BROOKINGS

RESEARCH

Gender, race, and intersectional bias in AI resume screening via language model retrieval

Kyra Wilson and Aylin Caliskan

April 25, 2025

- → Though the use of AI in the hiring process has continued to grow, few laws have been passed that require auditing of these systems to ensure they do not discriminate against some applicants.
- → In a simulation of resume screening, some systems resulted in significant gender and racial discrimination, especially for Black men.
- → Increased protections and transparency with these systems could protect against harmful effects, especially with intersectional identities, and empower applicants to act in the event of discrimination.

Artificial intelligence (AI) is now firmly a part of the hiring process. Some candidates use large language models (LLMs) to write cover letters and resumes, while employers use various proprietary AI systems to evaluate candidates. Recent estimates found as many as 98.4% of Fortune 500 companies ↗ leverage AI in the hiring process, and one company saved over a million dollars ↗ in a single year by incorporating AI into its interview process ↗. While this figure is lower for non-Fortune 500 companies, it is still expected to grow from 51% to 68% ↗ by the end of 2025 because of the potential time and cost savings for employers. However, when these systems are deployed at scale, they can introduce a myriad of biases that can potentially impact millions of job seekers annually.

With more companies choosing to use AI in employment screening, these systems should face more scrutiny to ensure they comply with laws against discrimination. The Equal Employment Opportunity Commission ↗ (EEOC) enforces various laws that make it illegal for employers to discriminate against employees or job applicants on the basis of their race, color, religion, sex (including gender identity, sexual orientation, and pregnancy), national origin, age (40 or older), disability, or genetic information. According to guidance ↗ published by the EEOC in 2022, using AI systems does not change employers' responsibility to ensure their selection procedures are not discriminatory, either intentionally or unintentionally. While this guidance was removed when President Donald J. Trump assumed office in January 2025, there has been no change in anti-discrimination laws. Investigations ↗ into AI hiring systems continue to be an important tool in evaluating the risks these systems pose and discovering ways to mitigate their potential societal harms. For example, in the U.K., an audit of AI recruitment software revealed multiple fairness and privacy vulnerabilities; in response to these findings, the Information Commissioner's Office issued ↗ nearly 300 recommendations for ways to improve hiring practices that model providers and developers used in their products.

WHY EMPIRICAL INVESTIGATIONS INTO AI USE IN HIRING ARE LIMITED

Empirical investigations of AI hiring systems are limited, despite being needed to avert discrimination. AI hiring systems are often proprietary, meaning independent researchers and auditors do not have the access necessary for relevant inquiry and testing. One study ↗ investigated public statements made by developers of these systems and found that, while many claimed to reduce bias and discrimination in hiring, they provided little evidence about how this was accomplished. Often reflective ↗ of prior inequalities in hiring processes based on historical discrimination, these biases likely propagate (https://www.brookings.edu/articles/challenges-for-mitigating-bias-in-algorithmic-hiring/) into the systems, which then replicate or even amplify them.

THE USE OF LARGE LANGUAGE MODELS IN HIRING

We recently conducted our own study ↗, which aimed to investigate large language models (LLMs) used for hiring at scale. LLMs are of particular interest in this domain as they are not only being used as part of proprietary systems for AI-mediated hiring but are also often open source and thus more widely available for public usage and testing. This means a simulation of how open-source LLMs perform hiring tasks could approximate their effect on proprietary systems as well, providing crucial insights into whether these systems are potentially discriminatory.

Our study investigated LLM-mediated hiring processes by simulating resume screening, an initial stage of candidate review where an automated system reduces a large set of applicants to identify those who are most suited for a particular role. By looking at evaluations of the same resumes with different names (signaling different gender or racial identities) in the context of a particular job, we determined whether an applicant's presumed social identity is a relevant factor in predicting whether they are suitable for a position. Using names to signal social identity is a common approach which has revealed discrimination in mortgage lending ↗, online ad delivery ↗, as well as hiring ↗. We used a set of over 550 unique job descriptions (covering nine diverse occupations) and 550 unique resumes, each augmented with 80 different names highly associated with Black women, Black men, white women, or white men.

Our procedure was inspired by real-world retrieval systems, in which a large set of documents is ranked based on how relevant the information they contain is to a user's request; then, the user only needs to look at the most highly ranked documents to find a suitable answer for their request. In our study, job descriptions were analogous to user requests and resumes were analogous to documents, and the suitability of a resume was determined by calculating its similarity to a particular job description using three distinct LLMs. Additional information about the study methodology is available in the Appendix.

