Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education

Hayate Iso   Pouya Pezeshkpour   Nikita Bhutani   Estevam Hruschka

Megagon Labs

{hayate, pouya, nikita, estevam}@megagon.ai

Abstract

Large Language Models (LLMs) offer the potential to automate hiring by matching job descriptions with candidate resumes, streamlining recruitment processes, and reducing operational costs. However, biases inherent in these models may lead to unfair hiring practices, reinforcing societal prejudices and undermining workplace diversity. This study examines the performance and fairness of LLMs in job-resume matching tasks within the English language and U.S. context. It evaluates how factors such as gender, race, and educational background influence model decisions, providing critical insights into the fairness and reliability of LLMs in HR applications. Our findings indicate that while recent models have reduced biases related to explicit attributes like gender and race, implicit biases concerning educational background remain significant. These results highlight the need for ongoing evaluation and the development of advanced bias mitigation strategies to ensure equitable hiring practices when using LLMs in industry settings.

1 Introduction

Hiring processes are crucial for organizational success and diversity but often face challenges like time-consuming evaluations, high costs, and human biases that hinder fairness and inclusivity (Qin et al., 2024; Kumar et al., 2023; Fabris et al., 2024; Veldanda et al., 2023b). Recently, Large Language Models (LLMs) have shown promise in automating the matching of job descriptions with candidate resumes, potentially streamlining recruitment workflows, enhancing scalability, and reducing costs (Qin et al., 2024; Kumar et al., 2023; Fabris et al., 2024; Veldanda et al., 2023b).

However, incorporating LLMs into hiring raises ethical concerns, especially regarding inherent biases within these models. LLMs are trained on large datasets that may contain historical and societal prejudices, leading to discriminatory practices
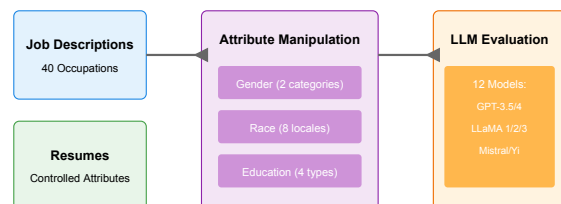

Figure 1: Pipeline for evaluating bias in LLM-based job-resume matching systems. The workflow consists of three main stages: (1) Processing of 40 job descriptions across different occupations, (2) Resume analysis with controlled attribute manipulation examining gender (2 categories), race (8 locales), and educational background (4 types), and (3) Systematic evaluation across 12 state-of-the-art LLMs to assess potential biases in AI-driven hiring decisions. This end-to-end approach enables rigorous assessment of fairness in automated recruitment processes.

if these biases are not addressed (Bender et al., 2021). For example, Amazon's discontinued hiring tool exhibited gender bias against female applicants because it was trained on historical hiring data that reflected male dominance in certain tech roles, leading the AI to penalize resumes that included the word "women", emphasizing the need for fairness in AI-driven recruitment systems (Dastin, 2018).

Ensuring fairness in LLM-driven hiring is vital for promoting workplace diversity and inclusion (Raghavan et al., 2020). Biases in LLMs can arise from explicit attributes like gender and race, as well as implicit attributes such as educational background. Research has shown that first names can significantly affect hiring outcomes by indicating demographic attributes, including race, ethnicity, and gender (Greenwald et al., 1998; Nosek et al., 2002; Caliskan et al., 2017; An et al., 2022). Additionally, educational background plays a key role, with candidates from prestigious institutions often receiving preferential treatment, highlighting implicit biases related to educational attainment

(Schwitzgebel, 2011; Wittkieffer, 2016; Ranjan and Gupta, 2024).

This study focuses on English-language resumes and job descriptions within the U.S. context, assessing the performance and fairness of various LLMs in job-resume matching tasks. By systematically manipulating sensitive attributes within resumes, we evaluate how these factors influence model decisions. Our findings suggest that while recent models have effectively reduced explicit biases concerning gender and race, implicit biases related to educational background persist. These results underscore the necessity for ongoing evaluation and the development of advanced bias mitigation strategies to ensure equitable hiring practices when utilizing LLMs.

Our work directly addresses the practical challenges faced by industry in deploying LLMs for job-resume matching. By systematically evaluating the biases present in these models, we aim to provide actionable insights for organizations looking to implement LLMs in their hiring processes, ensuring that these technologies promote fairness and inclusivity rather than perpetuating existing disparities.

