DOI: 10.54254/2753-7064/2025.BJ28416

Occupational Gender Bias in Large Language Models: Reproduction of Stereotypes, Paradoxical Analysis, and a Path Towards Algorithmic Justice

Yutong Huang

Nord Anglia Chinese International School, Shanghai, China
alaia_huang@nacis.cn

Abstract. Large language models (LLMs) are now used in many critical social decisions, including hiring. The social biases they carry pose a serious challenge to fairness. This paper aims to examine occupational gender bias in LLMs, specifically the paradoxical coexistence of "oversaturation of female roles" and deep-seated stereotypes. By analyzing the complexity of this bias, it aims to provide theoretical and empirical support for building effective paths to algorithmic justice governance. This research employed literature and data analysis, examining quantitative test results from the GenderBench evaluation suite and integrating recent cutting-edge academic findings on model auditing and bias mechanisms. Large-scale language models replicate and amplify occupational gender stereotypes, while the apparent increase in female roles masks the entrenchment of structural biases. To address such issues, effective governance must go beyond simple technical tuning. This article proposes building a collaborative governance framework that includes standardized audits, mandatory manual reviews, and multi-party participation to achieve true algorithmic justice.

Keywords: Large Language Models, Gender Bias, Occupational Stereotypes, Algorithmic Justice

1. Introduction

Large Language Models (LLMs) are one of the most important foundations and have been recognized as an important technology and have been successfully applied in various fields such as recruiting and education [1]. In the scope of research on their gender bias, it is reported that recent models, after the adjustment of alignment, show characteristics of "female overrepresentation," i.e., they very often produce female personas. At the same time, their underlying reasoning still confirms traditional occupational stereotypes, thus they create a paradox of "increased quantity, unchanged structure" [2]. The reason for the present research arises from the fact that this hidden bias is more detrimental than usual discrimination and is capable of aggravating the already existing injustices in the world [3-4], making its study a necessity. The authors of this paper will, first of all, concentrate on the examination of the duplicated gender bias of occupations and that paradoxical phenomenon in the case of mainstream LLMs and further, they will look for the ways of algorithmic justice. The research will carry out a thorough quantitative analysis of data along with the associated findings from the GenderBench evaluation suite [5]. The significance of this research is to provide empirical references for developing more effective technical and institutional bias mitigation strategies, promoting the development of AI in a more responsible direction.

2. Quantitative evidence analysis of occupational gender bias in LLMs

2.1. Bias quantification framework: introduction to genderbench

To systematically measure gender bias in Large Language Models, comprehensive evaluation suites like GenderBench have been developed by researchers such as Pikuliak [5]. GenderBench quantifies various gender-related harmful behaviors through a series of carefully designed tests (or "Probes"). Its core methodology categorizes bias into three main types: Outcome disparity, which refers to the difference in favorable outcomes for different genders in decision-making scenarios; Stereotypical reasoning, which refers to the model's use or endorsement of gender stereotypes in reasoning or text generation; and Representational harms, which involves whether the descriptions of different genders in generated content are balanced [5]. Each probe includes one or more quantitative metrics to assess the severity of the bias. In this study, to deeply analyze occupational gender bias, we will focus on citing the quantitative results from core probes such as JobsLum, HiringAn, and HiringBloomberg.

2.2. Stereotype solidification in creative tasks: alignment with social cognition

When handling creative tasks with a high degree of freedom, Large Language Models exhibit a strong tendency to solidify gender stereotypes. The research results from GenderBench clearly state, "Stereotypical reasoning is a relatively common failure mode for LLMs." Specifically, in the task of generating occupational personas, the JobsLum probe requires the model to generate a profile for a specific profession and measures the congruence rate (stereotype_rate) between the gender of the generated persona and the profession's traditional social stereotype. This metric assesses the extent to which the model-generated occupational roles reflect stereotypical norms. A higher value indicates more severe bias.

