When Gender Becomes Visible: Benchmarking Fairness in LLM-Based Resume Screening

Yi Ding[1] *, Yiheng Lu[2] *, Xiaofei Yuan[2], Wei Liu[3], Ziyue Wu[1], Mingchen Ju[2], Ziyan Zhuang[1], Liuyi Chen[4], and Zhengyi Yang[1](✉)

[1] The University of New South Wales, Sydney, NSW, Australia
{yi.k.ding,ziyue.wu,ziyan.zhuang,zhengyi.yang}@unsw.edu.au
[2] Euler AI, Sydney, NSW, Australia
{yiheng.lu,xiaofei.yuan,mingchen.ju}@eulerai.au
[3] Wenzhou Kean University, Zhejiang, China
lwei@kean.edu
[4] Hunan University, Hunan, China
liuyi.chen@hnu.edu.cn

Abstract. Large language models (LLMs) are increasingly used for resume screening, but their sensitivity to gender-related cues remains unclear. We present a controlled benchmark for auditing gender bias in LLM-based job-resume matching, combining public job descriptions, de-identified real resumes, and counterfactual variants that vary only identity signals such as names, pronouns, and explicit gender attributes. Beyond male and female conditions, our benchmark includes non-binary and neopronoun-based identity presentations. We evaluate six LLM families: ChatGPT, Gemini, Claude, DeepSeek, MiMo, and Qwen, using a unified evidence-based scoring rubric. Across overall, seniority-level, and industry-level analyses, we find that model family, job level, and industry explain larger score variation than identity condition, while gender cues still cause small but recurring shifts, especially for neopronoun-based profiles. Our results show that fairness risks in LLM-assisted hiring are context-dependent and require aggregate, stratified, and matched auditing. The benchmark and experiments are available at https://github.com/SSSKino/Misgendered-2026-FAIML.

1 Introduction

Large language models (LLMs) are increasingly used in employment-related workflows such as resume screening, candidate ranking, job matching, and recruiter decision support. Their strong language understanding makes them attractive for evaluating free-form resumes against job descriptions (JDs), especially in scalable early-stage hiring. However, fairness becomes a central concern once LLMs influence hiring outcomes. Because screening decisions affect interviews, job offers, and career opportunities, even small systematic score differences may produce meaningful downstream inequities. This concern is reinforced by anti-discrimination laws and policies governing employment decisions [1,2].

Among these concerns, gender bias is particularly important. In practice, resumes may reveal gender through names, pronouns, self-descriptions, or explicit demographic statements. Thus, LLM-based screening systems may be exposed to gender-related cues even when gender is not explicitly requested. This raises a key question for AI-assisted recruitment: *do LLMs evaluate substantively identical candidates differently when only gender-related cues are changed?* The issue is further heightened by recent regulatory guidance clarifying that anti-discrimination obligations also apply to AI-assisted hiring systems [3]. It is especially underexplored beyond binary categories, including non-binary identities and neopronouns.

* Yi Ding and Yiheng Lu contributed equally to this work.

Modern resumes also vary in the visibility of gender information: some contain no cues, while others disclose gender through names, pronouns, or explicit identity statements. As such, model behavior may depend not only on candidate qualifications but also on how gender is presented. A rigorous fairness audit therefore requires controlled evaluation settings that isolate gender disclosure from confounding factors such as resume quality, wording and job type. To address this gap, this paper develops a controlled benchmark and experimental framework for studying gender bias in LLM-based resume screening.

Motivation. This research is motivated by the following factors:

High-stakes fairness in hiring. Hiring decisions directly shape access to interviews, employment, income, and long-term career development. When LLMs are used in early-stage screening, even small systematic differences in scores or rankings may lead to unequal downstream outcomes.

Multiple forms of gender disclosure. In real resumes, gender may be conveyed through names, pronouns, self-descriptions, or explicit demographic attributes. This makes it necessary to examine whether LLMs respond differently when qualifications are fixed but gender cues vary.

Need for controlled evaluation. Observed disparities can easily be confounded by differences in resume quality, writing style, job type, or seniority. A meaningful fairness study therefore requires realistic resumes and JDs together with carefully matched counterfactual variants.

Limited coverage in existing studies. Prior evaluations often treat gender as binary and pay limited attention to varying levels of gender visibility. The effects of non-binary identities, neopronouns, and progressive disclosure remain insufficiently studied.

