AI Self-preferencing in Algorithmic Hiring: Empirical Evidence and Insights

Jiannan Xu[1], **Gujie Li**[2], **Jane Yi Jiang**[3]

[1]Robert H. Smith School of Business, University of Maryland
[2]School of Computing, National University of Singapore
[3]Max M. Fisher College of Business, Ohio State University
jiannan@umd.edu, gujieli@nus.edu.sg, jiang.3186@osu.edu

Abstract

As generative artificial intelligence (AI) tools become widely adopted, large language models (LLMs) are increasingly used on both sides of high-stakes decision processes, ranging from hiring to content moderation. This dual adoption raises a critical question: do LLMs systematically favor content that resembles their own outputs? Prior research in computer science has identified self-preference bias—the tendency of LLMs to favor their own generated content—but its real-world implications have not been empirically evaluated. We focus on the hiring context, where job applicants increasingly use LLMs to refine resumes, while employers integrate LLMs into their recruitment pipelines to screen those same resumes. Using a controlled resume correspondence experiment, we find that LLMs consistently prefer resumes generated by themselves over those written by humans or by alternative models, even when content quality is controlled. The bias against human-written resumes is particularly substantial, with self-preference rates ranging from 68 percent to 92 percent across a diverse set of commercial and open-source models. This behavior introduces a novel form of algorithmic unfairness, one that advantages users of specific AI tools and disadvantages others based on their tool choices or access. We further show that this bias can be significantly reduced through simple interventions based on LLMs' self-recognition capabilities, yielding reductions of over 60 percent. These findings highlight an emerging but previously overlooked risk in AI-assisted decision making and call for expanded frameworks of AI accountability that address not only demographic disparities but also biases in model-to-model interaction.

Extended Abstract

The rapid development and commercialization of generative artificial intelligence (AI) have made large language models more accessible, with growing adoption in professional tasks such as writing (Gero, Liu, and Chilton 2022) and coding (Moradi Dakhel et al. 2023). As LLMs become increasingly capable of producing fluent and persuasive text, job seekers are turning to them to enhance how they present qualifications to employers. Nearly 45% of job seekers have used generative AI to build, update, or improve their resumes (Revell 2024), highlighting a growing dependence on these tools to shape first impressions in the labor market.

On the employer side, AI tools are also transforming hiring. It is reported that 51% of companies are using AI in recruitment (ResumeBuilder 2024). In particular, LLMs are integrated into workflows such as resume screening (Gan, Zhang, and Mori 2024) or interview automation (Kim et al. 2024). While these applications promise efficiency and scalability, they elevate AI to a critical gatekeeping role—one that can significantly shape who gains access to employment opportunities.

A novel dynamic arises when both sides of the hiring process engage with AI: job seekers use LLMs to refine their resumes, while employers deploy the same or similar models to screen them. In such cases, LLMs effectively serve as both the evaluatee and the evaluator, creating conditions for a new form of algorithmic bias: self-preference, which refers to the tendency of LLMs to disproportionately favor content that resembles their own generative outputs (Panickssery, Bowman, and Feng 2024). It can manifest in two forms: (1) an LLM preferring its own output over human-written content, and (2) over content generated by another LLM. While these patterns have been observed in benchmark evaluation settings (Zheng et al. 2023; Panickssery, Bowman, and Feng 2024), their impact in high-stakes, real-world contexts like hiring remains underexplored.

In this paper, we provide the first empirical evidence that self-preference bias can distort candidate evaluations in algorithmic hiring. Specifically, we examine whether LLMs, when deployed as evaluators, systematically favor resumes they generated themselves over otherwise equivalent resumes written by humans or produced by alternative models. To test this, we conduct a large-scale resume correspondence experiment using a real-world dataset of $2,245$ human-written resumes, sourced from a professional resume-building platform. For each resume, we generate multiple counterfactual versions using a range of state-of-the-art LLMs, including GPT-4o, GPT-4o-mini, GPT-4-turbo, LLaMA 3.3-70B, Mistral-7B, Qwen 2.5-72B, and Deepseek-V3. By controlling content quality, we assess whether these LLMs exhibit systematic bias in favor of their own outputs when acting as evaluators. To mitigate AI self-preference, we propose two simple yet effective strategies based on the underlying mechanism of self-recognition—a model's ability to implicitly identify content it generated, achieving substantial reductions in bias.

References

Gan, C.; Zhang, Q.; and Mori, T. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315*.

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 1002–1019.

Kim, E.; Suk, J.; Kim, S.; Muennighoff, N.; Kim, D.; and Oh, A. 2024. LLM-AS-AN-INTERVIEWER: Beyond Static Testing Through Dynamic LLM Evaluation. *arXiv preprint arXiv:2412.10424*.

Moradi Dakhel, A.; Majdinasab, V.; Nikanjam, A.; Khomh, F.; Desmarais, M. C.; and Jiang, Z. M. J. 2023. GitHub Copilot AI pair programmer: Asset or Liability? *Journal of Systems and Software*, 203: 111734.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 68772–68802. Curran Associates, Inc.

ResumeBuilder. 2024. 7 in 10 Companies Will Use AI in the Hiring Process in 2025, Despite Most Saying It's Biased.

Revell, E. 2024. Nearly half of job seekers use AI to polish their resumes.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.