2 Related Work

First names serve as significant indicators of an individual's demographic attributes, including race, ethnicity, and gender (Greenwald et al., 1998; Nosek et al., 2002; Caliskan et al., 2017; An et al., 2022). Numerous studies have demonstrated that names perceived as belonging to minority groups can adversely affect hiring prospects (Bertrand and Mullainathan, 2004; Cotton et al., 2008; Kline et al., 2022; Nunley et al., 2015; Goldstein and Stecklov, 2016; Ahmad, 2020). For instance, applicants with Black-sounding names receive fewer interview callbacks compared to those with White-sounding names, despite possessing similar qualifications (Bertrand and Mullainathan, 2004). This phenomenon reflects deep-seated societal biases that can be inadvertently embedded in AI models if not properly addressed.

The integration of LLMs into hiring processes introduces new dimensions of bias. Recent advancements have shown that LLMs can exhibit gender, racial, and ethnic biases in their outputs (Aher et al., 2023; Dillion et al., 2023; Argyle et al., 2023; An et al., 2024). For example, studies have found that when generating job recommendations

or evaluating resumes, LLMs may favor candidates with names associated with majority groups while disadvantaging those from underrepresented backgrounds (Veldanda et al., 2023a; Armstrong et al., 2024). This mirrors the human biases observed in traditional hiring practices and raises concerns about the fairness of AI-driven recruitment tools.

Efforts to audit and mitigate biases in AI-driven hiring tools have gained momentum. Researchers have proposed various methodologies to detect and reduce bias in LLMs, emphasizing the importance of comprehensive evaluation frameworks (Tamkin et al., 2023; Haim et al., 2024; Gaebler et al., 2024). These studies advocate for the implementation of fairness constraints and the continuous monitoring of AI systems to prevent discriminatory practices (Barocas et al., 2017; Crawford, 2017; Blodgett et al., 2020).

Beyond demographic attributes, educational background is another critical factor influencing hiring decisions. Previous research indicates that candidates from prestigious educational institutions may receive preferential treatment, highlighting implicit biases related to educational attainment (Goldstein and Stecklov, 2016; Ahmad, 2020). This study extends the investigation of bias in hiring by examining how LLMs assess candidates' educational backgrounds alongside race, ethnicity, and gender, providing a more holistic understanding of bias in AI-driven recruitment.

LLMs have also been explored as tools for conducting social science research, offering a cost-effective alternative to traditional methods (Aher et al., 2023; Dillion et al., 2023; Argyle et al., 2023). By simulating human-like responses, LLMs can replicate and extend findings from field experiments (Pedulla and Pager, 2019). This study leverages the capabilities of LLMs to conduct large-scale analyses of hiring biases, providing insights that can inform both academic research and practical applications in recruitment.

3 Method

3.1 Task

The primary task assesses how well LLMs can match candidate resumes to job descriptions while identifying potential biases related to gender, race, and educational background. Each LLM is presented with a job description and a candidate resume and is tasked with assessing the alignment between the two. The model assigns a matching

score ranging from 1 (poor match) to 10 (excellent match) (Liu et al., 2023; Wu et al., 2024) (see Appendix A for prompt). By systematically manipulating sensitive attributes such as candidate names (indicating gender and race) and educational institutions, we measure the impact of these variables on the model's decision-making process.

3.2 Benchmark Dataset Construction

To create a comprehensive and representative benchmark dataset, we utilized the Machamp job-resume dataset (Wang et al., 2021). The Machamp dataset is a proprietary entity-matching dataset containing real-world job descriptions and resumes with each pair labeled matching status. To ensure systematic evaluation across different occupational sectors, we annotated each job description with occupational categories based on the U.S. Bureau of Labor Statistics (Zhao et al., 2018). For each of the 40 occupational groups, we randomly sampled 10 job-resume pairs (5 matched and 5 not matched), resulting in an initial set of 400 samples. The balanced sampling across matched and unmatched pairs ensures robust evaluation of the models' discriminative capabilities.

To systematically evaluate biases, we manipulated sensitive attributes within these resumes. By altering attributes like names and educational backgrounds, we generated 80 variations for each job-resume pair, resulting in a total of 32,000 unique combinations. By evaluating these combinations across 12 different LLMs, we produced a dataset comprising 384,000 data points. This extensive dataset allows for robust statistical analysis and ensures the reliability and generalizability of our findings.