RESULTS

The results of the research showed clear evidence of significant discrimination based on gender, racial identities, and their intersections. Out of 27 tests for discrimination across three LLMs and nine occupations, gender bias was evident: Men's and women's names were selected at equal rates in only 37% of cases. In the rest,

resumes with men's names were favored 51.9% of the time, while women's names were favored just 11.1% of the time. Racial bias was even more pronounced—resumes with Black- and white-associated names were selected at equal rates in only 6.3% of tests. White-associated names were preferred in 85.1% of cases, while Black-associated names led in just 8.6%. Disparities in resume selections did not necessarily correlate with existing disparities in workforce employment for gender or race, suggesting that using AI screening mechanisms could either alter or increase disparities in sectors and occupations where they do not already exist.

While these results offer evidence for significant differences based on single axes of identity, societal harm is often better quantified when considering intersectional identities. This lens considers how the combination of multiple identities can produce unique experiences and outcomes that differ from those associated with any single identity on its own. When considering gender and race together, we found that names associated with Black men led to the most significant disparities in outcomes—compared to resumes with Black women's names, they were selected only 14.8% of the time, and compared to white men's names, they were selected 0% of the time. Equal preference was found in 18.5% and 0% of comparisons between these groups respectively. This unique harm at the intersection of gender and race reflects broader societal patterns, where Black men are often the most disadvantaged group ↗ in employment settings. This finding is obscured when examining only single axes of identity, which would potentially underestimate the real-world harm and discrimination these models can perpetuate.

LIMITATIONS OF THE CURRENT DEBIASING APPROACHES

Current approaches to evaluating the role of AI in hiring may try to minimize discrimination by removing ↗ the most explicit references to race and gender when training models. However, this alone is unlikely to prevent discriminatory outcomes and could even lead to worse performance overall. Information about protected class membership can also be inferred from content that correlates with particular social identities. For example, names alone do not unambiguously signal a gender or racial identity, but they can be an implicit cue that signals identities that are more likely than others. Other implicit signals, such as locations and even word choice, can give ↗ information ↗ from which AI models can infer social identities. In 2018, Amazon

revealed that an AI recruiting tool it developed unfairly discriminated against graduates of all-women's colleges, suggesting that educational history can also be used to infer and discriminate against particular identities. Other empirical work has found that resumes mentioning awards or other honorary recognitions related to disability can lead to worse outcomes than those that include no awards at all.

Given the many ways social identities can be signaled in hiring materials, fully removing identifying information from training data or resumes under evaluation is infeasible. In some cases, it may even be inadvisable, as these features are often inseparable from achievements or activities that are directly relevant to hiring decisions. Therefore, further bias mitigation approaches are needed from both model developers and regulators tasked with ensuring fair hiring practices. Additionally, informing employers about how this nuanced information can signal identities and the potential consequences of using AI for hiring can both increase legal compliance and enable the creation of a more diverse workforce, which has been shown to improve productivity and employee performance.

ETHICAL AND EQUITABLE AI USE IN EMPLOYMENT

The findings from our research suggest that more work needs to be done, especially in empirical settings, which reveal the role and outcomes of increased AI use in hiring practices. Until more researchers, industry experts, and policymakers convene to determine strategies to improve the performance of these models through greater debiasing tools, we offer some programmatic and policy recommendations, which include greater auditing of these models, understanding the impact of intersectionality on model identification of suitable candidates, and greater transparency.

Greater auditing practices and regulation of AI hiring tools

Mandating regular auditing or reporting of models' performance can help protect against discriminatory outcomes and is a potential best practice for employers who choose to leverage AI products in hiring. Some policymakers have also considered the importance of using auditing to mitigate employment biases in AI models. Currently, New York City and Colorado are the only jurisdictions with a comprehensive law

mandating auditing of AI hiring systems, with Colorado's going into effect in 2026. New York City's law has been in effect since 2023, but it has weaknesses ↗ that have impacted its ability to meaningfully reduce discrimination in AI hiring.

One area of risk is automation bias—a phenomenon in which people perceive AI-generated decisions as objective and are more likely to trust them over conflicting judgments from non-automated sources. In the case of hiring decisions, if AI systems are discriminatory, adding human decision-makers into the process may not counteract the discrimination but instead further entrench it because humans may also be biased ↗ when making hiring decisions. The New York City law includes a disclosure exemption for firms that use AI systems alongside human decision-makers, leaving it up to the firms themselves to determine whether their systems qualify. In addition to the risk that this loophole could be exploited—undermining and significantly weakening the law—it also raises the possibility that some of the most severe and harmful discriminatory practices may go unreported.

