Cognitive Biases in Large Language Models for News Recommendation

Yougang Lyu[1], Xiaoyu Zhang[2], Zhaochun Ren[3] and Maarten de Rijke[1]

[1]*University of Amsterdam, Amsterdam, The Netherlands*
[2]*Shandong University, Qingdao, China*
[3]*Leiden University, Amsterdam, The Netherlands*

Abstract

Despite large language models (LLMs) increasingly becoming important components of news recommender systems, employing LLMs in such systems introduces new risks, such as the influence of cognitive biases in LLMs. Cognitive biases refer to systematic patterns of deviation from norms or rationality in the judgment process, which can result in inaccurate outputs from LLMs, thus threatening the reliability of news recommender systems. Specifically, LLM-based news recommender systems affected by cognitive biases could lead to the propagation of misinformation, reinforcement of stereotypes, and the formation of echo chambers. In this paper, we explore the potential impact of multiple cognitive biases on LLM-based news recommender systems, including anchoring bias, framing bias, status quo bias and group attribution bias. Furthermore, to facilitate future research at improving the reliability of LLM-based news recommender systems, we discuss strategies to mitigate these biases through data augmentation, prompt engineering and learning algorithms aspects.

Keywords
News recommender system, Large language models, Cognitive bias

1. Background

Large language models (LLMs) are becoming crucial components of recommender systems [1, 2, 3, 4]. In particular, news recommender systems rely heavily on LLMs to analyze vast amounts of textual data, ensuring that users receive news articles that align with their interests and preferences [5, 6, 7].

Despite their growing importance and effectiveness, the deployment of LLMs in news recommender systems is not without risks [8, 9, 10, 11, 12]. One significant issue that has emerged is the influence of cognitive biases in LLMs on the ability to make decisions correctly [13, 14, 15, 16]. Cognitive biases are systematic patterns of deviation from norm or rationality in judgment, can lead to the production of inaccurate or skewed outputs [17, 18]. Specifically, Itzhak et al. [19] find that LLMs fine-tuned on human-generated data are significantly affected by cognitive biases during the inference phase, which seriously affects the reliability of LLM-based news recommender systems.

Identifying and addressing the potential impact of cognitive biases in LLMs for recommender systems is critical, especially in the high-stake news recommendation task. News recommender systems play a crucial role in shaping public opinion, informing decision-making, and influencing societal discourse [20]. Therefore, any distortion in the news recommendation process caused by cognitive biases can have far-reaching consequences, potentially spreading misinformation [21, 22], reinforcing stereotypes [23], or contributing to echo chambers [24].

Consequently, there is a pressing need to discuss the potential risks of cognitive bias in LLMs for news recommender systems and to explore possible solutions. This leads to our two central research questions: (1) *How do cognitive biases influence the decision-making processes of LLMs for news recommendations?* (2) *What strategies can be employed to mitigate these cognitive biases?* To this end, we analyze the impact of various cognitive biases on LLM-based news recommender systems. Additionally, we discuss strategies for mitigating these biases through data augmentation, prompt engineering, and learning algorithms.

2. Risks of Cognitive Biases

News recommender systems aim to personalize the news articles presented to users, tailoring the recommendation based on individual preferences, behavior, and historical data. LLM-based news recommender systems often leverage LLMs to filter and rank news articles, determining which articles to prioritize. However, LLMs trained on large amounts of human-generated data may inherit human cognitive biases, reflecting similar patterns in their outputs [16, 19]. Cognitive biases in LLMs can profoundly influence LLM-based news recommender systems, potentially leading to several adverse effects:

- **Anchoring bias [17]:** The reliance on the first piece of information (the "anchor") when making decisions. Assuming a conversational presentation mode of LLM-based news recommender systems, LLMs are likely to be influenced by users' initial interaction with a certain type of news (e.g., a particular political viewpoint), which could disproportionately influence future recommendations.
- **Framing bias [25]:** The way information is presented (framed) can influence decision-making and judgments. For example, news headlines and summaries framed in a particular way can lead LLM-based news recommender systems to severely prefer these news articles.
- **Status quo bias [26]:** The tendency to favor familiar content that has appeared in previous experiences. In LLM-based news recommender systems, LLM may prefer news articles that they have seen in the pre-training or fine-tuning stages, resulting in users finding it hard to view the latest news articles and reducing the diversity of information consumed.
- **Group attribution bias [27]:** This type of bias refers

to the tendency to associate specific topics or opinions with particular demographic groups. E.g., LLM-based news recommender systems may disproportionately recommend certain news topics to specific ethnic or social groups, reinforcing harmful biases and deepening societal divisions.