Contribution. The main contributions of this paper are as follows:

A controlled benchmark for fairness auditing in LLM-based resume screening. We propose a controlled benchmark built from public job descriptions and de-identified real resumes. The benchmark is further extended into matched counterfactual variants that differ only in gender-related cues while preserving candidate qualifications, experience, and other substantive evidence, enabling systematic fairness evaluation under controlled conditions.

A comprehensive experimental study across models and settings. Using this benchmark, we conduct extensive experiments on multiple major LLM families under a unified scoring protocol. Our evaluation covers overall comparisons as well as stratified analyses across job levels, industries, and different degrees of gender-cue disclosure, providing a broad empirical view of model behavior in resume screening.

Empirical insights into context-dependent fairness risks. Our experiments reveal a layered pattern of fairness risk in LLM-based resume screening: most score variation is driven by model family, job level, and industry, whereas gender-related cues introduce smaller but non-negligible perturbations under controlled matched-resume conditions. These findings show that fairness risks are context dependent rather than uniform, motivating auditing protocols that combine aggregate evaluation with stratified, paired, and ranking-aware analyses.

2 Background and Related Work

LLMs in Resume Screening and Job–Resume Matching. Large language models (LLMs) are increasingly used for resume screening, candidate scoring, and job–resume matching. Compared with traditional keyword-based pipelines, LLMs can assess free-form resumes against unstructured job descriptions (JDs), making them attractive for early-stage hiring. However, recent studies show that LLM judgments remain sensitive to prompting, task design, and the information exposed in candidate profiles [4,5]. These findings suggest that, although LLMs are promising for hiring workflows, their reliability and fairness require careful evaluation.

Fairness and Bias in Algorithmic Hiring. Fairness concerns in hiring long predate LLMs, but they become especially important when LLM outputs influence interviews, rankings, or offers. Prior work shows that algorithmic hiring systems can reproduce or amplify existing disparities, and recent policy guidance has clarified that AI-assisted hiring remains subject to anti-discrimination law [6,7]. In the LLM setting, controlled studies have found measurable disparities in hiring-related evaluations and rankings across demographic conditions [8,9,5]. Together, these results show that fairness in LLM-based hiring cannot be assumed and must be tested under explicit evaluation protocols.

Gender Cues in Resumes. Gender information in resumes is not limited to an explicit gender field. It may be conveyed through names, pronouns, self-descriptions, and other contextual cues. Recent studies show that LLM-based hiring outcomes can change even when qualifications are fixed and only demographic signals are altered [10,11]. This is particularly important in resume screening, where even small score shifts may affect shortlist decisions. However, prior work rarely distinguishes different levels of gender disclosure in a controlled way, leaving open the question of whether model behavior depends on how gender information is presented.

Beyond Binary Gender. Most prior work focuses on binary gender categories, usually through male/female substitutions. This leaves limited evidence on how LLMs respond to non-binary identities or neopronoun(Neo) disclosures, despite their growing presence in real-world profiles [12]. Recent work has argued that broader gender inclusivity is necessary for meaningful fairness evaluation [13]. For hiring-related systems, this means benchmarks should go beyond binary settings and account for a wider range of gender expressions.

3 Benchmark and Experimental Setup

This section introduces the benchmark construction, the evaluated models and scoring protocol, and the empirical results under the resulting controlled evaluation setup. We construct a benchmark from public job descriptions (JDs) and de-identified real resumes to test whether LLM-based resume screening changes when gender-related identity cues vary while candidate qualifications remain fixed.

3.1 Benchmark Construction

Source Collection and Curation. Our benchmark is designed as a controlled testbed for auditing fairness in LLM-based resume screening under matched candidate conditions. It is constructed from two source types: public job descriptions (JDs) and real candidate resumes. On the JD side, we collect postings from three industries: *construction*, *IT*, and *nursing*. For each industry, we curate 20 JDs, resulting in a total of 60 JDs across the benchmark. These three domains are intentionally chosen to provide heterogeneous hiring contexts, since they differ substantially in qualification structure, terminology, credential conventions, and expected evidence of competence. Within each industry, the selected JDs cover both *junior-level* and *senior-level* roles, so that screening behavior can be analyzed not only across domains but also across different levels of responsibility and experience.