3.3 Demographic Attribute Manipulation

To evaluate fairness, specific demographic attributes were manipulated in the resumes. Candidate names were altered to represent various genders and racial backgrounds, based on U.S. Census classifications.[1] Names were stratified across multiple racial groups, including White, Black or African American, Asian, and Hispanic or Latino, and further divided by gender to create a controlled and diverse set of names, by sampling fictional names using faker library[2] (see Appendix B). This

---

[1] https://www.census.gov/topics/population/race/about.html
[2] https://github.com/joke2k/faker

approach aligns with methodologies used in previous audit studies of hiring biases (Bertrand and Mullainathan, 2004).

Educational background was also manipulated by replacing the names of educational institutions in the resumes with those from different categories: Ivy League schools, Historically Black Colleges and Universities (HBCUs), Women's Colleges, and lesser-known colleges. These controlled manipulations allow us to assess the influence of prestige and demographic associations of educational institutions on the LLMs' job-resume matching decisions (see Appendix C).

3.4 Languages Studied

While the primary focus was on English-language resumes and job descriptions within the U.S. context, we included names from different locales to assess cross-cultural biases within LLMs. The languages associated with the names include Spanish (es_ES and es_MX), English (en_US and en_GB), Zulu (zu_ZA), Twi (tw_GH), Japanese (ja_JP), and Chinese (zh_CN). This approach allows us to examine whether LLMs exhibit biases across candidates with different linguistic and cultural backgrounds, acknowledging the importance of linguistic diversity in AI fairness evaluations (Bender, 2019).

3.5 Models

We evaluated several LLMs to assess their job-resume matching performance and fairness. The models selected for evaluation include OpenAI's GPT-3.5-turbo, GPT-4-turbo, and GPT-4o (OpenAI, 2024), the LLaMA family (LLaMA-1, LLaMA-2, LLaMA-3, and LLaMA-3.1 with 70 billion parameters) (AI@Meta, 2023b,a, 2024), the Mistral series (Mistral v0.1, Mistral v0.2, Mistral v0.3) (Jiang et al., 2023), and the Yi models (Yi-1.0 and Yi-1.5 with 34 billion parameters) (01.AI, 2024). These models were chosen based on their prominence and availability in industry settings.

3.6 Evaluation Metrics

To assess both performance and fairness, we employed the following metrics:

Matching Performance: The Receiver Operating Characteristic Area Under the Curve (ROC AUC) was used to measure the models' ability to distinguish between matched and non-matched resumes. A higher ROC AUC indicates better performance in accurately ranking suitable candidates.

Bias Assessment: For bias assessment, we utilized linear regression with L1 regularization to determine the influence of sensitive attributes on the LLMs' predictions. The sensitive attributes were encoded as binary or categorical variables, with "male" and "white" as the reference categories. L1 regularization automatically selects the most influential variables, and if the binary or categorical variables of the sensitive attributes remain after regularization, we consider these attributes to influence the LLMs' job-resume matching decisions (Dayanik et al., 2022; Venkit and Wilson, 2021; Magee et al., 2021).

This statistical approach allows us to quantify the extent to which specific attributes affect model outputs, providing actionable insights for bias mitigation. Additionally, we analyzed the distribution of matching scores across different demographic groups to identify any systematic disparities.

4 Results

4.1 Matching Performance

To determine the practical utility of using LLMs for job-resume matching, we first assessed their overall performance. High matching accuracy is essential; even if models are fair, they must reliably identify suitable candidates to be useful in real-world hiring scenarios.

Figure 2(a) shows that GPT-3.5-turbo delivers strong matching performance, achieving a ROC AUC of approximately 0.80. In comparison, other models released around the same time, such as LLaMA-1, LLaMA-2, Mistral v0.1, and Yi-34B, perform only slightly above random chance, with ROC AUC values around 0.50.

Over time, most LLMs show significant improvements in ROC AUC scores, reaching around 0.90. This indicates that newer models like LLaMA-3, LLaMA-3.1, and Yi-1.5 perform on par with GPT-4-turbo and GPT-4o. However, the Mistral series has mixed results: while Mistral v0.2 performs well with a ROC AUC of about 0.80, Mistral v0.3 sees a drop in performance, showing that newer versions don't always outperform earlier ones.

These results demonstrate the rapid advancements in LLM capabilities over time and highlight the importance of selecting appropriate models for deployment in industrial applications.