In conclusion, cognitive biases in LLM-based news recommender systems can have significant effects on society. These biases may trap users in information bubbles, exposing them mostly to content that reinforces their existing beliefs and preferences. Over time, this can lead to increased social polarization, reduced critical thinking, and a less informed public. The risks posed by cognitive biases in LLM-based news recommender systems highlight the need for strategies that mitigate these biases and promote more balanced and diverse news recommendations.

3. Mitigating Strategies

LLM-based news recommender systems are trained on extensive human-generated datasets, which inherently reflect cognitive biases present in human society. LLM-based news recommender systems might inadvertently absorb and replicate these biases, leading to outputs that may not only reflect but also reinforce existing cognitive biases.

To address the cognitive biases of LLM-based news recommender systems, we propose several strategies to mitigate cognitive biases through data augmentation, prompt engineering and learning algorithms aspects. Below is a detailed list of mitigating strategies:

- **Synthetic data augmentation:** LLMs can be employed to generate synthetic datasets that are carefully crafted to break correlations between cognitive biases and irrational outputs [28, 29]. First, we can construct balanced synthetic datasets that counteract the skewed cognitive bias pattern existing in human-generated datasets. Then, we train LLM-based news recommender systems on balanced datasets to reduce the influence of cognitive biases and generate more rational outputs. This strategy helps to reduce specific biases in human-generated datasets, such as group attribution bias, by ensuring that the training data represents a more diverse and equitable distribution of content.
- **Self-debiasing via iterative refinement:** LLMs possess a self-refinement ability that can be harnessed through specially designed debiasing prompts. By iteratively refining the outputs, the model can be guided to recognize and correct biases in its recommendations. For example, in the context of LLM-based news recommender systems, LLMs can be prompted to reconsider their choices and adjust them to minimize the influence of cognitive biases. This process involves LLMs generating outputs, evaluating outputs against debiasing criteria, and revising outputs if necessary [30].
- **Cognitive debiasing through human feedback:** After an LLM generates a news recommendation, human evaluators can assess whether the output exhibits cognitive biases. If biases are detected, techniques such as direct preference optimization (DPO) [31] or reinforcement learning from human feedback (RLHF) [32] can be employed to adjust the model's behavior. This learning from the human feedback process helps to align the model's outputs more closely with human ethical standards and reduces the likelihood of biased recommendations in the

future. By integrating human feedback into the training loop, the model learns to prioritize objectivity and fairness, thus mitigating the effects of cognitive biases.

Overall, LLM-based news recommender systems fine-tuned on extensive human-generated data are susceptible to cognitive biases. However, developing strategies such as synthetic data augmentation, self-debiasing via iterative refinement, and cognitive debiasing through human feedback can help mitigate these biases, fostering more rational and equitable outputs. These strategies are essential for ensuring that LLMs serve as reliable components in decision-making processes for news recommendations.

4. Related Work

Large language models (LLMs) have become essential components of news recommender systems due to their advanced generation abilities. Li et al. [33] directly prompt ChatGPT [34] to generate news recommendations. Prompt4NR [35] introduces the prompt learning paradigm to news recommendation by reframing the task of predicting user clicks on news articles as a cloze-style mask-prediction problem. PGNR [36] transfers the personalized news recommendation task into a text-to-text generation task for LLMs, following a generative training and inference paradigm that directly generates recommendations. ONCE [37] investigates the integration of both open- and closed-source LLMs to improve news recommendation. While the deployment of LLMs into news recommender systems has led to significant improvements, previous works ignore the potential risks associated with cognitive bias in LLMs for news recommender systems.

Recent studies have shown that LLMs are affected by cognitive biases when making decisions [13, 38, 39, 40, 41, 42, 43, 44]. Jones and Steinhardt [13] identify that error patterns of GPT-3 [45] resemble human cognitive biases. Similarly, Agrawal et al. [38] discover the framing effect bias of GPT-3 [45] in the clinical information extraction task. Itzhak et al. [19] suggest that LLMs develop emergent cognitive biases after fine-tuning on extensive human-generated data. Furthermore, Echterhoff et al. [16] introduce a framework for the quantitative evaluation of cognitive biases in LLMs within a student admissions task. To the best of our knowledge, we are the first to discuss the influence of cognitive biases and mitigation strategies in LLM-based news recommender systems.

