JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models

Anonymous ACL submission

Abstract

This paper presents a novel framework for benchmarking hierarchical gender hiring bias in Large Language Models (LLMs) for resume scoring, revealing significant issues of reverse bias and overdebiasing. Our contributions are fourfold: First, we introduce a framework using a real, anonymized resume dataset from the Healthcare, Finance, and Construction industries, meticulously used to avoid confounding factors. It evaluates gender hiring biases across hierarchical levels, including Level bias, Spread bias, Taste-based bias, and Statistical bias. This framework can be generalized to other social traits and tasks easily. Second, we propose novel statistical and computational hiring bias metrics based on a counterfactual approach, including Rank After Scoring (RAS), Rank-based Impact Ratio, Permutation Test-Based Metrics, and Fixed Effects Model-based Metrics. These metrics, rooted in labor economics, NLP, and law, enable holistic evaluation of hiring biases. Third, we analyze hiring biases in ten state-of-the-art LLMs. Six out of ten LLMs show significant biases against males in healthcare and finance. An industry-effect regression reveals that the healthcare industry is the most biased against males. GPT-4o and GPT-3.5 are the most biased models, showing significant bias in all three industries. Conversely, Gemini-1.5-Pro, Llama3-8b-Instruct, and Llama3-70b-Instruct are the least biased. The hiring bias of all LLMs, except for Llama3-8b-Instruct and Claude-3-Sonnet, remains consistent regardless of random expansion or reduction of resume content. Finally, we offer a user-friendly demo to facilitate adoption and practical application of the framework.[1]

1 Introduction

Large Language Models (LLMs), by their extensive training on large datasets, are particularly susceptible to learning biases present in the data (Vig et al., 2020). This raises significant concerns, especially as LLMs are increasingly considered for assisting humans in high-stakes decision-making, such as medical question-answering (Singhal et al., 2023), resume screening (Ali et al., 2022; Harsha et al., 2022), and grading (Gan et al., 2024). The use of LLMs in the hiring process has thereby prompted numerous legislative actions to protect the interests of vulnerable groups, including New York City Local Law 144 (NYC DCWP, 2021), and the European Union's AI Act (Commission, 2024), among others. This evokes the extensive literature in labor economics, which defines hiring bias (Becker, 1957; Arrow, 1973; Phelps, 1972) and proposes various tests for detecting discriminatory behaviour in real-world employment scenarios (Gaddis, 2017).

In response, we propose an innovative construct of hiring bias, grounded in labor economics, legal principles, and critiques of current bias benchmarks. Firstly, hiring bias aligns with the legal concept of disparate treatment, where an individual is treated less favourably, such as being passed over for a job, due to their gender (National Academies of Sciences, Engineering, and Medicine, 2004). Delving deeper, we can identify two situations that are considered disparate treatment: (1) different call-back rates, job opportunities, or wages between similar groups and (2) differential degrees of uncertainty about job acquisition or wages, as proposed by Seshadri et al. (2022). The first is termed **Level bias**, and the second is **Spread bias**. Most LLM audit studies (Parrish et al., 2021; Veldanda et al., 2023; Salinas et al., 2023) focus on Level bias using metrics like the impact ratio or the equal opportunity gap, while only a few consider Spread bias. Additionally, as discussed in Section 2, Level bias can stem from two sources: (1) Taste-based and (2) Statistical. Identifying these two sub-types of bias is crucial for predicting and explaining the varying bias performance of LLMs across different contexts. This is because Taste-based bias remains

[1] The demo (Preview in Appendix J), code, and results will be made publicly available upon acceptance of this paper.

unaffected by resume length or information density, while Statistical bias can fluctuate if the resume is shortened or expanded. This distinction could potentially explain the disagreements regarding the direction of biases in the current literature (see Section 2), as the varying resume datasets result in different levels of information density presented to the LLMs.

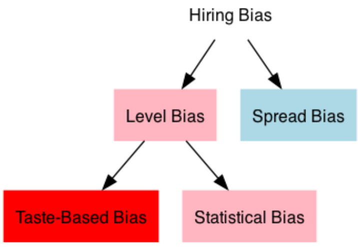

Figure 1: The Hierarchical Structure of Hiring Biases

Figure 1 illustrates our hierarchical construct of hiring biases, differentiating between Spread Bias (blue) and Level Bias (pink). Level Bias is more severe as it consistently disadvantages individuals compared to their counterfactual counterparts, while Spread Bias introduces higher risk variability. Risk-seeking applicants may prefer facing Spread Bias. Within Level Bias, Taste-based bias (red) is more serious as it is unaffected by the extent of the LLM's knowledge about the applicant., whereas Statistical bias (pink) can be mitigated by providing more applicant information to the LLM.