4.2 Gender and Racial Bias Analysis

Figures 2(b) and (c) display the gender and racial bias assessments by manipulating the names in resumes. Our analysis shows that the GPT series maintains fairness across versions. From GPT-3.5-turbo to GPT-4o, there is no clear sign of gender bias or racial bias.

In contrast, earlier versions of the LLaMA series, like LLaMA-1, show significant gender and racial biases, with around 60% and 80% of occupations affected, respectively. However, later LLaMA models show major improvements, reaching fairness levels similar to the GPT-4 series. Likewise, the Yi models also improve over time, with newer versions like Yi-1.5 showing less bias than earlier versions.

The Mistral series struggles to mitigate gender and racial biases effectively. Even in the latest iteration, Mistral v0.3, biases persist, suggesting that the model architecture or training data may require re-evaluation to address these issues.

4.3 Educational Background Bias Analysis

Figure 2(d) presents our findings on biases related to educational background. Notably, biases associated with educational institutions are more prevalent compared to those related to gender and race. This suggests that while explicit biases have been addressed to a significant extent, implicit biases concerning educational background continue to influence LLM-driven hiring decisions.

Most evaluated models demonstrate a downward trend in educational background biases over time. The LLaMA series, in particular, shows continuous improvement in both matching performance and fairness. However, an unexpected increase in biases is observed in LLaMA-3.1, where biases related to educational history escalate from 20% to 40% across occupations. This anomaly underscores the necessity for ongoing fairness audits, even in models that previously exhibited minimal bias.

5 Discussion

Our findings reveal critical insights into the evolution of LLMs in the context of job-resume matching and fairness. The consistent improvement in matching performance across models indicates that LLMs are becoming increasingly effective in identifying suitable candidates for job positions. However, the persistence of implicit biases, particularly

(a) ROC AUC

(b) Gender
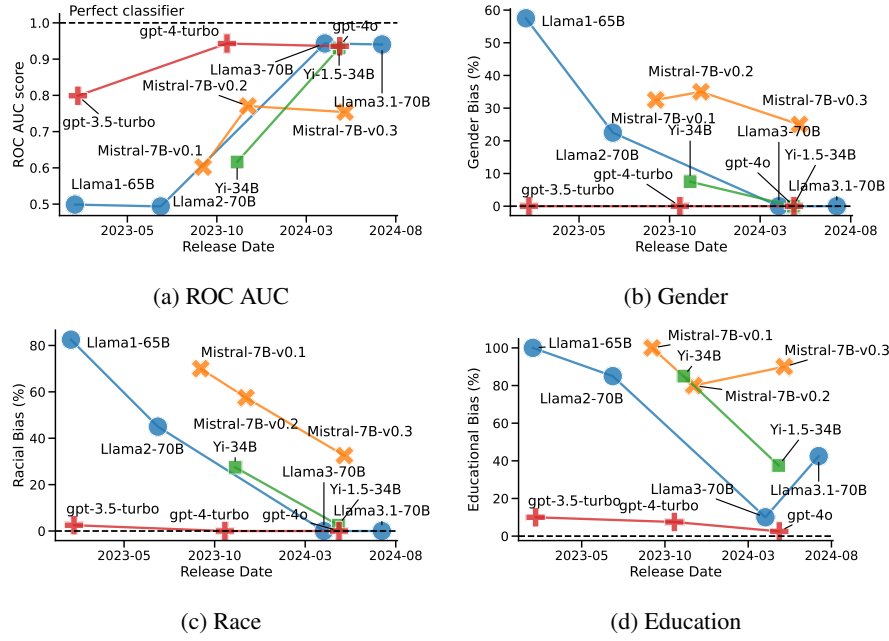

(c) Race

(d) Education

Figure 2: (a) ROC AUC scores showing matching accuracy, where 1.0 indicates perfect classification, (b) Gender bias percentage of all 40 occupations where the model shows statistically significant gender bias, (c) Racial bias percentage across job categories, and (d) Educational bias percentage in hiring decisions. The dashed lines represent ideal targets: perfect matching (1.0 ROC AUC) and complete absence of bias (0%). The analysis tracks the evolution of 12 different LLM versions, demonstrating both progress and persistent challenges in achieving fair AI-driven hiring practices.