5. Conclusion

In this study, we focus on exploring possible risks of cognitive biases in LLM-based recommender systems and discussing potential mitigating techniques. Given the critical role news recommender systems play in creating public opinion and influencing social discourse, any distortion brought about by cognitive biases could have far-reaching effects. To explore the risks of cognitive bias, we have introduced the general definitions of anchoring bias, framing bias, status quo bias and group attribution bias, and how they specifically affect the decision-making process of LLM-based news recommender systems. Our discussion of the cognitive biases inherent in LLMs reveals that, without careful deployment, these biases can lead to skewed and incorrect outputs in news recommendations. Our analysis of the cognitive biases in LLM-based news recommender

systems shows that these biases can cause biased and inaccurate news recommendations without careful deployment of LLMs. To address these challenges, we have proposed mitigating strategies from three aspects, including data augmentation, prompt engineering and learning algorithms.

While LLMs demonstrate the potential to enhance the personalization and relevance of news recommendations, their implementation must be carefully managed to avoid exacerbating cognitive biases. By implementing the proposed strategies, we can create more reliable and equitable news recommender systems, better meeting the diverse needs of users and contributing to a healthier news recommendation ecosystem. The teaser video is available at https://github.com/youganglyu/CB_News

Acknowledgments

References

[1] L. Li, Y. Zhang, D. Liu, L. Chen, Large language models for generative recommendation: A survey and visionary discussions, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 10146–10159. URL: https://aclanthology.org/2024.lrec-main.886.

[2] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, E. Chen, A survey on large language models for recommendation, CoRR abs/2305.19860 (2023). URL: https://doi.org/10.48550/arXiv.2305.19860.

[3] A. Vats, V. Jain, R. Raja, A. Chadha, Exploring the impact of large language models on recommender systems: An extensive review, CoRR abs/2402.18590 (2024). URL: https://doi.org/10.48550/arXiv.2402.18590.

[4] X. Zhang, R. Xie, Y. Lyu, X. Xin, P. Ren, M. Liang, B. Zhang, Z. Kang, M. de Rijke, Z. Ren, Towards empathetic conversational recommender systems, arXiv preprint arXiv:2409.10527 (2024).

[5] Z. Zhang, B. Wang, Prompt learning for news recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 227–237. URL: https://doi.org/10.1145/3539618.3591752.

[6] Q. Liu, N. Chen, T. Sakai, X. Wu, ONCE: Boosting content-based recommendation with both open- and closed-source large language models, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024, ACM, 2024, pp. 452–461. URL: https://doi.org/10.1145/3616855.3635845.

[7] X. Li, Y. Zhang, E. C. Malthouse, Prompt-based generative news recommendation (PGNR): accuracy

[8] and controllability, in: Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II, volume 14609 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 66–79. URL: https://doi.org/10.1007/978-3-031-56060-6_5.

[8] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, X. He, Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023, ACM, 2023, pp. 993–999. URL: https://doi.org/10.1145/3604915.3608860.

[9] W. Hua, L. Li, S. Xu, L. Chen, Y. Zhang, Tutorial on large language models for recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023, ACM, 2023, pp. 1281–1283. URL: https://doi.org/10.1145/3604915.3609494.

[10] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, X. He, Large language models for recommendation: Progresses and future directions, in: Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, ACM, 2024, pp. 1268–1271. URL: https://doi.org/10.1145/3589335.3641247.

[11] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, X. He, Item-side fairness of large language model-based recommendation system, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 4717–4726.

[12] Y. Deldjoo, Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency, ACM Transactions on Recommender Systems (2024).

[13] E. Jones, J. Steinhardt, Capturing failures of large language models via human cognitive biases, Advances in Neural Information Processing Systems 35 (2022) 11785–11799.

[14] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do, Nature Machine Intelligence 4 (2022) 258–268.

[15] L. Tjuatja, V. Chen, S. T. Wu, A. Talwalkar, G. Neubig, Do LLMs exhibit human-like response biases? A case study in survey design, CoRR abs/2311.04076 (2023). URL: https://doi.org/10.48550/arXiv.2311.04076.

