Evaluating the Gender Bias of Large Language Models in Academic Knowledge Production

Judit Hermán*, Budapest University of Technology and Economics, Budapest, Hungary

Kíra Diána Kovács*, Budapest University of Technology and Economics, Budapest, Hungary

Yajie Wang, Center for Collective Learning, CIAS, Corvinus University of Budapest, Budapest, Hungary

Orsolya Vásárhelyi, Center for Collective Learning, CIAS, Corvinus University of Budapest, Budapest, Hungary; Department of Network Science, Institute of Data Analytics and Information Systems, Corvinus University of Budapest, Budapest, Hungary; orsolya.vasarhelyi@uni-corvinus.hu

Keywords: LLM, Gender inequality, algorithmic bias, evaluation

Reference Format:

Hermán, J., Kovács K.D., Wang,Y., and Vásárhelyi, O.. 2025. Example EWAF'25 Submission. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25)*. Proceedings of Machine Learning Research, 5 pages.

1 Introduction

Gender inequality in science is a complex issue that affects every stage of a scientific career, from education to professional advancement [1]. Despite progress in recent decades, women remain underrepresented in most scientific fields, particularly in leadership roles and prestigious research positions [2].

On November 30, 2022, OpenAI released ChatGPT, an AI language model that has rapidly transformed how we communicate, learn, and conduct research [3]. Since then, numerous companies have launched their own models, integrating them into various products. ChatGPT's use has also expanded into knowledge production and scientific publication [4]. It holds promise for addressing long-standing inequalities in academia, such as assisting non-native English speakers in articulating their scientific discoveries more clearly and efficiently [5,6].

However, large language models (LLMs) have been shown to exhibit biases, failing to represent men and women equally in image generation [7,8,9,10]. They also generate factually incorrect responses and fabricate non-existent references [11,12].

In this paper, we examine references generated by the ChatGPT-4o model across 26 research areas within four main scientific domains: Physical Sciences, Health Sciences, Social Sciences, and Life Sciences [13]. Specifically, we designed a prompt that instructed ChatGPT to generate literature reviews in various subfields and provide references including authors' full names, article titles, journals, publication years, and DOIs. We then compared these references across research areas to OpenAlex, an open-source database containing over 250 million scientific publications [13].

Our analysis focuses on the publication years of cited articles, the gender composition of co-author teams, and the occurrence of hallucinated references provided by ChatGPT. Additionally, we assessed the consistency of literature reviews generated on the same topic. While previous studies have identified bias through small-sample analyses within

Authors' Contact Information: Orsolya Vásárhelyi, Corvinus University of Budapest, orsolya.vasarhelyi@uni-corvinus.edu

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

specific subfields, our approach systematically compares hundreds of literature reviews across 26 subtopics, providing a broader evaluation and laying the groundwork for a standardized framework.

2 Data and Methods

Data Our analysis relies on two main data sources. The first is OpenAlex, a free and open dataset of global scientific publications, which contains over 250 million papers dating back to the 1800s. We used its topic classification system to generate literature reviews for the 20 most frequently published topics within 26 subfields, spanning four main scientific research fields [13].

To account for the skewed nature of citation practices—where 95% of papers have fewer than two citations, 93% have fewer than one citation, and 97% have never been cited—we included only papers cited at least twice. We then instructed ChatGPT to generate literature reviews for 497 topics (23 subfields with 20 topics each, plus three subfields with fewer than 20 topics). Each prompt was repeated 10 times to assess the consistency of the results. Aggregated outputs were computed from these repeated attempts, and derived metrics were based on the average for each prompt. Literature reviews were generated using the GPT-4o model via an API, and all analyses were performed in Python.

Prompting Our prompt was created to be as specific as possible to get the desired result, which is a 500-word long literature review with at least 20 references listed after the review in a specific format to process the resulting reviews easier. As related work suggests [14], giving ChatGPT a role can result in more accurate responses. We also specify the field and subfield names since there are topics which can belong to multiple main fields (e.g.: Contraception is both a topic in Medical Sciences and Sociology).[1]

Gender Inferring To identify the authors' gender in references, we used a Python package, called gender guesser to determine authors' genders on the basis of first names. It classifies first names into the following categories based on a large dictionary of first names and gender: male, female, andy, mostly_female, mostly_male or unknown. We are aware of the limitations that name-based gender inferring methods can cause, especially yielding lower accuracy for Asian names, and handling gender binary [15].

3 Results

Similarly, to previous studies [7], we also found that ChatGPT generally cites more recent papers, with over 80% of its references published after 2000, compared to 60-65% in OpenAlex. The difference is significant in the Physical, Health, and Life Sciences, but not in the Social Sciences.