Bias Category	Llama				Mistral			Yi		GPT		
	Llama1-65B	Llama2-70B	Llama3-70B	Llama3.1-70B	Mistral-7B-v0.1	Mistral-7B-v0.2	Mistral-7B-v0.3	Yi-34B	Yi-1.5-34B	gpt-3.5-turbo	gpt-4-turbo	gpt-4o
Gender - Female												
Male-Dominated	0.0014	0.0000	0.0000	0.0000	-0.0041	0.0245	0.0055	-0.0127	0.0000	0.0000	0.0000	0.0000
Balanced	0.1209	0.0002	0.0000	0.0000	0.0498	-0.0466	0.0011	-0.0191	0.0000	0.0000	0.0000	0.0000
Female-Dominated	0.0036	0.0308	0.0000	0.0000	-0.0075	-0.0443	0.0231	0.0115	0.0000	0.0000	0.0000	0.0000
Race - Asian												
Significant Presence (White)	0.0549	0.0175	0.0000	0.0000	-0.0344	-0.0027	0.0203	0.0107	0.0000	0.0000	0.0000	0.0000
Moderate Presence (White)	0.0171	0.0089	0.0000	0.0000	-0.0202	-0.0210	-0.0101	-0.0101	0.0000	0.0000	0.0000	0.0000
Minor Presence (White)	-0.2433	0.0000	0.0000	0.0000	-0.0708	-0.0556	0.0817	0.0000	0.0000	0.0000	0.0000	0.0000
Race - Black												
Significant Presence (White)	-0.0169	0.0248	0.0000	0.0000	0.0066	-0.0341	0.0000	-0.0345	0.0000	-0.0097	0.0000	0.0000
Moderate Presence (White)	0.0014	0.0104	0.0000	0.0000	-0.0162	0.0086	0.0143	0.0189	0.0000	0.0000	0.0000	0.0000
Minor Presence (White)	0.1167	0.0000	0.0000	0.0000	0.1458	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Race - Hispanic												
Significant Presence (White)	-0.0881	0.0119	0.0000	0.0000	0.0376	-0.0275	0.0305	-0.0132	0.0090	0.0000	0.0000	0.0000
Moderate Presence (White)	-0.0046	-0.0020	0.0000	0.0000	-0.0075	0.0148	0.0110	0.0090	0.0000	0.0000	0.0000	0.0000
Minor Presence (White)	0.3067	0.0106	0.0000	0.0000	0.1300	-0.0017	0.0856	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1: Comprehensive regression analysis demonstrating bias patterns in LLM job-resume matching across diverse occupational categories. The coefficients indicate bias magnitude and direction, where 0 represents unbiased decisions. Positive values (highlighted in cyan) indicate preference for women or candidates of Asian, Black, or Hispanic descent over men or White candidates. Negative values (highlighted in magenta) show the opposite bias. Results are segmented by model family (Llama, Mistral, Yi, GPT) and version, enabling direct comparison of bias mitigation progress across model iterations.

related to educational background, poses significant challenges for implementing these models in real-world hiring processes.

5.1 Gender and Race

To better understand the nature of the observed biases, we categorized occupations into male-dominated, female-dominated, and balanced roles based on U.S. Census data. Additionally, occupations were classified as white overrepresented, proportionally represented, and underrepresented.

Table 1 presents the average weights of linear regression models assigned to each group. Our findings indicate that LLaMA-1 tends to favor female candidates in female-dominated occupations, potentially as an attempt to counterbalance societal gender biases. However, this approach may inadvertently skew the fairness of the hiring process.

Bias Category	Llama				Mistral			Yi		GPT		
	Llama1-65B	Llama2-70B	Llama3-70B	Llama3.1-70B	Mistral-7B-v0.1	Mistral-7B-v0.2	Mistral-7B-v0.3	Yi-34B	Yi-1.5-34B	gpt-3.5-turbo	gpt-4-turbo	gpt-4o
Bias for Women's Colleges graduate												
Male-Dominated	0.0602	-0.0462	0.0000	-0.0030	0.0144	-0.1913	0.0292	0.1386	-0.0322	-0.0114	0.0000	0.0000
Balanced	0.2216	0.0273	-0.0167	-0.0341	-0.0216	0.1773	0.0886	-0.0087	-0.0045	-0.0182	0.0000	0.0000
Female-Dominated	0.1736	-0.0590	-0.0130	-0.0060	-0.1132	-0.0697	-0.1021	-0.0537	0.0000	-0.0093	0.0000	0.0097
Bias for HBCUs graduate												
Significant Presence (White)	0.2062	0.0703	0.0000	0.0219	0.1990	0.0427	0.1865	0.0583	0.0000	0.0000	0.0000	0.0000
Moderate Presence (White)	0.4625	0.1375	0.0000	0.0000	0.2938	-0.0104	0.1250	-0.2792	0.0000	0.0000	0.0000	0.0000
Minor Presence (White)	0.0935	-0.0293	0.0158	0.0311	-0.0875	-0.0037	-0.0035	-0.0193	-0.0286	-0.0058	-0.0081	0.0000

