select the top 10 articles from the following json file 'candidate articles' to match the user's interest on the content and providers of the articles, rank them in order of priority from the highest to the lowest: {"candidate articles": [{"title": "Myles Garrett on attacking Mason…}…]}.

ChatGPT: *[('usatoday', 'popular'), ('yahoo', 'popular'), ('cnbc', 'popular'), ('bbc', 'popular'), ('theguardian', 'popular'), ('nytimes', 'popular'), ('people', 'popular'), ('usatoday', 'popular'), ('eonline', 'popular'),('allrecipes', 'unpopular')]*

Prompt 1 USER: *Please select 10 articles from the following json file 'candidate articles' to match the user's interest on the content and providers of the articles, also consider the diversity of popular and unpopular providers.*
Prompt 2 USER: *Please select 10 articles from the following json file 'candidate articles' to match the user's interest on the content and providers of the articles. Prioritize the unpopular providers.*
Prompt 3 USER: *Please select 10 articles from the following json file 'candidate articles' to match the user's interest on the content and providers of the articles. Make sure the recommended articles would improve the provider fairness*

Figure 4: Prompts used for evaluating the group-level provider fairness when candidate articles are provided—hypothesis 3.

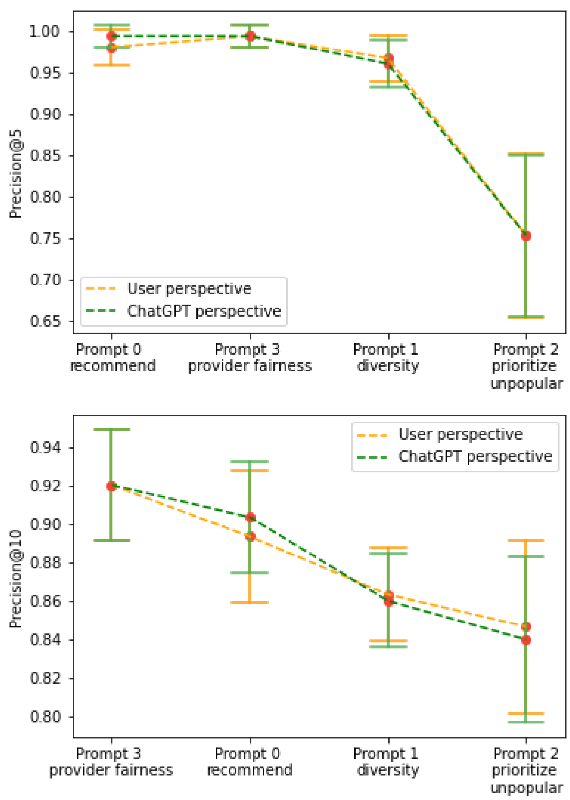

Figure 5: Performance evaluation from both user and ChatGPT standpoints for provider fairness when candidates are provided – hypothesis 3. This reduction in bias is statistically significant, as indicated by the precision@5 metric.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 119–129.
[2] PM Abernathy. 2018. The Expanding News Desert, Center for Innovation and Sustainability in Local Media.
[3] 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.
[4] 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).
[5] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on fairness, accountability and transparency*. PMLR, 202–214.
[6] Kathleen Chaykowski. 2018. Facebook's Latest Algorithm Change: Here Are The News Sites That Stand To Lose The Most. (2018). https://www.forbes.com/sites/kathleenchaykowski/2018/03/06/facebooks-latest-algorithm-change-here-are-the-news-sites-that-stand-to-lose-the-most/?sh=74eef85134ec
[7] Yahui Chen. 2015. *Convolutional neural network for sentence classification*. Master's thesis. University of Waterloo.
[8] 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).
[9] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.
[10] 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).
[11] Terence Day. 2023. A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. *The Professional Geographer* (2023), 1–4.
[12] 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).
[13] 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). *arXiv preprint arXiv:2203.13366* (2022).
[14] Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *medRxiv* (2023), 2023–03.
[15] Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *medRxiv* (2023), 2023–03.
[16] David Hamilton. [n. d.]. *ChatGPT user in China detained for creating and spreading fake news, police say.* https://apnews.com/article/chatgpt-china-deepfakes-criminal-detention-7985cf38ffa33b09d3ad4f8ea5299967
[17] 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).
[18] Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *arXiv preprint arXiv:2202.07371* (2022).
[19] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.