[16] J. M. Echterhoff, Y. Liu, A. Alessa, J. J. McAuley, Z. He, Cognitive bias in high-stakes decision-making with LLMs, CoRR abs/2403.00811 (2024). URL: https://doi.org/10.48550/arXiv.2403.00811.

[17] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty., Science 185 (1974) 1124–1131.

[18] A. Wilke, R. Mata, Cognitive bias, in: Encyclopedia of human behavior, Academic Press, 2012, pp. 531–535.

[19] I. Itzhak, G. Stanovsky, N. Rosenfeld, Y. Belinkov, Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias, CoRR abs/2308.00225 (2023). URL: https://doi.org/10.48550/arXiv.2308.00225.

[20] N. Helberger, On the democratic role of news recommenders, in: Algorithms, Automation, and News, Routledge, 2021, pp. 14–33.

[21] D. Spina, M. Sanderson, D. Angus, G. Demartini,

D. Mckay, L. L. Saling, R. W. White, Human-AI co-operation to tackle misinformation and polarization, Communications of the ACM 66 (2023) 40–45.

[22] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, J. Cook, Misinformation and its correction: Continued influence and successful debiasing, Psychological science in the public interest 13 (2012) 106–131.

[23] T. Yunkaporta, Right Story, Wrong Story: Adventures in Indigenous Thinking, Text Publishing, 2023.

[24] X. Wang, A. D. Sirianni, S. Tang, Z. Zheng, F. Fu, Public discourse and social network echo chambers driven by socio-cognitive biases, Physical Review X 10 (2020) 041042.

[25] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, Science 211 (1981) 453–458.

[26] W. Samuelson, R. Zeckhauser, Status quo bias in decision making, Journal of Risk and Uncertainty 1 (1988) 7–59.

[27] D. L. Hamilton, R. K. Gifford, Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments, Journal of Experimental Social Psychology 12 (1976) 392–407.

[28] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, S. Joty, Data augmentation using LLMs: Data perspectives, learning paradigms and challenges, arXiv preprint arXiv:2403.02990 (2024).

[29] Y. Lyu, L. Yan, S. Wang, H. Shi, D. Yin, P. Ren, Z. Chen, M. de Rijke, Z. Ren, Knowtuning: Knowledge-aware fine-tuning for large language models, arXiv preprint arXiv:2402.11176 (2024).

[30] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al., Self-refine: Iterative refinement with self-feedback, Advances in Neural Information Processing Systems 36 (2024).

[31] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, Advances in Neural Information Processing Systems 36 (2024).

[32] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, Advances in Neural Information Processing Systems 33 (2020) 3008–3021.

[33] X. Li, Y. Zhang, E. C. Malthouse, A preliminary study of ChatGPT on news recommendation: Personalization, provider fairness, fake news, arXiv preprint arXiv:2306.10702 (2023).

[34] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: Proceedings of NeurIPS, 2022.

[35] Z. Zhang, B. Wang, Prompt learning for news recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 227–237.

[36] X. Li, Y. Zhang, E. C. Malthouse, Prompt-based generative news recommendation (PGNR): Accuracy and controllability, in: European Conference on Information Retrieval, Springer, 2024, pp. 66–79.

[37] Q. Liu, N. Chen, T. Sakai, X.-M. Wu, Once: Boosting content-based recommendation with both open-and closed-source large language models, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 452–461.

[38] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are zero-shot clinical information extractors, 2023.

[39] R. Lin, H. T. Ng, Mind the biases: Quantifying cognitive biases in language model prompting, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5269–5281.

[40] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, arXiv preprint arXiv:2104.08786 (2021).

[41] S. Schmidgall, C. Harris, I. Essien, D. Olshvang, T. Rahman, J. W. Kim, R. Ziaei, J. Eshraghian, P. Abadir, R. Chellappa, Addressing cognitive bias in medical language models, arXiv preprint arXiv:2402.08113 (2024).

[42] L. Tjuatja, V. Chen, S. T. Wu, A. Talwalkar, G. Neubig, Do LLMs exhibit human-like response biases? a case study in survey design, arXiv preprint arXiv:2311.04076 (2023).

[43] Y. Lyu, P. Li, Y. Yang, M. de Rijke, P. Ren, Y. Zhao, D. Yin, Z. Ren, Feature-level debiased natural language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13353–13361.

[44] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of NLP models with checklist, arXiv preprint arXiv:2005.04118 (2020).

[45] T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).