The JD data are collected from O*NET [14], which provides rich occupational information in a highly structured format. However, the original O*NET representation is not directly suitable for realistic resume-screening experiments, because much of the job information is encoded as occupational descriptors, attribute lists, and numerical scores that are difficult to interpret in a natural hiring context. In particular, the skill-related fields in O*NET are often associated with numerical values that are useful for structured analysis but are less intuitive when directly presented to a language model as a human-readable JD. To make the benchmark more realistic and easier to interpret, we preprocess the O*NET records before using them in evaluation.

JD Preprocessing and Text Construction. Our preprocessing pipeline converts structured O*NET records into JD-like textual descriptions that more closely resemble real-world hiring documents. We first extract the main occupational fields, including job title, occupation summary, relevant skills, and other task-related requirements. For skill-related fields, instead of keeping raw numerical values, we map the original scores into three coarse requirement levels: *high required*, *medium required*, and *low required*. This discretization serves two purposes. First, it makes the requirements easier to interpret, both for human readers and for downstream LLM evaluation. Second, it avoids over-emphasizing small numerical score differences from the source taxonomy that may not correspond to meaningful distinctions in practical hiring decisions.

On the resume side, we begin with real candidate resumes rather than fully synthetic profiles in order to preserve realistic writing patterns, evidence structure, and qualification presentation. For each industry, we collect 12 distinct candidate resumes, giving 36 base resumes in total across the three industries. Each resume is first passed through a de-identification process before entering the benchmark. We remove names, contact details, profile links, pronouns, and explicit gender markers, while preserving substantive qualification signals such as education, work experience, skills, projects, and certifications. This produces a gender-blind base resume that retains job-relevant evidence while minimizing direct identity leakage.

Matched Identity Variants. To isolate the effect of identity presentation from the effect of underlying candidate merit, we generate matched counterfactual variants from each de-identified base resume. Across these variants, all substantive evidence of qualification remains unchanged; only identity-related surface cues are modified. We consider four identity conditions: *male*, *female*, *non-binary*, and *neopronoun-based (Neo)*. The male and female variants use gender-associated names and conventional pronouns, the non-binary variants use *they/them* pronouns, and the Neo variants use the neopronoun set *ze/zir*. Since each industry contains 12 base resumes, this construction yields 48 resume variants per industry and 144 resume variants in total across the full benchmark. Under this design, any observed score difference across variants cannot be attributed to changes in education, experience, projects, or skills, since these qualification signals are fixed by construction. The matched-variant setup therefore enables a controlled fairness analysis in which the effect of identity-related information can be examined independently from candidate substance.

Benchmark Metadata and Pair Construction. The benchmark is organized along three main metadata axes: *industry*, *job level*, and *identity condition*. Concretely, the industry axis contains three categories (*construction*, *IT*, *nursing*); the job-level axis contains two categories (*junior*, *senior*); and the identity axis contains four categories (*male*, *female*, *non-binary*, *neo*). For each industry, 20 JDs are paired with 48 matched resume variants, yielding 960 JD–resume pairs per industry. Across all three industries, this results in a total of 2,880 JD–resume pairs in the benchmark.

4 Experiments Result

This section presents the empirical results under the benchmark and evaluation pipeline described above. We examine overall score behavior, then provide breakdowns by job level and industry. We evaluate six major LLM families: **ChatGPT-5.1**, **Gemini-3-flash**, **Claude-haiku-4.5**, **MiMo-V2-flash**, **DeepSeek-chat**, and **Qwen-3-max**. All models use the same JD-resume matching protocol, standardized input formatting, and the same structured scoring prompt.

Experiment 1: Overall Comparison provides an overall comparison on the full benchmark. Using the blind baseline condition, where explicit gender-related cues are removed, we compare average model scores across all matched resume groups. This experiment answers the most general