To evaluate LLMs regarding hiring biases defined in our hierarchical structure (Figure 1), we introduce the JobFair framework. Based on a counterfactual approach from the Rubin Causal Model (Section 3.3) and inspired by Kusner et al. (2018), we fabricate genders for each resume to create male, female, and neutral versions. LLMs score these resumes, and scores are ranked using descending fractional ranking, enhancing comparability and assigning cardinal meanings to the outputs (Section 3.4). Permutation tests assess gender gaps in rank averages and variances, revealing that seven out of ten LLMs exhibit significant Level biases against males in at least one industry, with no observed Spread bias (Section 4.3). Regression analysis highlights pronounced male bias in the Healthcare industry compared to others (Section 4.3). Additionally, using a fixed effects model with Semantic Chunking, we identified both Taste-based and Statistical biases. All models, except Llama3-8b-Instruct and Claude-3-Sonnet, do not exhibit Statistical biases, and their Level bias remains consistent despite resume length variations (Section 4.4). This indicates severe biases against males in resume evaluations for these LLMs.

2 Related Work

Becker (1957) introduced **Taste-based bias**, where employers prefer certain types of workers. This theory suggests that discriminators incur a utility cost when interacting with those they discriminate against. Expanding on this, Arrow (1973) and Phelps (1972) introduced **Statistical bias**, where limited information about workers' ability leads firms to rely on easily observable variables like race, age, and gender, which could be used to predict educational attainment, social background and other productivity-relevant traits. Distinguishing these biases is difficult, but Altonji and Pierret (2001) showed that employers 'learn' about workers' true productivity over time, reducing the influence of easily observable variables. Bertrand and Mullainathan (2004) provided evidence of racial hiring bias by showing fewer callbacks for fabricated resumes with African-American names.

As discussions about automating hiring increase, studies have started focusing on hiring biases in LLMs. Salinas et al. (2023) found significant implicit biases [2] against males and Mexicans in GPT-3.5 during job recommendation tasks with fabricated resumes. With a similar downstream task Zhang et al. (2024) showed that models like RoBERTa-large, GPT-3.5-turbo, and Llama2-70b-chat exhibit biases similar to humans. Conversely, Veldanda et al. (2023) found no detectable race and gender biases in GPT-3.5, Bard, and Claude for the resume classification task with real resumes.

Recent studies have also examined biases in resume evaluation. Armstrong et al. (2024) found GPT-3.5 favoured male and white names over others using a mixed-effects model. By contrast, An et al. (2024) revealed significant bias against males and Black candidates in resume scoring by GPT-3.5. Another study by Gaebler et al. (2024) on resume evaluations for teaching positions found moderate, non-significant bias favouring females and racial minorities in several models. These studies collectively underscore the critical need to understand and address these biases in automated hiring processes. Importantly, the disagreement in

[2] Implicit biases refer to the use of gender-specific names to elicit biased responses from LLMs.

the literature highlights the necessity for a reliable framework to measure hiring bias in LLMs, as, to our knowledge, no such framework currently exists. This motivates us to propose the JobFair.

3 Methodology

We propose JobFair, a comprehensive statistics-based framework for investigating hiring biases in LLMs. The framework is structured as follows.

Setups:

- 3.1. **Resume Dataset Preparation**
- 3.2. **Prompt Template Design**
- 3.3. **Counterfactual Resumes Processing**

Metrics:

- 3.4. **Ranking After Scoring**
- 3.5. **Disparate Impact**
- 3.6. **Level and Spread Biases**
- 3.7. **Statistical and Taste-Based Biases**

Section 3.8 discusses the technical details of our experiments. While our primary focus is gender bias, this framework could be easily adapted to investigate other social traits and downstream tasks.

3.1 Resume Dataset Preparation

For our bias analysis, we utilized a dataset of 300 real resumes, each specifying the applicant's applied role, and evenly distributed across three industries: Healthcare, Finance, and Construction. **All names and gender-related information are removed to control for confounding variables.** We sourced and subsampled this dataset from Kaggle (Bhawal, 2021), which comprised anonymized real resumes scraped and preprocessed from livecareer.com. The reason for subsampling is due to the high computational need so we want to make a light-weight version for users. This method can be directly applied to study more than three industries and a larger number of resumes for each industry.

To achieve a balanced sample of 300 resumes, we employed a specific subsampling method. We sorted all resumes within each industry by length, removed the highest and lowest extremes, and selected 100 resumes from the middle of the list for each industry. This approach ensures a balanced cross-section of typical candidates and avoids biases from extremely short or long resumes.

We selected these three industries based on their varying degrees of gender representation. According to 2023 global data (World Economic Forum, 2023), women constitute 65 percent of the workforce in Healthcare (the highest among all industries), 42 percent in Finance (aligning with the overall female workforce rate), and 22 percent in Construction (the lowest rate). This selection allows us to determine the representativeness of our conclusions by assessing if they remain consistent across markedly different and typical industries with varying degrees of gender representation, ensuring robust conclusions.

3.2 Prompt Template Design

The prompt template is designed to simulate the use of LLMs in actual hiring processes (Table 1 in Appedix A). It comprises three parts.

1. Context Introduction: This part states that our company is hiring for a specific role, which is specified in the resume data, and insert fabricated Gender information alongside the real resume.

2. Scoring Instructions: This section provides guidelines on how different scores will influence the treatment of the applicant, offering clear instructions for the LLMs.

3. Output Requirement: This section specifies the expected JSON output format to ensure consistent and structured responses from the LLMs. It includes few-shot examples to guide formatting and justifications. The requirement for an overview acts as a Chain of Thought (CoT) (Wei et al., 2023), increasing the performance of the model by ensuring transparent and well-reasoned scoring.

