https://doi.org/10.51408/issi2025_182

# Evaluating Large Language Models for Gender Bias in Academic Knowledge Production

Judit Hermán[1], Kíra Diána Kovács[2], Yajie Wang[3], Orsolya Vásárhelyi[4]

[1]*hermanjudit01@gmail.com*, [2]*kira20020111@gmail.com*
Budapest University of Technology and Economics, Faculty of Natural Sciences, Budapest, Hungary

[3]*yajie.wang998@gmail.com*
Center for Collective Learning, Corvinus Institute for Advanced Studies, Corvinus University, Budapest, Hungary

[4]*orsolya.vasarhelyi@uni-corvinus.hu*
Center for Collective Learning, Corvinus Institute for Advanced Studies, Corvinus University; Institute of Data Analytics and Information Systems, Corvinus University, Budapest, Hungary

## Introduction

Gender inequality persists in science, with women being underrepresented in leadership and disadvantaged in hiring, funding, and publishing. While Large Language Models (LLMs) like ChatGPT and Gemini offer new tools for research support, they also risk reinforcing existing biases. Prior studies show LLMs can reproduce gender and racial stereotypes, hallucinate references, and generate inconsistent outputs. This study evaluates references produced by nine advanced LLMs across 26 research subfields and four major domains, comparing them to the OpenAlex database to assess accuracy, gender balance, publication trends, and consistency.

## Related Work

Women remain underrepresented in senior academic roles, especially in STEM, due to barriers like unequal access to resources, limited mentorship, and work-life conflicts (Legewie & DiPrete, 2014; Winslow, 2010; Vásárhelyi, 2020; Hopkins et al., 2013; Huang et al., 2020). LLMs may seem promising for reducing inequalities by equally representing the work of men and women, but they may actually worsen these disparities by reproducing gender and racial biases present in their training data (Ferrara, 2023; Smith & Rustagi, 2021; Zhou et al., 2024; Ghosh & Caliskan, 2023). They also hallucinate references (Metze et al., 2024; Buchanan et al., 2023) and overcite highly cited, male-authored works (Algaba et al., 2024; Antu et al., 2023), potentially reinforcing existing inequalities. Ensuring equity requires critically evaluating AI outputs (Zimmermann et al., 2024; Kotek et al., 2023; Pfohl et al., 2024). Based on these findings, we hypothesized that LLMs undercut women's work.

## Data

We analyzed outputs from nine LLMs and used the OpenAlex database of 250+ million publications. From OpenAlex's classification, we selected the 20 most-published topics in 26 subfields across four disciplines, yielding 497 topics. To reduce the size of our data, we included only articles that were cited at least twice within our OpenAlex baseline database for these topics.

## Methods

We prompted each LLM with a standardized query to generate literature reviews and references. Hallucinated references were detected using fuzzy string matching based on Levenshtein distance and a Jaccard index filter, with a threshold of 0.86. We inferred authors' gender using a name-based gender and ethnicity inference method, Ethnea (Torvik & Agarwal, 2016). For each paper in both the OpenAlex dataset and the LLM-

generated outputs, we calculated the ratio of female authors. We then analyzed these ratios by averaging the proportion of women at both the subfield and major academic domain levels. Statistical differences were tested using the Mann-Whitney U test ($\alpha = 0.05$) on female authorship and reference matching rates.

## Results

Some LLMs (e.g., Claude 3.5 Sonnet, ChatGPT 4o) slightly overcite women, while others (e.g., Gemini models, Llama 3.3 70b, DeepSeek R1) tend to undercite them—often significantly. Citation patterns varied by field: Gemini 2.0 Pro and Llama 3.3 70b cited more women in Health Sciences, while other Gemini models less in Social Sciences. Gender bias persisted even when considering only recent publications, especially in Physical and Life Sciences, indicating model-driven citation patterns.