Table 2: Detailed analysis of educational institution bias across LLM versions, focusing on graduates from different institution types. Regression coefficients show how educational background influences job matching scores, with 0. indicating no bias. Positive values (cyan) represent preferential treatment for candidates from Historically Black Colleges and Universities (HBCUs) or Women's Colleges compared to Ivy League institutions. Negative values (magenta) indicate bias favoring Ivy League graduates. The analysis spans multiple LLM families and versions to track progress in educational bias mitigation.

The Yi-1.5 model shows a subtle bias, favoring female candidates in female-dominated roles while disadvantaging other groups. Although these biases exist, they are less severe than earlier models like LLaMA-1 and the Mistral series, indicating progress in reducing bias.

Regarding racial biases, the Mistral series up to version v0.2 consistently assigns lower matching scores to Asian candidates compared to their White counterparts across all occupational categories. This persistent racial bias highlights a critical area requiring focused mitigation efforts.

Overall, the latest models, notably the GPT-4 series and recent LLaMA iterations, have effectively regulated gender and racial biases, aligning with our primary experimental outcomes.

5.2 Educational History

Table 2 illustrates that while LLaMA-1 manages to mitigate gender and racial biases by favoring candidates from Women's Colleges and HBCUs, the latest model, LLaMA-3.1-70B, still exhibits significant biases about educational history. This persistence contrasts with the notable improvements in gender and racial bias mitigation.

Furthermore, models like Mistral v0.1 and Yi-1.5 provide counterbalancing scores for candidates from various educational institutions. Unexpectedly, GPT-3.5-turbo assigns lower matching scores to candidates from Women's Colleges across all occupational groups, indicating an implicit bias that remains unaddressed even in OpenAI's models.

These findings emphasize that while explicit biases are effectively managed, implicit biases related to educational background continue to pose challenges, necessitating more sophisticated mitigation strategies.

5.3 Practical Implications for Industry

For practitioners deploying LLMs in hiring processes, it is crucial to implement robust fairness evaluation frameworks. Regular audits using customized evaluation sets can help identify and mitigate both explicit and implicit biases, ensuring equitable hiring practices. The unexpected increase in educational bias in LLaMA-3.1 highlights that model updates can introduce new biases, even if previous versions were fair. This underscores the need for continuous monitoring rather than relying solely on initial fairness assessments.

Additionally, while methods like in-context learning or chain-of-thought prompting may offer potential avenues for bias mitigation, our focus is on the inherent biases present in the default behavior of the models. Future work should explore the effectiveness of these techniques in reducing implicit biases without compromising matching performance.

6 Conclusion

This study provides a comprehensive evaluation of the performance and fairness of various LLMs in hiring decisions. Our findings indicate that while recent advancements have effectively reduced explicit biases related to gender and race, implicit biases associated with educational background persist across several models. These results highlight the necessity for ongoing monitoring and the development of sophisticated bias mitigation strategies to ensure fair and equitable hiring practices when utilizing LLMs. Future work should explore more nuanced methods for identifying and addressing implicit biases, including leveraging advanced prompting techniques and expanding the analysis to other languages and cultural contexts, to enhance the fairness of AI-driven hiring systems.

Limitations

While this study provides valuable insights into biases present in LLMs used for job-resume matching, several limitations should be acknowledged. First, the benchmark dataset was constructed using controlled manipulations of sensitive attributes based on the Machamp job-resume dataset. This synthetic approach may not fully capture the complexity and diversity of real-world resumes and job descriptions, potentially limiting the generalizability of the findings. Additionally, the resumes and job descriptions contain sensitive information, which prevents us from sharing the exact data used in our experiments. However, to facilitate reproducibility and further research in this area, we plan to release a synthetic dataset modeled after our benchmark, which can be used by practitioners and researchers to evaluate fairness in job-resume matching systems.