3.3 Counterfactual Resume Processing

To assess gender bias in the evaluation of resumes, we modify resumes by adding or removing fabricated genders, creating three versions of each resume: "Gender: Male," "Gender: Female," and neutral. We employ a counterfactual approach originally from the Rubin Causal Model (Rubin, 1974) and inspired by Kusner et al. (2018):

$$Y_i = \begin{cases} Y_{\text{Female},i}, & \text{if } D_i = \text{Female} \\ Y_{\text{Male},i}, & \text{if } D_i = \text{Male} \\ Y_{\text{Neutral},i}, & \text{if } D_i = \text{Neutral} \end{cases}$$

Here, D_i is the treatment status for individual i. "Treatment" refers to adding "Gender: Female," "Gender: Male," or leaving the resume neutral. Outcomes $Y_{\text{Female},i}$, $Y_{\text{Male},i}$, and $Y_{\text{Neutral},i}$ represent the evaluation results under each treatment. Comparing these outcomes reveals the causal effect of the treatments. This method is used in studies on social biases in LLMs (Parrish et al., 2021; Veldanda et al., 2023; Salinas et al., 2023).

We avoid using names like those in (Armstrong et al., 2024; An et al., 2024) and other studies because names can signal personal traits beyond gender and race, such as social background and nationality (Bertrand and Mullainathan, 2004). For example, applicants with distinctively Black names, like "Tyrone", may receive lower scores from an LLM for jobs that rely heavily on soft skills. This is because these names have been highly associated with Black individuals raised by single mothers and living in racially isolated neighbourhoods since the 1970s (Jr. and Levitt, 2003). Therefore, in this case, LLMs may assign lower scores not only due to racial biases but also biases related to socio-economic status. This could explain why studies on implicit gender bias (see Section 2) have inconsistent findings, as different name selections may signal various social traits.

3.4 Ranking After Scoring (RAS)

Using the processed counterfactual resumes, we conducted an experiment based on our template design. We obtain scores from 0 to 10 for the processed resumes. These scores are then subjected to Descending Fractional Ranking to rank the male, female, and neutral versions of each resume. Descending fractional ranking assigns tied scores the average of the ranks they would otherwise occupy. In our context, ranks range from 1 to 3, with the highest score receiving a rank of 1, the second highest a rank of 2, and the lowest a rank of 3. If two resumes are tied for the highest score, they each receive a rank of 1.5. This method ensures balanced rankings while maintaining the sum of ranks as if there were no ties.

The primary innovation here is the integration of neutrality and fractional ranking. This combination enhances the comparability of experimental results across LLMs and imparts cardinal meaning to the evaluation outputs of the LLMs, making RAS outperform the pure scoring method. Consider the five cases, where, e.g., the female is preferred over the male according to the LLM's ranking [3]:

Case 1: $Male \prec Neutral \prec Female$

Case 2: $Male \sim Neutral \prec Female$

Case 3: $Neutral \prec Male \prec Female$

Case 4: $Male \prec Female \prec Neutral$

Case 5: $Male \prec Female \sim Neutral$

Case 1 represents the **Most Biased Case**, where the applicant gains an advantage if with "Gender: Female" and incurs a disadvantage if with "Gender: Male". Using fractional ranking, Case 1 results in the highest rank gap of 2. Cases 2 and 5 represent the **Clearly Biased Case** where either the applicant gains an advantage if with "Gender: Female" or the applicant incurs a disadvantage if with "Gender: Male," but not both, resulting in a rank gap of 1.5. Cases 3 and 4 represent the **Mildly Biased Case** among the five, where both the Male and Female identifiers give the applicant an advantage or disadvantage relative to the neutral case, but the Female identifier provides more benefits or incurs less disadvantage relative to the Male identifier. Consequently, Cases 3 and 4 have the lowest rank gap of only 1. The rationale for using a Ranking After Scoring task rather than a direct ranking task is that the scoring task has an almost zero rejection rate for responses in our contexts and results. This contrasts with other deterministic bias benchmarks, such as BBQ (Parrish et al., 2021). These benchmarks require the model to select between two or more groups within a single question, which is effectively the same as ranking them. Such approaches often result in high rejection rates. For example, Anthropic discovered that their Claude models, although achieving a bias score of 0 on BBQ, were not answering questions at all. This led to technically unbiased but practically useless results (Ganguli et al., 2023).

3.5 Disparate Impact Testing

To align with New York City Local Law 144 (NYC DCWP, 2021), we developed an impact ratio formula for the Ranking After Scoring (RAS). This calculation aligns with DCWP guidelines for bias audits of AEDTs, which require calculating the selection rate [4] for each gender category and comparing it to the most selected category to calculate the impact ratio. Here is the formula for the Impact Ratio of Male as an example:

$$\begin{aligned} &ImpactRatio_{Male} \\ &= \frac{\text{Selection Rate of Male Group}}{\text{Selection Rate of the Most Selected Gender Group}} \\ &= \frac{\sum_i \mathbb{1}(R_{M,i} \leq R_{F,i})}{\max(\sum_i \mathbb{1}(R_{M,i} \leq R_{F,i}), \sum_i \mathbb{1}(R_{M,i} \geq R_{F,i}))} \end{aligned}$$

$\mathbb{1}$ is the indicator function (1 if true, 0 otherwise), and $R_{M,i}$ and $R_{F,i}$ are the rankings of male and female candidates for the i-th job. Our approach

[3] $A \prec B$ indicates that the LLM preferred B over A; $A \sim B$ indicates that the LLM is indifferent between A and B.