It is necessary to adopt policies at various levels of government which encourage both model developers and employers to monitor their AI hiring systems for discriminatory outcomes and disclose the results to the public. While the federal government sets a minimum standard ↗ for employment discrimination, state and local laws may offer expanded protections for certain groups that should not be omitted from audit requirements. These regulations should include clear guidelines on how systems will be evaluated for compliance and how that compliance will be monitored and enforced. They should also not create exemptions for systems that work in collaboration with human decision-makers. Instead, jurisdictions should prioritize more support and infrastructure for independent monitoring and auditing of these AI products by increasing access to proprietary systems as well as the development of standardized evaluation materials and procedures.

Understanding the impact of intersectionality

Because intersectional identities can lead to greater disadvantages in hiring than single identities alone, it is crucial to raise awareness among hiring managers, policymakers, enforcers, and judicial officers about how these identities can be inferred and exploited by AI models. In September 2024, California became the first state to officially recognize ↗ intersectionality as a protected identity in addition to

single axes of identity. This means that Californians will not be required to prove that they have been discriminated against on the bases of only a single identity, which might be more difficult to show than discrimination based on a combination of identities or might not accurately reflect their lived experiences.

In the 1994 decision *Lam v. University of Hawaii*, the Ninth Circuit recognized that discrimination based on the combination of race and gender could not be reduced to discrimination based on either characteristic alone. Since then, however, other Ninth Circuit courts have applied Lam inconsistently—for example, in *White v. Wilson*, the court limited intersectionality to race and gender but applied the standard to a Black man rather than an Asian woman, as in *Lam*.

Another decision, *Bala v. Oregon Health & Sciences University*, acknowledged *Lam*'s mandate to examination combinations of identities but simultaneously separated the plaintiff's intersectional race-and-sex claim to be based on sex alone. By explicitly including the intersection of multiple identities in anti-discrimination legislation, as California has done, there are clearer standards for what qualifies as discrimination. As a result, plaintiffs may be more likely to file discrimination lawsuits based on the intersection of multiple identities, potentially leading to greater awareness and reforms around hiring discrimination.

In addition, explicitly considering intersectionality as a protected characteristic will encourage more research and testing of systems for harms against people with combinations of identities. Currently, our study is the only one that has investigated intersectionality in the context of AI and hiring, and it was limited to axes of gender and race. Other identities—such as sexuality, disability, or national origin—are also important and highly relevant in employment contexts. These should be considered in both anti-discrimination laws and in future monitoring and auditing of AI hiring systems.

Greater transparency in the use of AI hiring tools

An additional way to protect job seekers from discrimination—and to manage risk for employers—is to provide notice and obtain consent before using AI tools in the hiring process. This ensures that applicants are aware of the use of AI and can appeal

adverse decisions made by automated systems. Maryland, ↗ Illinois, ↗ Colorado ↗, and New York City ↗ require employers to obtain applicant consent before using AI to analyze application or interview materials, and Colorado also allows ↗ applicants to appeal adverse decisions made by AI systems.

For over 50 years, the Fair Credit and Reporting Act ↗ has required employers to notify applicants when background checks or credit reports are used and to disclose if that information led to an adverse employment decision. It also outlines procedures ↗ for applicants to dispute inaccurate information, including requiring reporting agencies to investigate and correct errors, and provides avenues for private litigation or government enforcement if disputes are not resolved. Similar mechanisms could empower job seekers to contest adverse impacts from AI systems that may inaccurately evaluate application materials. As the use of AI in hiring grows, policymakers could consider comparable processes to help applicants and employers better understand or appeal hiring decisions.

CONCLUSION

The increasing use of AI tools in the hiring process raises the risk of widespread employment discrimination if these systems are not properly developed, audited, and regulated. In our simulation of resume screening, we found that large language models (LLMs) caused significant gender and racial discrimination, particularly against Black men. Current technical efforts to mitigate biases are limited by an incomplete understanding of how protected characteristics like gender and race can be signaled on application materials or inferred by LLMs and employers.

To ensure the safety and legality of these systems, policy solutions are necessary, including broader support for independent audits of hiring systems; applying the same scrutiny to systems with human-AI collaboration as those using AI alone; considering how harmful effects are amplified for people with overlapping identities; and encouraging transparency when these systems make adverse decisions, empowering job applicants. Addressing the large-scale harms AI can inflict on people's economic and life opportunities must be a top priority for model developers, employers, and policymakers.

APPENDIX – RESEARCH STUDY METHODOLOGY

Parts of this section were pulled directly from the authors' report.