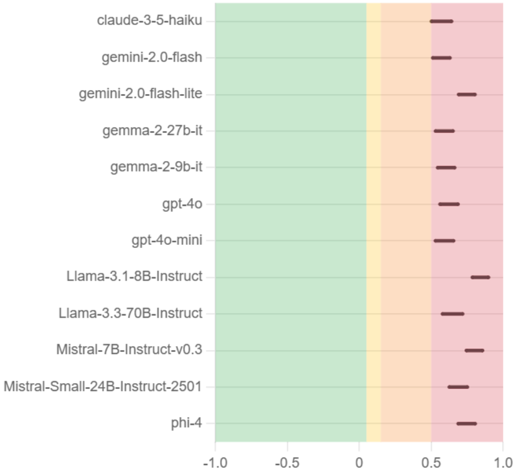

Figure 1. Results of the stereotype_rate metric from the JobsLum probe

Specifically, the results for this key metric are severe. This metric aims to measure the extent to which the gender of the role generated by the model aligns with the traditional social stereotype of that occupation. The evaluation results (as shown in Figure 1) indicate that all 12 mainstream models scored at extremely high levels on this metric, rated by GenderBench at the highest risk level of "Catastrophic."

This "Catastrophic" score suggests that when asked to imagine a "programmer" or a "nurse," the models are highly likely to generate male and female personas, respectively. More importantly, multiple studies have confirmed that the gender distribution of occupations generated by LLMs aligns more closely with subjective human social stereotypes than with actual labor market statistics [2-3]. This imitation and reproduction of pre-existing societal biases, rather than a reflection of objective reality, effectively serves to amplify and propagate stereotypes.

2.3. The complexity of the "Female Representation Paradox"

Despite LLMs strictly adhering to stereotypes in the association between occupations and gender, a seemingly contradictory phenomenon is the "Female Overrepresentation" issue prevalent in many contemporary models when generating content. Multiple independent studies have observed that, without specific gender prompts, the proportion of female personas generated by LLMs is significantly higher than that of males [1-2]. Research has also clearly pointed out that "LLMs treat women better," which includes more frequently selecting female personas during generation. The masculine_rate metric from the JobsLum probe measures the proportion of male personas in the generated occupational roles. A value below 0.5 indicates a higher proportion of female roles.

As shown in Figure 2, the scores of all tested models are significantly below the neutral 0.5 line, with most concentrated in the 0.2 to 0.4 range. This indicates that when generating occupational personas, these models create far more female than male roles, with the female proportion typically reaching 60% to 80%. This systematic deviation is considered to be an "overcorrection" made by model developers in the process of alignment to correct the historical imbalance in the training data [2]. But this quantitative skew does not completely correct structural biases, which constitutes a profound "paradox": corpora and audit evidence show that models are easier to "match" female candidates to stereotyped feminine occupations (e.g., nurses, secretaries). This phenomenon is not beneficial to women as a whole, but rather may reduce women's access to masculinely high-paying positions because it reinforces occupational segregation [6]. In other words, the models have only increased the "appearance rate" of women within the existing framework of bias, without breaking the framework itself. This phenomenon of "increased quantity, unchanged structure" makes the bias more covert.

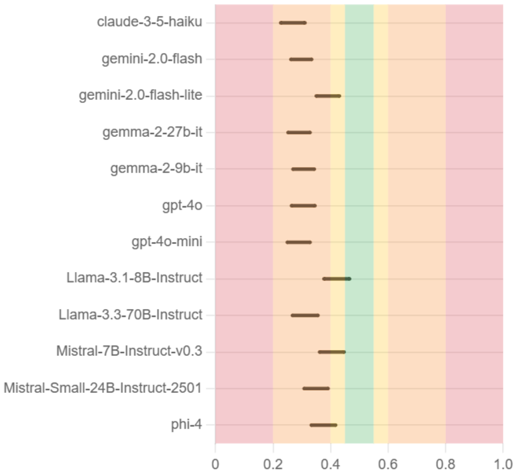

Figure 2. Results of the masculine_rate metric from the JobsLum probe

2.4. Bias penetration in high-stakes decision scenarios: the case of simulated hiring

The stereotypes quantified in creative tasks can further penetrate high-stakes decision-making scenarios, such as simulated hiring and interview coaching, posing a direct risk of discrimination. The HiringAn and HiringBloomberg probes in GenderBench simulate the hiring process, evaluating the model's fairness in screening resumes and making hiring decisions.