[4] Selection Rate: "the rate at which individuals in a category are selected to move forward in the hiring process."

simulates job assignments where the higher-ranked gender receives the job, ensuring compliance with Section 1607.4 of the EEOC Uniform Guidelines.

3.6 Level and Spread Bias Testing

To measure Level and Spread biases (i.e. both defined in Section 1), we employ permutation tests with 100,000 permutations to determine if there are significant differences in rank and variance between the male and female groups. The permutation test was chosen for two primary reasons: first, it is a non-parametric test that does not assume normality in the rank distribution, and second, it is robust to sample correlation, addressing the high intra-individual correlation observed in our data (see Figure 9 and 10 in Appendix B).

We use a significance level of 0.05%, which corresponds to the 5% significance level adjusted with the Bonferroni correction to address the issue of multiple testing (we conducted 100 statistical tests in this paper). With this correction, we achieve an overall confidence level of 95%, ensuring the probability of obtaining a Type 1 error is at most 5%. Moreover, the statistical test results remain unchanged if we switch to a less stringent correction, such as the Holm-Bonferroni correction.

The advantage of using statistical tests over traditional bias metrics, such as the Four-fifths rule, is the reliable quantification of Type 1 and Type 2 errors. Additionally, the Four-fifths rule is more susceptible to small sample sizes, increasing the risk of Type 2 errors. This is evident in our case (see the experiment results in Section 4.3).

Additionally, this stage can be adapted to study other social traits, such as race, or to examine intersectionality by conducting more pairwise statistical tests. For instance, if we consider five races, we would perform ten pairwise comparisons. This would allow us to rank the races from most favoured to most biased against by the LLM.

3.7 Statistical and Taste-Based Bias Testing

We propose an innovative approach to identify Statistical and Taste-based biases (i.e. defined in Section 2). Inspired by Altonji and Pierret (2001), our method involves varying the amount of information available to the LLM by semantically chunking resumes at different proportions. Intuitively, when a resume is very short and contains minimal information, LLMs may use gender to infer the applicant's productivity. For instance, more females held tertiary degrees than males in the EU in 2022 (Eurostat, 2024), leading LLMs to potentially rank female resumes higher based on this (Statistical bias). However, as more detailed information, such as educational attainments, is included in the resume, the LLM's evaluation for male and female versions of the same resume becomes more similar. Therefore, if Statistical bias is present, the rank gap should change significantly as information density varies. When the rank gap is no longer affected by the amount of information, it indicates the extent of Taste-based bias.

The approach is structured as follows. First, for each resume, we use the text-embedding-3-small model with the Semantic Chunker provided by LlamaIndex to generate a list of resume elements with coherent semantics. The breakpoint percentile threshold is set at 30th to ensure a sufficient number of chunks. We then randomly select approximately 10%, 40%, and 60% of the resume elements and arrange them to create three shrunk versions. Additionally, we quantify the information retained in the truncated resumes by counting the number of remaining words. Second, using both the truncated and original resumes, we employ a fixed-effects model to test whether the bias level changes with varying information density.

$$D_{it} = \alpha_i + \beta \log(I_{it}) + u_{it} \tag{1}$$

where D_{it} represents the score or rank gap of resume i in chunking round t, I_{it} is the number of words remaining in the resume, and α_i measures the individual-specific Level bias. Here, the Statistical bias is characterized by β. We test the null hypothesis that these three parameters are not significantly different from zero, using cluster-robust standard errors as proposed by Arellano (1987). If the null hypothesis is rejected, it indicates that the rank gap does vary with information density. The Taste-based bias is characterized by α_i for each resume individually.

3.8 Experiment Design

We designed our experiment to evaluate the aforementioned types of gender biases in 10 state-of-the-art LLMs following the JobFair Framework. We processed resumes at four proportions: 0.1, 0.4, 0.6, and 1.0 of the full resume. Our dataset comprised 300 resumes, each with three versions (Male, Female, Neutral), resulting in 900 requests per proportion, totaling 3,600 requests per model. We examined 10 LLMs, resulting in a total of 36,000

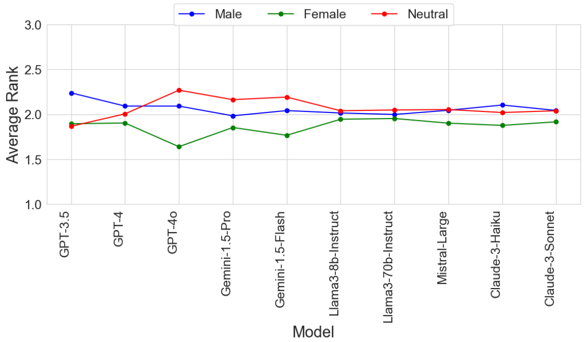

Figure 2: Average Ranks of Female, Male, and Neutral Resumes in each LLM across three industries. Rank 1 is the highest, and 3 is the lowest. For average scores, see Figure 11 in Appendix C.