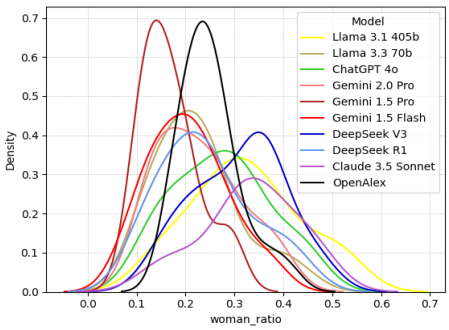


**Figure 1. Density of the ratio of women in the OpenAlex and LLMs' references.**

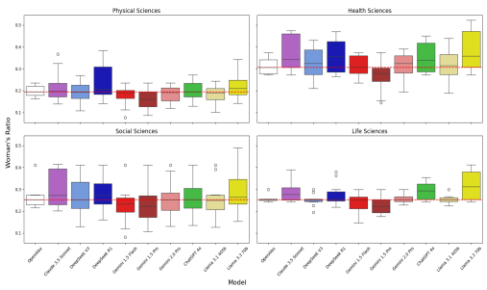


**Figure 2. Distribution of female author ratios in LLM and OpenAlex references across four scientific fields. Boxes show quartiles; whiskers indicate non-outlier ranges. Dotted line marks OpenAlex median.**

Contrary to earlier findings (Antu et al., 2023), our analysis shows that all examined LLMs now favor recent publications, except Gemini models in Social Sciences. The analysis of moving averages across the years reveals that large language models often show statistically significant differences from OpenAlex in the ratio of women authors across disciplines—especially in Physical Sciences—and these disparities become more pronounced in papers published after 2000, indicating increasingly widespread gender citation gaps.

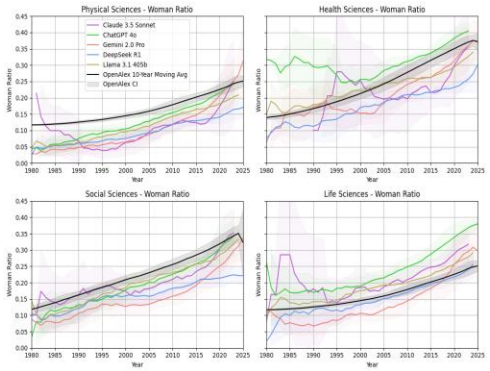


**Figure 3. Ratio of women in references in the 4 main fields with moving averages and 95% confidence intervals.**

Over 70% of LLM references were hallucinated, with ChatGPT 4o reaching 93% and Gemini 2.0 Pro and DeepSeek R1 the lowest (~70%), underscoring the need for citation caution. Even among real references, models like Gemini 1.5 Flash and Llama 3.1 405b undercited women, while Llama 3.3 70b overcited them—especially in Health and Life Sciences—indicating persistent gender bias.

## References


Algaba, A., et al. (2024). Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias.

Antu, S. A., et al. (2023). How ChatGPT Selects Sources: A Study of Bias in Scientific Literature Recommendations.

Buchanan, B., et al. (2023). ChatGPT Hallucinates Non-existent Citations: Evidence from Economics.

Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models.

Ghosh, S., & Caliskan, A. (2023). Stable Diffusion reflects social bias and stereotypes in image generation.
Hopkins, N., et al. (2013). Gender Disparities in Academic Productivity and Representation.
Huang, J., et al. (2020). Historical comparison of gender gaps in scientific publishing and impact.
Kotek, H., et al. (2023). Gender Stereotypes in LLM-generated Texts: An Empirical Investigation.
Legewie, J., & DiPrete, T. A. (2014). The High School Path to STEM: Gendered Influences on Science Orientation.
Metze, K., et al. (2024). Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination.
Pfohl, S. R., et al. (2024). Language Bias in AI-generated Reference Letters.
Smith, S., & Rustagi, J. (2021). Gender and Racial Bias in AI Systems: An Audit of 133 Models. Berkeley Haas Center for Equity, Gender, and Leadership.
Torvik, V. I., & Agarwal, S. (2016). Ethnea—an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. Paper presented at International Symposium on Science of Science, Washington DC, United States.
Vásárhelyi, O. (2020). Barriers to Women's Retention in Technology and Engineering.
Vásárhelyi, O., et al. (2021). Gendered Patterns in Online Science Dissemination.
Winslow, S. (2010). Gender Inequality and Time Allocation in Academia.
Zhou, Y., et al. (2024). Stereotypical Visual Representations in AI-generated Images.
Zimmermann, R., et al. (2024). Evaluating ChatGPT's Ability to Generate Literature Reviews.