Data and models

To measure bias in resume screening, 554 resumes were augmented with a name—consisting of a variable first name and a constant last name—by prepending the complete name to the beginning of the document. Williams was selected as the last name because it is both frequent (the third most common name in the U.S.) and approximately equally likely to be used either by a Black or white person (47.68% vs. 45.75%). The last name was kept constant across all resumes to maximize experimental control and document realism while also minimizing required computation.

We use the name database introduced in Elder and Hayes (2023) to select names associated with one of four groups: Black males, Black females, white males, or white females. Among these, the Black male group contained the fewest potential names, so the 20 most distinctive names—representing 33% of all Black male names in the database—were chosen for resume augmentation. An equal number of names corresponding to the other groups were then selected to closely match or be proportional to the corpus frequencies of the Black male names. Corpus frequencies were determined using Infini-gram, a tool that facilitates n-gram searches for arbitrarily large corpora, and the DOLMA corpus.

The set of names was selected to reflect the relative population differences between Black and white people in the United States, replicating the distribution of names likely to appear in real-world resume screening. According to 2023 U.S. Census estimates, individuals who identify as white alone comprise 75.5% of the U.S. population, while those who identify as Black alone comprise 13.6%. Accordingly, we selected white male and female names that were approximately 5.5 times more frequent in the corpus than the corresponding Black male names, as well as Black female names that were approximately equally frequent to Black male names.

We also gathered a selection of 571 job descriptions ↗ across nine occupations: chief executive, marketing and sales manager, miscellaneous manager, human resources worker, accountant and auditor, miscellaneous engineer, secondary school teacher, designer, and miscellaneous sales and related worker).

These resumes were encoded by three massive text embedding models (MTEs)—E5-Mistral-7b-Instruct ↗, GritLM-7B ↗, and SFR-Embedding-Mistral ↗—along with 10 variations of instructions for the resume screening task. In total, we computed nearly 40,000 comparisons of resumes and job descriptions for each model, providing a sufficiently representative assessment of the potential impacts these models could have when deployed at scale.

Resume screening experiments

Zero-shot dense retrieval, which uses contextualized embeddings to compare documents rather than relying on exact term matches, provides a natural analog for resume screening. In the initial stages of retrieval, relevance scores computed from text embeddings are used to select a set of documents from a large corpus that best match a user request, with cosine similarity commonly used ↗ as the relevance metric. Similarly, in resume screening, resumes that are most similar to a job description can be identified via the cosine similarity of their respective embeddings. Furthermore, using a retrieval-based approach for resume screening enables direct analysis of textual embeddings to determine whether the representations are potentially biased in ways that could influence model outputs. If the resumes most similar to a particular job description consistently belong to a certain group, this provides evidence that the representations are biased in favor of that group.

To simulate candidate selection, we select a percentage of the most similar of resumes for each job description for further analysis. A chi-square test is used to determine whether the selected resumes are distributed uniformly among relevant groups or whether certain groups are represented at significantly higher rates than others, indicating bias in resume screening outcomes. Results for resume screening outcomes are presented primarily in terms of difference in selection rates.

Gender and race groups were formed by combining names—selected with population-proportional frequencies from the four intersectional groups—into four groups corresponding to a single race or gender identity (Black, white, male, or female). Each group contained 40 names. Embeddings for job descriptions and name-augmented resumes were generated using the three MTE models, and cosine similarities were computed.

For each model and occupation, we performed a bias test by selecting the top 10% of the most similar resumes for every job description and determining whether race or gender groups were represented at significantly higher rates. At this threshold, a minimum of 160 resumes were selected for each job description, and a total of 27 bias tests were conducted for both gender and race.

Using the 20 names with population-proportional frequencies from each intersectional group (Black female, Black male, white female, white male), we repeated the embedding procedures, selection of the top 10% of resumes, and 27 chi-square bias tests from the gender and race experiments for each pair of intersectional identities, excluding those in which no race or gender dimension was shared.

AUTHORS

Kyra Wilson **Ph.D. Student** - University of Washington

Aylin Caliskan **Nonresident Fellow** - Governance Studies, Center for Technology Innovation (CTI) 𝕏 @aylin_cim

---

Acknowledgements and disclosures

Amazon is a general, unrestricted donor to the Brookings Institution. The findings, interpretations, and conclusions posted in this piece are solely those of the authors and are

not influenced by any donations.

The Brookings Institution is committed to quality, independence, and impact.
We are supported by a diverse array of funders (/about-us/annual-report/) . In line with our values and policies (/about-us/research-independence-and-integrity-policies/) , each Brookings publication represents the sole views of its author(s).

Copyright 2025 The Brookings Institution