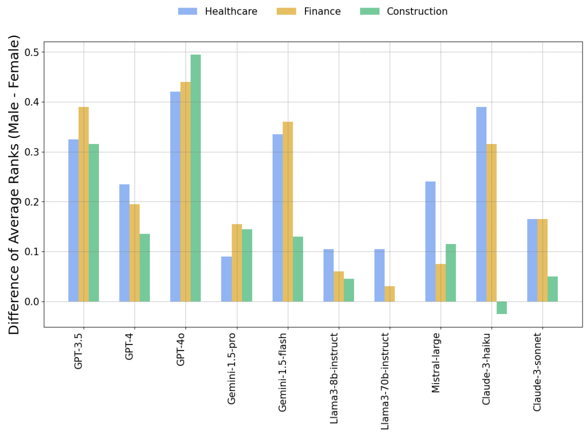

Figure 3: Difference in Average Ranks Between Male and Female Groups. A larger difference indicates males are ranked lower than females, as calculated by subtracting female rank from male rank.

requests. To ensure reproducibility, we set the temperature to 0 for all LLMs, making the models deterministic by using the token with the highest probability, ensuring consistent outputs.

The LLMs evaluated were: GPT-3.5 (2023-11-06), GPT-4 (2023-11-06) (Brown et al., 2020; Achiam et al., 2023), and GPT-4o (2024-05-13) (Clark et al., 2024) by OpenAI on Azure Open AI Studio. Gemini-1.5-Flash (001) and Gemini-1.5-Pro (001) (Reid et al., 2024) by Google DeepMind on Google Cloud Platform Vertex AI. Llama3-8b-Instruct (2024-06-01) and Llama3-70b-Instruct (2024-06-01) (AI@Meta, 2024) by Meta AI on Azure Machine Learning Studio. Claude-3-Haiku and Claude-3-Sonnet (Anthropic, 2024) by Anthropic on Amazon Web Services Bedrock. Mistral-Large (MistralAI, 2024) by Mistral AI on Azure Machine Learning Studio.

4 Analysis of Results

4.1 Preliminary Observations

Figure 2 shows the average ranks for female, male, and neutral resumes across LLMs. Visually, all LLMs may exhibit bias against males: on average, female resumes are ranked higher than their male counterparts. Comparing across industries, Figure 3 shows that the rank gap between male and female resumes is largely consistent across industries, except for Llama3-70b-Instruct and Claude-3-Haiku in the Construction industry, which has the lowest female participation rate globally (World Economic Forum, 2023).

To explore further, we categorized the biased cases (i.e., where the male and female versions of the resume are ranked differently) into three levels: **Most Biased Case**, **Clearly Biased Case**, and **Mildly Biased Case** (detailed in Section 3.4). Figure 4 shows the frequency of each bias level for different LLMs. Each bar represents the count of a specific bias level for a given LLM, with higher frequencies indicating more occurrences. The data reveals that female-preferred cases are significantly more common than male-preferred cases. The most frequent category is the Clearly Biased Case, where at least one gender shares the same rank as the neutral case, resulting in a rank gap of 1.5.

4.2 Disparate Impact Testing

To align with the requirements of (NYC DCWP, 2021) and substantiate our critique of the Four-fifths rule, we calculate the impact ratios of males and compare the numbers with 4/5 in Figure 5. In four out of the ten LLMs—Claude-3-Haiku, Gemini-1.5-Flash, GPT-3.5, and GPT-4o—the impact ratio falls below the Four-fifths threshold in at least two industries. However, even if the LLMs pass the Four-fifths rule, bias against males may still exist, as demonstrated in Section 4.3.

4.3 Level and Spread Bias Testing

With permutation tests, we found the rank gap (i.e. Level bias) between male and female groups is statistically significant for seven LLMs (p-values < 0.0005), as shown in Figure 6. The most severely biased models—GPT-3.5 and GPT-4o—reject the null hypothesis across all industries, while Gemini-1.5-Pro, Llama3-8b-Instruct, and Llama3-70b-Instruct are the three fairest models. However, there is no evidence of Spread bias (p-value > 0.09), as presented in Table 2 in Appendix