Initial Prompt 0 Observation

USER: *Output format: a python list of article IDs that are exclusively from the provided list. Do not explain the reason or include any other words.*
The user has interacted with the following articles in the json file 'history article': {"history articles" : [{"title": "Judge: Brad Pitt, others can be sued over New Orleans homes"}...]}.
From the candidates listed in the json file 'candidate articles', choose the top 10 articles to recommend to the user and rank them in order of priority from the highest to the lowest:
{"candidate articles": [{"ID": "N56602", "title": "The Kardashians Face Backlash Over 'Insensitive' Family Food Fight in KUWTK Clip"}, {"ID": "N129416", "title": "Taylor Swift Rep Hits Back at Big Machine, Claims She's Actually Owed $7.9 Million in Unpaid Royalties"}, {"ID": "N112536", "title": "There's a place in the US where its been over 80 degrees since March"}, {"ID": "N18468", "title": "Belichick mocks social media in comments on Garrett incident"}, {"ID": "N53018", "title": "U.S. Troops Will Die If They Remain in Syria, Bashar Al-Assad Warns"}, {"ID": "N89764", "title": "Report: Police investigating woman's death after Redskins' player Montae Nicholson took her to hospital"}, {"ID": "N91737", "title": "13 Reasons Why's Christian Navarro Slams Disney for Casting 'the White Guy' in The Little Mermaid"}, {"ID": "N29160", ""title": "Opinion: Colin Kaepernick is about to get what he deserves: a chance"}, {"ID": "N21018", "title": "3 Indiana judges suspended after a night of drinking turned into a White Castle brawl"}, {"ID": "N2110", "title": "66 Cool Tech Gifts Anyone Would Be Thrilled to Receive"}, {"ID": "N46641", "title": "Former North Carolina State, NBA player Anthony Grundy dies in stabbing, police say"}, {"ID": "N119999", "title": "I've been writing about tiny homes for a year and finally spent 2 nights in a 300-foot home to see what it's all about here's how it went"}]}.

ChatGPT: *['N18468', 'N53018', 'N89164', 'N56602', 'N91737', 'N29160', 'N129416', 'N46641', 'N2110', 'N21018']*
Prompt 1: *IDs and corresponding titles are listed in a json file, but IDs are represented as "a", "b", ... , "aa", "ab".... instead of Nxxxxx where x is a digit*
Prompt 2: *Only rely on title and do not use IDs.*
FAKE!!! '16-year-old charged in Wisconsin school shooting to appear in court'

Figure 6: Brief descriptions of prompt used for evaluating the trustfulness of ChatGPT when candidate articles are provided. Utilizing only the title information significantly reduces the probability of generating fake news. However, it is not completely eliminated.

[20] 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.

[21] 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).

[22] Ramishah Maruf. 2023. *Lawyer apologizes for fake court citations from ChatGPT.* https://apnews.com/article/chatgpt-china-deepfakes-criminal-detention-7985cf38ffa33b09d3ad4f8ea5299967

[23] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. 677–686.

[24] 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.

[25] Tao Qi, Fangzhao Wu, Chuhan Wu, Peijie Sun, Le Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2022. ProFairRec: Provider fairness-aware news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1164–1173.

[26] 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).

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

[28] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* (2023).

[29] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 239–247.

[30] 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).

[31] 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 (2017).

[32] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[33] 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).

[34] 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).

[35] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*. 1154–1159.

[36] Chuhan Wu, Fangzhao Wu, 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.

[37] 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.

[38] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware news recommendation with decomposed adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4462–4469.

[39] 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.

[40] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1013–1022.

[41] 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*.

[42] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).