Previous research, based on smaller samples within specific fields, found that ChatGPT-3 cites fewer women than men [16]. Thus, we initially assumed that ChatGPT-4o would cite papers with lower representation of women. However, we found that the average ratio of women authors is higher than what the OpenAlex database suggests. *Figure 1.* shows

[1] **Final Prompt:** *"You are my very smart research assistant who knows well the scientific literature. The topic of our research is 'display name'. It is a subtopic of 'subfield name' which belong to the main topic of 'field name'. Please write a literature review which contains 500 words (and not shorter!) WITHOUT!!! references. References are not counted in the 500 words. Please cite at least 20 scientific papers (and not less!!!) with numbers in the text and add them as references at the end of the text. Do not create fake citations! List the correct scientific references at the end of your literature review and their links in recent literature in the following form: citation number (as in the text)| authors | reference titles | journal names| publication years| digital object identifiers | reference links. Separate these 7 things with |. Don't write other information, just these 7 things are needed. First write the review, then in new line write the word 'References:' and then in separate lines the 20 references in the form given above. Please write the name of all the authors (in the form: first name (middle name) last name), avoid writing the 'et al.' expression. Write down the whole name of the authors. And it's very important to separate the 7 elements of the references with |, please don't forget it. Do not write a title and do not write '#'-s before the word 'References:'! It is important to write at least 20 correct scientific references!"*

the distribution of the ratio of women authors in ChatGPT and OpenAlex, indicating significant differences between the two in all fields except the Social Sciences. Moreover, the degree of this difference is not constant; it varies depending on the field. The over-representation of papers with a higher ratio of women authors is especially notable in the Health and Life Sciences.

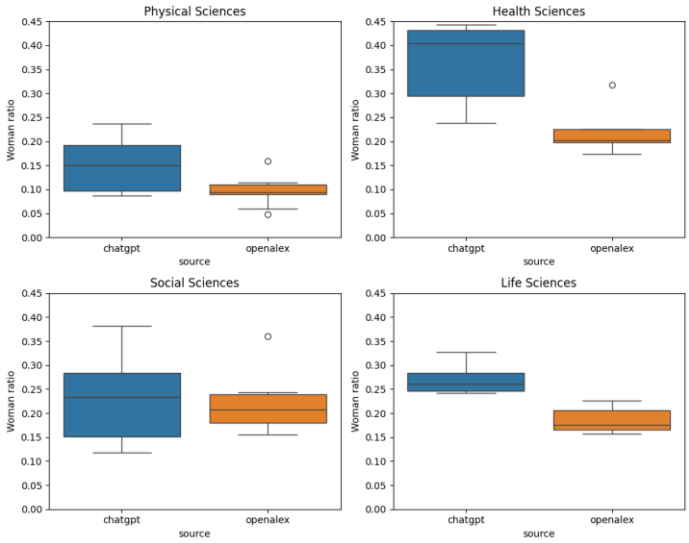

Fig. 1. Distribution of the ratio women in references aggregated by major academic domains by ChatGPT (blue) and OpenAlex (orange) by the 4 main scientific fields. The boxplots span from the first to the third quartile with a line at the median; whiskers extend to the furthest points within 1.5× the interquartile range. References generated by ChatGPT contains hallucinations as well.

To test whether the observed overall difference is due to the increasing representation of women in science and ChatGPT's tendency to cite newer papers, we visualized the ratio of women in references by ChatGPT and OpenAlex over the past 65 years (see Figure 2). The difference between ChatGPT and OpenAlex in referencing papers with a higher ratio of women authors is especially visible after 2010. This gap remains significant in the Physical and Life Sciences, suggesting that the over-citation of recent papers does not fully explain the difference.

To detect hallucinated citations, we compared ChatGPT-generated reference titles to OpenAlex titles using pairwise cosine similarity, applying stopword removal and manually defined subfield-specific thresholds to classify matches. These thresholds were set based on manual inspection, including checking ambiguous cases (e.g., book vs. chapter, missing subtitles) and validating potential matches via Google Scholar. Our method yielded very high hallucination rates: 87% for Physical and Health Sciences, 82% for Social Sciences, and 91% for Life Sciences. These numbers are significantly higher than those reported in a previous study in Psychology (36.0%) by Lehr et al. [18]. We assume this difference is due to (1) the restriction of our baseline OpenAlex database to only include papers cited at least twice, and (2) our automated method for detecting hallucinations, which was optimized to minimize false positives. When analyzing the non-hallucinated papers—those considered "real"—the higher-than-expected ratio of women authors

remains. After filtering out hallucinations, the results showed that ChatGPT cites more women in the fields of Physical, Health, and Life Sciences. In these fields, ChatGPT appears to replace male authors with female authors.