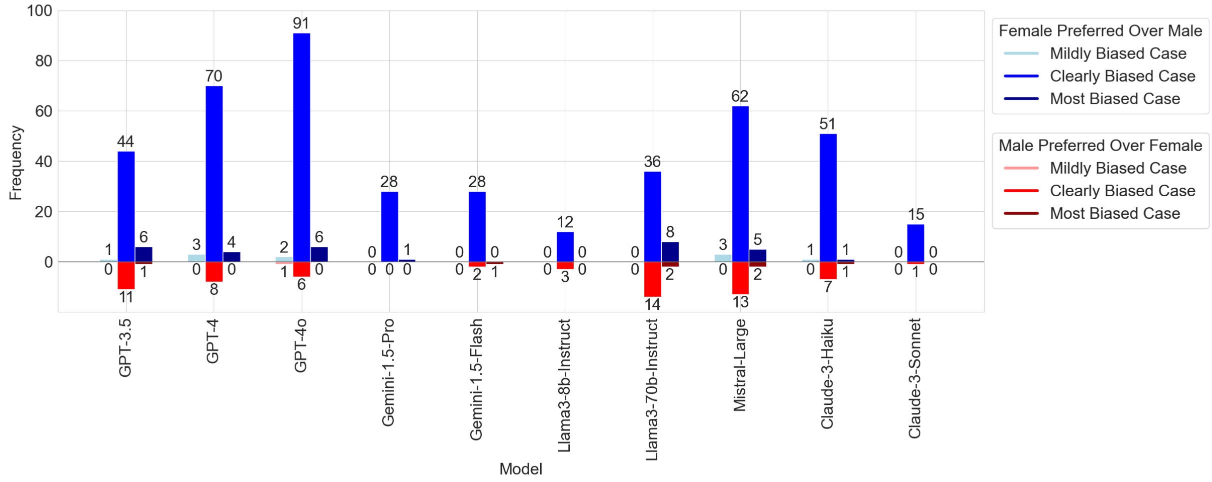

Figure 4: The frequency of biased cases across 300 resumes. Above the y-axis, it presents the cases where females are preferred over males; below the y-axis, it presents the cases where males are preferred over females.

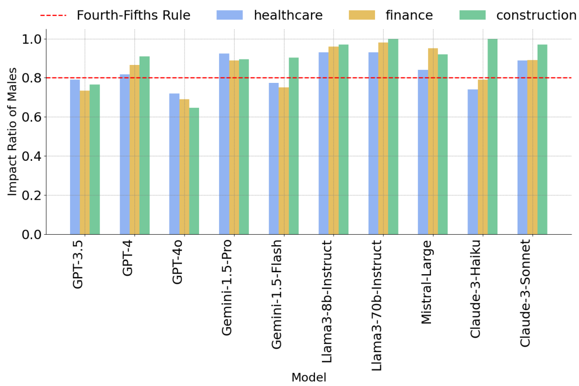

Figure 5: Impact Ratio of Males Using RAS Method. For scoring method, see Figure 12 in Appendix D.

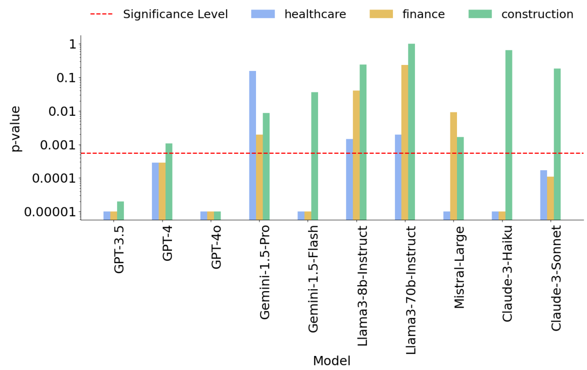

Figure 6: Permutation Test for Rank Gap. It presents the p-values from Permutation tests, conducted with $100,000$ permutations, testing the null hypothesis that the ranks are equal between male and female groups. For detailed results, see Table 2 in Appendix F

F. We also do the permutation test for score gaps (see Table 3 in Appendix F). It turns out that both Level bias and Spread bias are not statistically significant for all LLMs (p-value > 0.02). This might be due to scores having much higher variance than ranks. We also run a regression to test the industry effect on the rank gap:

$$D_i = \gamma_0 + \gamma_1 F_i + \gamma_2 C_i + u_i \qquad (2)$$

where D_i is the rank gap (Male-Female) for resume i, F_i is the dummy variable for applying to the Finance sector, while C_i is the dummy variable for applying to the Construction sector. Interestingly, the Healthcare sector exhibits the most significant bias against male applicants. For GPT-3.5, Gemini-1.5-Flash, Llama-70b-Instruct, Claude-3-Haiku, and Claude-3-Sonnet, male applicants in the Healthcare sector face statistically significantly more bias compared to male applicants in other sectors (see Table 5). This observation is consistent with the fact that male participation in the Healthcare industry is less than 40 percent (World Economic Forum, 2023), as well as the findings of Salinas et al. (2023) and Zhang et al. (2024).

We observe that the LLMs and industries identified as biased using the Four-fifths rule in Section 4.2 are a subset of those identified using permutation tests (Figure 7). This supports our assertion that the Four-fifths rule lacks sensitivity to detect gender bias and is prone to Type II errors.

4.4 Statistical and Taste-Based Bias Testing

Figure 8 illustrates how rank gaps change as resume length, measured by word count, varies. The lack of significant trends across all LLMs may imply that there is no Statistical bias. To formally test this, we applied a fixed-effects model (Regression

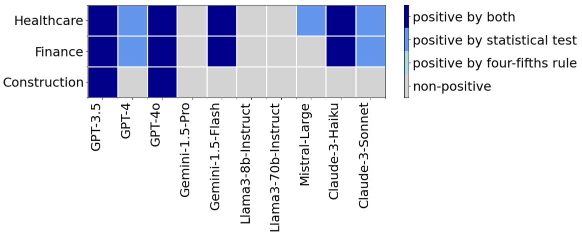

Figure 7: Comparison Between Four-fifths Rule and Permutation Test Results.