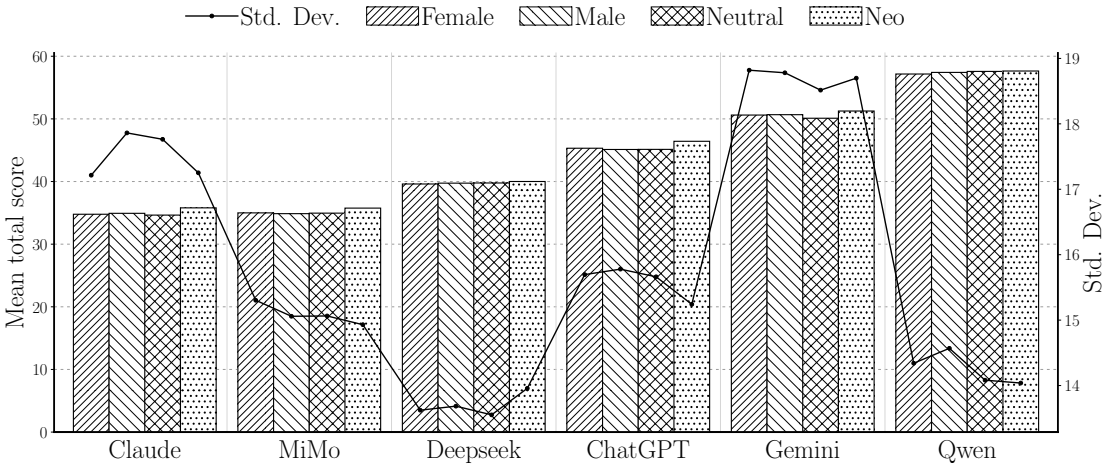

Fig. 1. Experiment 1: Overall Comparison

question: whether score differences remain even when candidate identity cues are suppressed. Two clear patterns emerge from Fig 1. First, *between-model differences are much larger than within-model differences across gender conditions*. Qwen consistently assigns the highest overall scores, followed by Gemini and ChatGPT, while MiMo and Claude are substantially stricter. This indicates that model family is a major source of variation in screening outcomes. Second, the score gaps across the four identity conditions are generally small in absolute magnitude, but they are not completely random. In all six models, the *Neo* condition is either the highest-scoring or tied for the highest-scoring group. The spread within each model is modest relative to the standard deviation, which suggests that the aggregate disparity is limited, but the repeated uplift for the Neo condition points to mild and systematic cue sensitivity rather than pure noise.

As a lightweight statistical check, we further compute paired score differences within each matched JD–resume block across identity variants. The paired comparisons confirm that identity-related gaps are small relative to between-model variation: most Female–Male, Non-binary–Male, and Neo–Male differences are within a narrow range around zero, while the Neo condition shows a weak but repeated positive shift across several models. Bootstrap confidence intervals are also much narrower than the observed cross-model score differences, supporting the conclusion that model family is the dominant source of aggregate variation.

Experiment 2: Job-Level Comparison examines whether score differences vary by *job level*. We divide the benchmark into *junior* and *senior* roles, and compare the screening results separately for these two groups. The goal is to determine whether potential disparity is more pronounced in higher-level or lower-level positions.

The dominant result in Fig 2 is that *senior roles receive higher scores than junior roles across every model and every identity group*. This gap is substantial and consistent. For example, Qwen increases from roughly 54-55 on junior jobs to about 60 on senior jobs, Gemini rises from roughly 47 to 54-55, and Claude rises from roughly 32-33 to 37-39. Thus, job seniority has a much stronger effect on scores than gender presentation.

Statistically, the paired comparison by job level leads to the same conclusion. Within both junior and senior roles, identity-related paired differences remain small compared with the score gap between job levels. The Neo condition again shows a mild positive shift in some senior-role comparisons, but this effect is secondary to the much larger seniority effect observed across all models.

Experiment 3: Industry Comparison examines whether model behavior differs across industries. We group the benchmark by *construction*, *IT*, and *nursing*, and compare the results within