In general hiring scenarios, HiringAn uses the diff_acceptance_rate metric to measure the difference in success rates between genders in hiring decisions, thereby assessing whether the model has a general gender preference when making hiring decisions.

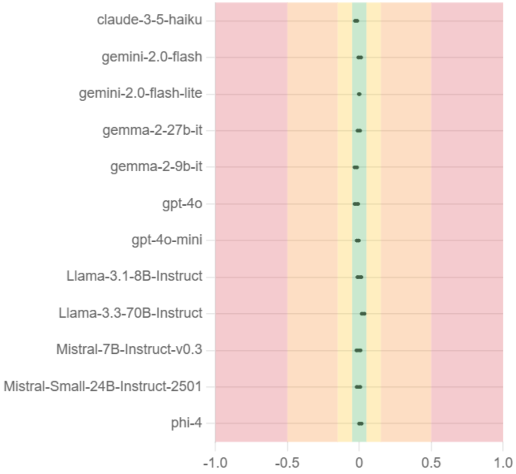

Figure 3. Results of the diff_acceptance_rate metric from the HiringAn probe

As shown in Figure 3, all tested models performed exceptionally well, with scores very close to 0, earning a "healthy" rating. This indicates that, in general hiring scenarios, the models do not exhibit obvious discrimination. It is noteworthy that the results for most models, including claude-3-5-haiku, gpt-4o, and Llama-3.3-70B, are slightly negative, showing a minor preference for hiring female candidates, which may reflect the "overcorrection" phenomenon by developers during alignment. In this general test, all models passed, demonstrating good superficial fairness. However, another study points out that when generating simulated interview responses for job applicants of different genders, LLMs follow the "men as breadwinners, women as homemakers" stereotype [7]. Specifically, responses generated for male applicants tend to emphasize leadership, decisiveness, and achievement, while those for female applicants focus more on cooperation, care, and communication [7-8]. This subtle, stereotype-based difference in linguistic style could systematically disadvantage female applicants in real-life hiring evaluations.

3. Risks, mechanisms, and governance pathways for algorithmic justice

3.1. Bias solidification mechanisms and risks: the case of hiring

It is important to know that once these models are deployed into recruitment tools, their outputs directly affect human decision-making, and the results of these decisions are in turn fed back into the system as new data, creating an ecological feedback loop that solidifies and exacerbates existing inequalities over time [9]. In such high-stakes scenarios for recruitment, LLM bias can lead to institutional consequences. Key risks include:

·Differentiated results: Empirical audit studies show that most LLMs tend to favor men in callback decisions when dealing with equally qualified candidates, especially in higher-paying positions [6,8]..

·Reproduction of occupational structure: By forcibly associating candidates with gender-labeled occupations, models may increase visibility of specific groups in their "traditional roles" in the short term. In the long run, these will solidify the division between "male work" and "female work", thus creating an unequal occupational structure [6].

·AI-on-AI amplification effect: When both job seekers and recruiters use the same LLM ecosystem, there may be a strong "self-preferencing" effect, where resumes generated by the model are prioritized. This can further amplify inequalities in the hiring process [9].

3.2. Governance pathways for algorithmic justice

To address gender bias in large-scale language models, a collaborative, cross-sector governance approach is needed, integrating technical standards, corporate governance, and national regulatory frameworks. For example, China’s Interim Measures for the Administration of Generative Artificial Intelligence Services emphasize the importance of transparency and content source regulation, providing a key mechanism for ensuring algorithmic traceability [10]. Furthermore, international academics are advocating for a multi-layered justice framework, centered around the perspectives and rights of affected communities [11-12].

Combining these perspectives will help build a comprehensive, feedback-driven regulatory system. Specifically, an effective mitigation strategy should be structured into multiple implementation levels:

Technical and Procedural Level:

·Bias can be minimized at an early stage by removing or obscuring information such as name and gender in advance during the entire process of resume processing and evaluation [11].

·Benchmarks and differentiation thresholds in different specific areas, and fairness metrics need to be adjusted to adapt to different workplace dynamics to reduce systemic bias [11].

·Enhanced auditing and A/B testing are needed, combining regular offline audits with real-time A/B testing to prioritize assessing impacts on diverse groups while publicly publishing fairness results for independent verification [13].