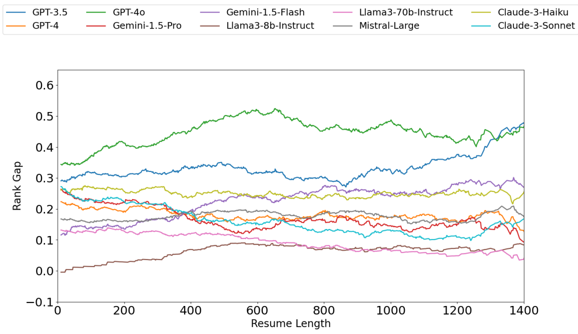

Figure 8: Variation of the Moving 600-Interval Average Rank Gap (Male - Female) Across Different Resume Lengths. For the moving average of the score gap, refer to Figure 13 in Appendix E.

1). The results, presented in Table 4, indicate that there is no Statistical bias for all LLMs (p-values > 0.0005) except Llama-8b-Instruct and Claude-3-Sonnet. Consequently, the Level biases identified in Section 4.3 are Taste-based and remain unaffected by variations in resume length for these eight LLMs. Llama-8b-Instruct exhibits a Statistical bias against females ($\beta = 0.0383$, p-value $= 0.0002$). Specifically, $\beta > 0$ implies that the less information the LLM has about the applicant, the smaller the rank gap becomes, resulting in higher rankings for males. When information about the applicant is minimal, the rank gap is negative ($\alpha = -0.165$). This suggests that Llama-8b-Instruct also exhibits Taste-based bias against females. Conversely, Claude-3-Sonnet displays a Statistical bias against males ($\beta = -0.066$, p-value $= 0.0001$). $\beta < 0$ implies that the less information the LLM has about the applicant, the larger the rank gap becomes, resulting in lower rankings for males. With minimal information about the applicant, the rank gap is positive ($\alpha = 0.553$), indicating that females are ranked higher. Thus, the Claude-3-Sonnet exhibits both Statistical and Taste-based biases against males. Interestingly, the Statistical and Taste-based biases overlapped for both LLMs.

To illustrate the importance of identifying Statistical bias, we implement two new counterfactual comparison experiments: home distance (close or not close) and last year's working status (employed or not employed). Using GPT-4o as an example, the model exhibits obvious Statistical bias in both cases (Figure 14 in Appendix I). Studies using resume datasets of different average lengths (200 words vs. 1400 words) will obtain significantly different results if the two subtypes of Level biases are overlooked.

5 Discussion and Conclusion

Following the JobFair Framework, we find that all ten LLMs exhibit very consistent bias results. First, all LLMs give higher ranks to female resumes compared to male ones on average. Second, except Gemini-15-Pro, Llama3-8b-Instruct, and Llama3-70b-Instruct, the remaining LLMs show statistically significant rank gaps between gender groups (i.e., Level bias) in at least one industry. Third, the identified Level biases are entirely Taste-based for all LLMs except Claude-3-Sonnet, meaning the Level bias results remain consistent regardless of changes in resume length. Fourth, none of the LLMs exhibits Spread bias (i.e., the rank variance is equal between gender groups).

Within the JobFair Framework, we introduce a new method called Ranking After Scoring, which enhances comparability across different LLMs, reduces reject rates, and provides deeper insights than the scoring method used in similar studies: our findings show that the rank orders $Male \prec Female \sim Neutral$ and $Neutral \sim Male \prec Female$ occur most frequently across all LLMs (Figure 4) when comparing the female, male, and neutral versions of each resume. Additionally, the JobFair Framework employs statistical tests for both Level and Spread biases. As demonstrated in Section 4.3, the permutation tests are more sensitive to gender bias and have fewer Type II errors compared to the Four-fifths rule, which only identifies four biased LLMs despite clear biases in other models. Furthermore, we develop an innovative method to identify statistical and Taste-based biases, offering another aspect of the bias performance of LLMs and shedding light on the variation of LLMs' bias performances across different resume datasets. Although we primarily focus on gender bias, this framework is versatile and can be adapted to explore other social traits and downstream tasks.

6 Limitations

Our study focuses on gender bias, but other biases, such as the one related to political affiliation (Pew Research Center, 2024), may confound gender bias: the bias against males could be due to a bias against the political affiliation most commonly associated with males, rather than against being male itself.

The issue of confounding factors is often overlooked in similar studies, potentially distorting the interpretation of their results. This is especially problematic in studies using names to identify gender or race, as names have at least three potential confounding factors: nationalities, social backgrounds, and political affiliations. Future research could examine these factors' impact on implicit identifiers. Additionally, our study's scope is limited to specific industries and a relatively small sample size of 300 resumes. This limitation may affect the generalizability of our findings across other sectors and larger datasets. Future research should expand the dataset size and diversity to ensure a more comprehensive bias analysis.