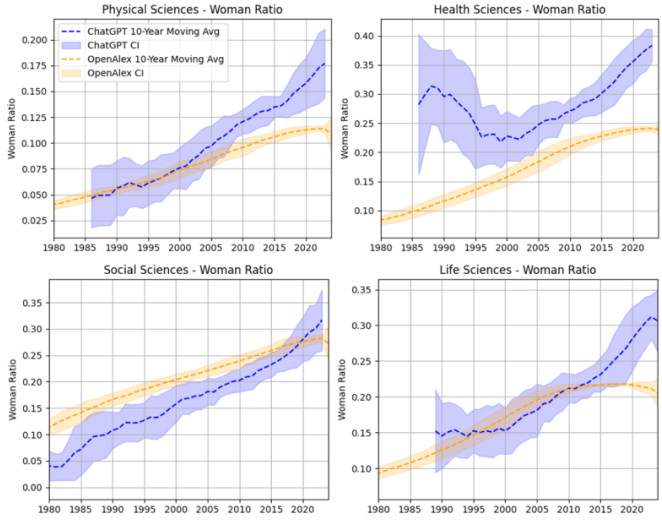

Fig.2. Average ratio of women in references by ChatGPT (blue) and OpenAlex (orange) in the 4 main fields with moving averages and 95% confidence intervals. References generated by ChatGPT contains hallucinations as well.

4 Conclusions and next steps

Our study showed that ChatGPT-4o tends to cite more recent publications and papers with a higher ratio of women authors. It also highlights a high hallucination rate in generated citations, with ChatGPT -4o showing no significant improvement over previous versions, emphasizing the need for a critical evaluation of AI-generated references. We recognize the limitations of this study, such as the use of binary gender categories, relatively small sample sizes, and reliance on a single version of ChatGPT. However, we believe this case study demonstrates the potential of our ongoing efforts toward standardized evaluations for LLMs in scientific knowledge production.

The 2024 AI Index Report published by Stanford University [17] stated that one of the key challenges in responsible AI development is the lack of standardization in AI reporting, benchmarks, and evaluation frameworks. Our ongoing research aims to develop a standardized framework for evaluating LLMs in academic knowledge production by systematically comparing AI-generated literature reviews with real databases of published work.

This work is ongoing, and we are currently applying our framework to DeepSeek V3 and R1, ChatGPT-o1, Llama 3.2, 3.3, Gemini 1.5 Pro and Flash, and Claude 3.5 Sonnet and Haiku, with plans to add more models in the near future. We are also refining our methods to better detect hallucinations and extending our analyzed variables beyond gender to include other demographic factors (ethnicity, country of origin, affiliation) and publication-related factors (publication type, impact factor, online visibility). Our goal is to create a tool that provides a comprehensive overview of how well

LLM-generated literature reviews represent scholarly publications. Further research directions include expert evaluations to benchmark the quality of AI-generated results.

Acknowledgements

This project was funded by the European Union under the European Research Executive Agency project LearnData, 101086712E. The authors gratefully acknowledge the financial support of the Women and Science Chair, a Paris Dauphine-PSL University and its Foundation Chair, in partnership with Amundi, Fondation L'Oréal, La Poste, Generali France, Safran and Talan.

5 References

[1] Charlesworth, T.E.S. B., & Banaji, M.R. (2019). Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions. Journal of Neuroscience.

[2] Sugimoto, C. R., & Larivière, V. (2023). Equity for women in science: Dismantling systemic barriers to advancement. Harvard University Press.

[3] ChatGPT. (n.d.). Wikipedia. Retrieved October 29, 2024, from https://en.wikipedia.org/wiki/ChatGPT

[4] Xames, M. D., & Shefa, J. (2023). ChatGPT for Research and Publication: Opportunities and Challenges. Journal of Applied Learning and Teaching.

[5] Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. Journal of the Association for Information Science and Technology.

[6] Bao, H., Sun, M., & Teplitskiy, M. (2024). Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. arXiv.

[7] Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language Model) to Improve Efficiency in Literature Review for Undergraduate Research.

[8] Grabb, D. (2023). The impact of prompt engineering in large language model performance: a psychiatric example. Journal of Medical Artificial Intelligence.

[9] Suppadungsuk, S., Thongprayoon, C., & et al. (2023). Examining the Validity of ChatGPT in Identifying Relevant Nephrology Literature: Findings and Implications. J. Clin. Med.

[10] Metze, K., Morandin-Reis, R. C., Lorand-Metze, I., & Florindo, J. B. (2024). Bibliographic research with ChatGPT may be misleading: the problem of hallucination. Journal of Pediatric Surgery.

[11] Chelli, & Mikaël. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*.

[12] Siontis, K. C., Attia, Z. I., Asirvatham, S. J., & Friedman, P. A. (2024). ChatGPT hallucinating: can it get any more humanlike? *European Heart Journal*.

[13] OpenAlex Topics. (n.d.). https://help.openalex.org/how-it-works/topics.

[14] Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv.

[15] Vásárhelyi, O., & Brooke, S. (2022). Computing gender. Available at: https://osf.io/preprints/socarxiv/admcs/

[16] Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. Proceedings of the ACM collective intelligence conference.

[17] Perrault, Ray, and Clark, Jack (2024). Artificial intelligence index report 2024. Available at: https://hai.stanford.edu/ai-index/2024-ai-index-report

[18] Lehr, Steven A., et al. 35 (2024). "ChatGPT as research scientist: probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor." *Proceedings of the National Academy of Sciences* 121.: e2404328121