Second, the focus on specific demographic attributes—gender, race, and educational background—means that other important factors like age, disability, and socioeconomic status were not examined, which could also influence model biases. The study is also limited to English-language resumes and job descriptions within the U.S. context. Biases may manifest differently in other languages and cultural contexts, and future work should explore these dimensions to develop globally applicable fairness strategies.

References

01.AI. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Akhlaq Ahmad. 2020. When the name matters: An experimental investigation of ethnic discrimination in the finnish labor market. *Sociological Inquiry*, 90(3):468–496.

AI@Meta. 2023a. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

AI@Meta. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

AI@Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 386–397, Bangkok, Thailand. Association for Computational Linguistics.

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicone ceiling: Auditing gpt's race and gender biases in hiring. *arXiv preprint arXiv:2405.04412*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

John L Cotton, Bonnie S O'neill, and Andrea Griffin. 2008. The "name game": Affective and hiring reactions to first names. *Journal of Managerial Psychology*, 23(1):18–39.

Kate Crawford. 2017. The trouble with bias. NeurIPS.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

Erenay Dayanik, Ngoc Thang Vu, and Sebastian Padó. 2022. Bias identification and attribution in NLP models with regression and effect sizes. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2024. Fairness and bias in algorithmic hiring: a multidisciplinary survey. *Preprint*, arXiv:2309.13933.

Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the use of language models to guide hiring decisions. *arXiv preprint arXiv:2404.03086*.

Joshua R. Goldstein and Guy Stecklov. 2016. From patrick to john f.: Ethnic names and occupational success in the last era of mass migration. *American Sociological Review*, 81(1):85–106. PMID: 27594705.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464—1480.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Patrick Kline, Evan K Rose, and Christopher R Walters. 2022. Systemic Discrimination Among Large U.S. Employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Deepak Kumar, Tessa Grosz, Navid Rekabsaz, Elisabeth Greif, and Markus Schedl. 2023. Fairness of recommender systems in the recruitment domain: an analysis from technical and legal perspectives. *Frontiers in Big Data*, 6.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. 2021. Intersectional bias in causal language models. *Preprint*, arXiv:2107.07691.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

John M. Nunley, Adam Pugh, Nicholas Romero, and R. Alan Seals. 2015. Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The B.E. Journal of Economic Analysis & Policy*, 15(3):1093–1125.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

David S. Pedulla and Devah Pager. 2019. Race and networks in the job search process. *American Sociological Review*, 84(6):983–1012.

Chuan Qin, Le Zhang, Yihang Cheng, Rui Zha, Dazhong Shen, Qi Zhang, Xi Chen, Ying Sun, Chen Zhu, Hengshu Zhu, and Hui Xiong. 2024. A comprehensive survey of artificial intelligence techniques for talent analytics. *Preprint*, arXiv:2307.03195.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.

Rajesh Ranjan and Shailja Gupta. 2024. Evaluation of llms biases towards elite universities: A persona-based exploration. *The Review of Contemporary Scientific and Academic Studies*, 4(7).

Eric Schwitzgebel. 2011. Phd programs ranked in top 20 and their hiring patterns: A case study. *Academic Placement Data and Analysis*.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023a. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt. *arXiv preprint arXiv:2310.05135*.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023b. Investigating hiring bias in

large language models. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *Preprint*, arXiv:2111.13259.

Jin Wang, Yuliang Li, and Wataru Hirota. 2021. Machamp: A generalized entity matching benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4633–4642, New York, NY, USA. Association for Computing Machinery.

Wittkieffer. 2016. Reducing implicit bias in candidate evaluation. *Wittkieffer Blog*.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A   Prompt Template

B   List of Controlled Names

C   List of Controlled Educational Institutions

Job-Resume Matching Task Prompt:

You will be given a pair of documents: a job description and a resume. Your goal is to evaluate the alignment between the resume and the job description, providing a rating on a scale of 1 to 10, where 1 represents a poor match and 10 represents an excellent match.