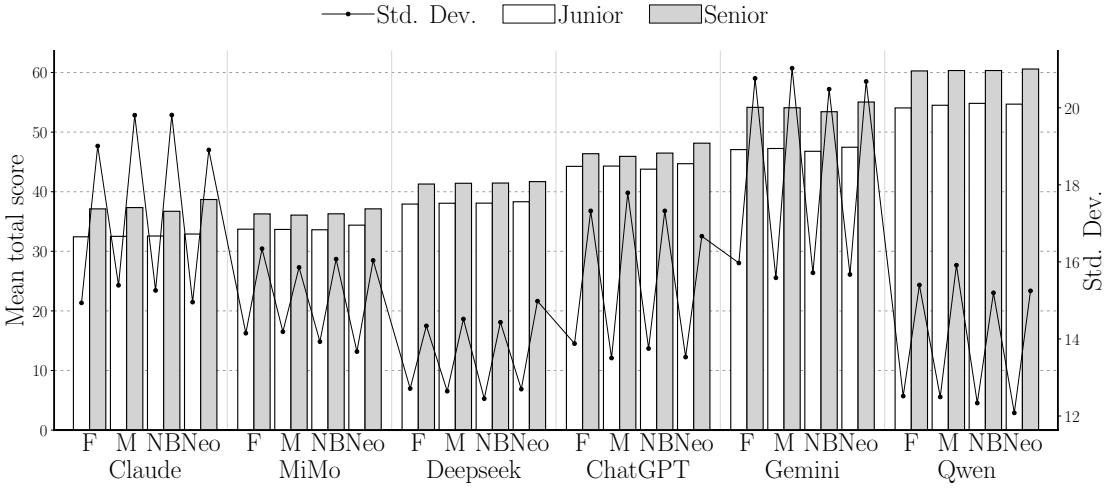

Fig. 2. Experiment 2: Job-Level Comparison

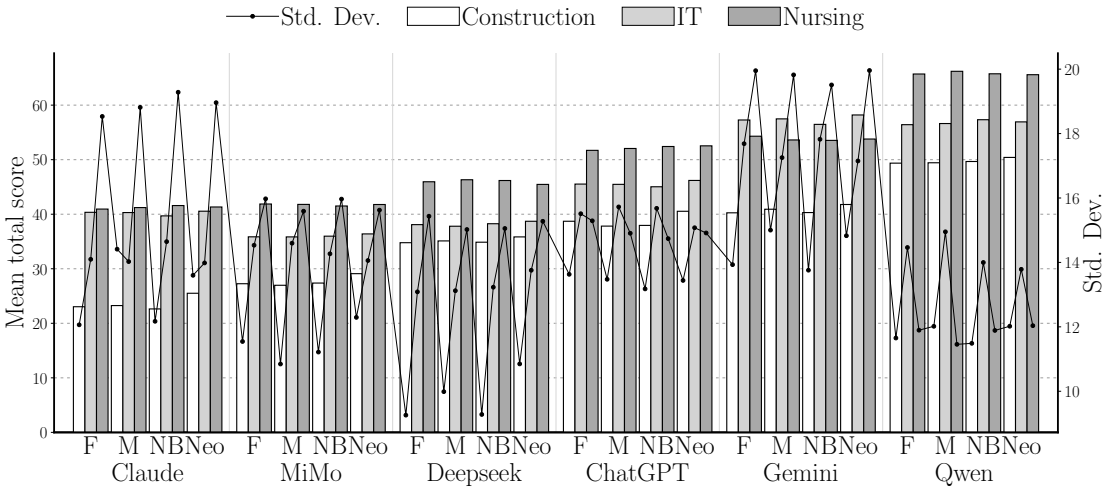

Fig. 3. Experiment 3: Industry Comparison

each industry. This allows us to test whether gender-related score variation is domain-dependent rather than uniform across hiring contexts.

These patterns indicate that model judgments are strongly shaped by domain context. The effect of industry is much larger than the effect of gender presentation within the same domain. At the same time, the slight Neo uplift remains visible in several industry slices, especially in construction and IT, which suggests that identity sensitivity persists but is moderated by domain. Overall, the results suggest that fairness evaluation should not rely only on pooled averages, because domain composition can dominate the observed score distribution.

We also repeat the paired comparison within each industry. The results show that industry effects are substantially larger than identity effects: construction, IT, and nursing produce clearly different score distributions even when identity variants are matched. Identity-related paired differences remain modest within each industry, although the slight Neo uplift is still visible in some construction and IT slices.

Discussion. Taken together, the three experiments suggest that variation in LLM-based resume screening is driven primarily by *model family*, *job level*, and *industry*, while the effect of gender-

related identity cues is generally smaller in magnitude but still not negligible. This conclusion is consistent across the overall comparison, the job-level comparison, and the industry comparison. At the same time, the experiments do not imply that gender-related cues are behaviorally irrelevant. Across all three experiments, score gaps across identity conditions are usually small in absolute terms, but they are not entirely random. In particular, the recurring slight advantage of the *Neo* condition appears in the overall comparison and remains visible in several job-level and industry-specific slices. Because the matched-resume design holds qualifications constant by construction, these repeated differences are more plausibly interpreted as mild cue sensitivity than as noise caused by substantive resume variation. The results therefore do not support a claim of large uniform disparity, but they do suggest that identity presentation can still perturb model outputs in a systematic way.

More broadly, the three experiments together show that fairness risks in LLM-based screening are *context dependent*. The magnitude and visibility of identity-related effects depend on which model is used, which roles are being evaluated, and which industries are represented in the benchmark. This means that fairness auditing should move beyond single aggregated summaries and instead include stratified analyses across models, job levels, and domains. Even modest identity-related score shifts may become consequential when candidates are close to shortlisting thresholds or when ranking decisions are sensitive to small margins.