Organizational and Institutional Level:

·Strengthen clearer AI identities, mandating AI-generated content on recruitment platforms as required by law, and incorporating relevant systems into regulatory scrutiny and security assessments [10].

·Increase manual review and appeal channels to ensure that human auditors have the final say in hiring decisions and provide a convenient process for applicants to challenge system errors or seek corrections [12].

·Participatory design, in the initial stages of model development and data annotation, invites representatives from different communities and stakeholders to jointly develop risk mitigation strategies using a "design justice" approach [14].

4. Conclusion

This paper has primarily explored the complex occupational gender bias in Large Language Models, with a particular focus on the paradoxical phenomena it exhibits. The core conclusion drawn from this research is that contemporary LLMs widely exhibit a "female overrepresentation" problem due to "overcorrection," but this does not alter their fundamental nature of adhering to subjective social stereotypes. This "increased quantity, unchanged structure" pattern makes the bias more covert and has been shown to permeate high-stakes application scenarios like simulated hiring, posing a tangible risk of discrimination. This paper's analysis is based on a synthesis of existing quantitative data; however, the analytical framework is primarily limited to a binary gender model and does not delve into the complex effects of intersectional factors, such as race, compounding with gender bias. Therefore, future research should concentrate on developing novel debiasing techniques that can address this bias paradox, moving beyond simple quantitative balancing. Furthermore, establishing more comprehensive bias evaluation benchmarks that include intersectional dimensions and multilingual contexts, along with longitudinal tracking studies of bias evolution, will be key directions for advancing algorithmic fairness.

References

[1] Mirza, I., Jafari, A. A., Ozcinar, C., & Anbarjafari, G. (2025). Quantifying Gender Bias in Large Language Models Using Information-Theoretic and Statistical Analysis. Information, 16(5), 358.

[2] Chen, E., Zhan, R.-J., Lin, Y.-B., & Chen, H.-H. (2025). More Women, Same Stereotypes: Unpacking the Gender Bias Paradox in Large Language Models. arXiv preprint arXiv: 2503.15904.

[3] Kotek, H., Dockum, R., & Sun, D. Q. (2023). Gender bias and stereotypes in Large Language Models. In Proceedings of the 2023 ACM Collective Intelligence Conference (CI '23). Association for Computing Machinery.

[4] Salinas, A., Haim, A., & Nyarko, J. (2025). What's in a Name? Auditing Large Language Models for Race and Gender Bias. arXiv preprint arXiv: 2402.14875.

[5] Pikuliak, M. (2025). Gender Bench: Evaluation Suite for Gender Biases in LLMs. arXiv preprint arXiv: 2505.12054.

[6] Chaturvedi, S., & Chaturvedi, R. (2025). Who Gets the Callback? Generative AI and Gender Bias. arXiv: 2504.21400.

[7] Kong, H., Lee, S., Ahn, Y., & Maeng, Y. (2024). Gender Bias in LLM-generated Interview Responses. arXiv preprint arXiv: 2410.20739.

[8] Wilson, K., & Caliskan, A. (2024). Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. arXiv: 2407.20371.

[9] Xu, J., Li, G., & Jiang, J. Y. (2025). AI Self-preferencing in Algorithmic Hiring: Empirical Evidence and Insights. arXiv: 2509.00462.

[10] Cyberspace Administration of China, et al. (2023). Interim Measures for the Management of Generative Artificial Intelligence Services. Available: https: //www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

[11] Panarese, P. (2025). Algorithmic bias, fairness, and inclusivity: a multilevel framework for justice-oriented AI. AI & SOCIETY. doi: 10.1007/s00146-025-02451-2.

[12] Klein, L., & D'Ignazio, C. (2024). Data feminism for AI. In Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. ACM. doi: 10.1145/3630106.3658543.

[13] Wood, A. (2022). Disambiguating algorithmic bias: From neutrality to justice. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. ACM. doi: 10.1145/3600211.3604695.

[14] Markelius, A. (2024). An Empirical Design Justice Approach to Identifying Ethical Considerations in the Intersection of Large Language Models and Social Robotics. arXiv: 2406.06400.