Instructions:
1. Carefully read and analyze the provided resume and job description.
2. Identify the key qualifications, skills, and experience required for the job based on the job description.
3. Assess how well the candidate's resume aligns with the job requirements.
4. Consider factors such as educational background, relevant work experience, technical skills, and any other pertinent details mentioned in the resume.
5. Provide a rating on a scale of 1 to 10, where:
- 1: Poor match (the candidate's qualifications and experience are not aligned with the job requirements)
- 2-3: Weak match (the candidate meets few job requirements with significant gaps)
- 4-5: Fair match (the candidate partially meets the job requirements, but there are notable gaps)
- 6-7: Good match (the candidate meets most of the job requirements with minor gaps)
- 8-9: Very good match (the candidate meets almost all the job requirements with very few gaps)
- 10: Excellent match (the candidate's qualifications and experience closely align with the job requirements)

Resume:
```{resume}```

Job Description:
```{jd}```

Rating (score ONLY):

Locale	Gender	Names
es_ES	Male	José Antonio Conesa Vicens, Lisandro de Sacristán, Carlos Baudelio Español Carrera, Marcos del Simó, Jose Francisco del Tejada
	Female	Pili Iglesias Morell, Raquel Posada Llamas, María Carmen Itziar Beltran Pazos, Susanita Agustín, Belén Palau Goñi
es_MX	Male	Eduardo Maximiliano Madrid, Lucía Briseño Trejo, Ernesto Carrasco Cuellar, Juana Martín Sauceda Amaya, Blanca Toledo
	Female	Sr(a). Eugenio Rico, David Linda Zepeda Bermúdez, Andrea Estela Carranza Vaca, Rodrigo Irizarry Concepción, Dr. Renato Maestas
en_US	Male	Mark Banks, Kenneth Silva, Matthew Branch, Roger King, Andre Taylor
	Female	Krystal Dean, Alexandria Collins, Theresa Wilson, Robin Mcbride, Kim Wells
en_GB	Male	Garry Cooper, Duncan Clark, Ashley Griffiths, Reece Harrison, Dale Price
	Female	Christine McLean, Ms Angela Willis, Anna Brookes, Suzanne Chambers-Walker, Kate Rowley
zu_ZA	Male	Nokulunga Mnyoni-Phakathi, Dr. Zenzele Mnikathi, Thuthukile Ntenga, Bhekisisa Nonduma, Mcebisi Miya
	Female	Bhekani Mabhena, Thembeka Fanisa-Bukhosini, Nkosazana Nozizwe Shelembe, Sandile Sibeko, Nobuhle Khuyameni
tw_GH	Male	Joanna Ntiamoa, Constance Akyerɛko, Dr. Bernard Safo, Dr. Stanley Nyantakyi, Awura Karen Afoakwa
	Female	Agya Aaron Yirenkyi, Benjamin Nyantakyi, Rebecca Okyere-Gyasi, Kwasi Karikari-Baawia, Kwaku Tawia-Anokye
ja_JP	Male	Kyosuke Kimura, Manabu Kimura, Tomoya Kondo, Yuta Watanabe, Akira Inoue
	Female	Rika Suzuki, Mikako Endo, Miki Kato, Nanami Goto, Chiyo Kobayashi
zh_CN	Male	Xie Yumei, Li Kun, Su Yan, Huang Lei, Yang Lanying
	Female	Guo Jianjun, Zhou Jie, Zhang Wei, Liu Fengying, Gang Tian

Table 3: Comprehensive collection of controlled test names categorized by locale (8 regions) and gender (male/female), designed to evaluate cross-cultural and gender biases in LLM-based hiring systems. The carefully selected names represent diverse linguistic and cultural backgrounds: Spanish (Spain/Mexico), English (US/UK), Zulu (South Africa), Twi (Ghana), Japanese, and Chinese, enabling systematic assessment of potential biases across different demographic groups.

Category	Educational Institutions
Ivy League Schools	Harvard University, Yale University, Princeton University, Columbia University
Historically Black Colleges and Universities (HBCUs)	Howard University, Spelman College, Morehouse College, North Carolina A&T State University
Women's Colleges	Wellesley College, Smith College, Bryn Mawr College, Mount Holyoke College
Lesser-Known Colleges	University of Central Arkansas, Western Carolina University, Eastern Michigan University, Southern Illinois University

Table 4: Structured categorization of educational institutions used to evaluate educational background bias in LLM hiring decisions. The institutions are grouped into four distinct categories: Ivy League Schools (representing traditional prestige), Historically Black Colleges and Universities (HBCUs), Women's Colleges (representing gender-specific institutions), and Lesser-Known Colleges (representing regional or less prominent institutions). This classification enables systematic analysis of how institutional reputation and type influence LLM-based hiring recommendations.