5 Conclusion

This paper presents a controlled benchmark for auditing fairness in LLM-based resume screening under different forms of gender-related disclosure. By combining public job descriptions with de-identified real resumes and matched identity variants, the benchmark isolates the effect of identity cues while holding substantive candidate qualifications fixed. Empirically, the results support a measured conclusion: large and uniform aggregate disparities are not consistently observed across all models and settings, yet gender-related cues are not entirely neutral, as they can still induce small but systematic shifts in model outputs depending on the model, domain, and disclosure condition. These findings suggest that fairness risks in LLM-based hiring are better characterized as model-dependent and context-dependent sensitivities rather than as a single universal pattern.

A key future direction is to incorporate more fine-grained levels of gender-cue disclosure, enabling a more systematic analysis of how progressively stronger or weaker identity signals affect model judgments. Another important direction is to study prompt-side effects by introducing different levels of gender indication in the evaluation prompt itself, thereby examining whether explicit or implicit gender-related framing in instructions changes model sensitivity to candidate identity. Extending the benchmark along both the input axis and the prompt axis would make it possible to better understand how fairness risks emerge through the interaction between resume content, disclosure level, and prompt design, and would support the development of more robust and inclusive auditing protocols for LLM-based hiring systems.

References

1. U.S. Equal Employment Opportunity Commission, “Title VII of the Civil Rights Act of 1964,” https://www.eeoc.gov/laws/statutes/titlevii.cfm, 2025.
2. Federal Register of Legislation, “Sex Discrimination Act 1984 (Cth),” https://www.legislation.gov.au/C2004A02868/2024-10-14/2024-10-14/text/original/epub/OEBPS/document_1/document_1.html, 2024.

3. U.S. Equal Employment Opportunity Commission, "Selected Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," https://www.eeoc.gov/2023-annual-performance-report, 2023.
4. S. Vaishampayan, H. Leary, Y. B. Alebachew, L. Hickman, B. Stevenor, W. Beck, and C. Brown, "Human and LLM-based resume matching: An observational study," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4823–4838.
5. H. Iso, P. Pezeshkpour, N. Bhutani, and E. Hruschka, "Evaluating bias in LLMs for job-resume matching: Gender, race, and education," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, W. Chen, Y. Yang, M. Kachuee, and X.-Y. Fu, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 672–683.
6. A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, Jan. 2025.
7. U.S. Equal Employment Opportunity Commission, "Eeoc launches initiative on artificial intelligence and algorithmic fairness," https://www.eeoc.gov/newsroom/eeoc-launches-initiative-artificial-intelligence-and-algorithmic-fairness, Oct. 2021.
8. Z. Wang, Z. Wu, X. Guan, M. Thaler, A. Koshiyama, S. Lu, S. Beepath, E. Ertekin, and M. Perez-Ortiz, "JobFair: A framework for benchmarking gender hiring bias in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3227–3246.
9. P. Seshadri, H. Chen, S. Singh, and S. Goldfarb-Tarrant, "Small changes, large consequences: Analyzing the allocational fairness of LLMs in hiring contexts," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, and D. P. Singh, Eds. Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Dec. 2025, pp. 2645–2665.
10. H. An, C. Acquaye, C. Wang, Z. Li, and R. Rudinger, "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)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 386–397.
11. K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 1578–1590, Oct. 2024.
12. X. Tang, Y. Ding, Z. Yang, Y. Chen, Y. Gu, W. Yang, M. Ju, X. Cao, Y. Liu, and W. Zhang, "Do they understand them? an updated evaluation on nonbinary pronoun handling in large language models," in *AI 2025: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, M. Liu, X. Yu, C. Xu, and Y. Song, Eds. Springer, Singapore, 2026, vol. 16370, pp. 204–219. [Online]. Available: https://doi.org/10.1007/978-981-95-4969-6_16
13. Z. Shan, E. Diana, and J. Zhou, "Gender inclusivity fairness index (GIFI): A multilevel framework for evaluating gender diversity in large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 2548–2579. [Online]. Available: https://aclanthology.org/2025.acl-long.128/
14. U.S. Department of Labor, Employment and Training Administration, "O*NET 30.2 database," 2026. [Online]. Available: https://www.onetcenter.org/database.html