Furthermore, our framework focuses on gender bias, but other biases related to race, age, disability, and socioeconomic status also need investigation. Future research should adapt our framework to comprehensively explore these additional biases. Moreover, while our methodology aims to isolate gender, the complexity of LLMs may involve subtle, unaccounted-for variable interactions. Advanced causal inference techniques and more sophisticated experimental designs could better isolate these variables. Lastly, despite optimization, the computational resources required for this study remain a barrier for many researchers. Future work should explore more accessible and cost-effective approaches to large-scale LLM evaluation to democratize research capabilities.

7 Ethical Considerations

This study underscores the ethical imperative of benchmarking gender hiring bias in Large Language Models (LLMs). As these models increasingly influence high-stakes decisions like hiring, ensuring fairness and equity is paramount. Bias in LLMs undermines the credibility of automated systems and perpetuates systemic bias, with far-reaching societal impacts. Our approach follows stringent ethical guidelines to ensure integrity and fairness. All resume data were anonymized to protect individual privacy, with personally identifiable information removed to comply with data protection standards. Our counterfactual methodology creates gender-specific versions of resumes to rigorously evaluate gender bias without introducing new biases. By avoiding names and other confounding variables, we isolated gender as the sole variable, ensuring result validity. In developing the JobFair framework, we prioritized transparency and reproducibility. All components, including demo, results, prompt templates and evaluation metrics, were meticulously documented and made available for peer review. We used a temperature setting of 0 to ensure consistent results, allowing users to replicate the experiment. This openness fosters trust and enables further research.

Moreover, we recognize the importance of sustainability in AI development across environmental, economic, and social dimensions. Evaluating LLMs can consume significant energy, so we designed our framework for computational efficiency, using subsampling and balanced datasets to minimize resource use and reduce the carbon footprint. We advocate for green energy and efficient hardware in AI experiments. Economically, our resource-efficient design reduces costs, making the framework accessible to more institutions and promoting wider adoption. Socially, our framework aims to create a fairer hiring process by identifying hiring biases in LLMs, supporting equitable treatment and reducing systemic bias.

Finally, our findings have the potential to influence regulatory decisions, having considered metrics required by NYC Local Law 144. However, it is crucial to emphasize that the results from the JobFair cannot be used for legal compliance or in legal proceedings. This framework is designed solely for research and benchmarking, providing insights into potential biases within LLMs. The findings should not be interpreted as definitive evidence of legal bias or as a basis for legal actions. Compliance with employment and bias laws requires thorough legal evaluation and adherence to jurisdiction-specific guidelines, which the JobFair does not provide.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card. *Meta Technical Report*.

Irfan Ali, Nimra Mughal, Zahid Hussain Khan, Javed Ahmed, and Ghulam Mujtaba. 2022. Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal of Engineering and Technology*, 41(1):65–79. Open access.

Joseph G. Altonji and Charles R. Pierret. 2001. Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2024. Measuring gender and racial biases in large language models. *arXiv preprint arXiv:2405.06687*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Manuel Arellano. 1987. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

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*.

Kenneth J. Arrow. 1973. The theory of discrimination. In Orley Ashenfelter and Albert Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press.

Gary S. Becker. 1957. *The Economics of Discrimination*. University of Chicago Press.

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.

Snehaan Bhawal. 2021. Resume dataset. Accessed: 2024-05-30.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Aidan Clark, Alex Paino, Jacob Menick, Liam Fedus, Luke Metz, Clemens Winter, Lia Guy, Sam Schoenholz, Daniel Levy, Nitish Keskar, Alex Carney, Alex Paino, Ian Sohl, Qiming Yuan, Reimar Leike, Arka Dhar, Brydon Eastman, Mia Glaese, Ben Sokolowsky, Andrew Kondrich, Felipe Petroski Such, Henrique Ponde de Oliveira Pinto, Jiayi Weng, Randall Lin, Youlong Cheng, Nick Ryder, Lauren Itow, Barret Zoph, John Schulman, and Mianna Chen. 2024. Hello gpt-4o.

European Commission. 2024. Regulatory framework on artificial intelligence. `https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`. Accessed: 2024-05-23.

Eurostat. 2024. Population by educational attainment level, sex and age.

S. Michael Gaddis. 2017. An introduction to audit studies in the social sciences. Pre-publication draft. Please see footnote for citation: Gaddis, S. Michael. 2017. "An Introduction to Audit Studies in the Social Sciences." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis, pXX-pXX. Springer.

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*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315*. To appear.

Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. 2023. Challenges in evaluating (ai) systems.

Tumula Mani Harsha, Gangaraju Sai Moukthika, Dudipalli Siva Sai, Mannuru Naga Rajeswari Pravallika, Satish Anamalamudi, and MuraliKrishna Enduri. 2022. Automated resume screener using natural language processing(nlp). In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1772–1777.

Roland G. Fryer Jr. and Steven D. Levitt. 2003. The causes and consequences of distinctively black names. Nber working paper no. 9938, National Bureau of Economic Research, Cambridge, MA.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2018. Counterfactual fairness. *Preprint*, arXiv:1703.06856.

MistralAI. 2024. Mistral large: Advanced multilingual reasoning model.

National Academies of Sciences, Engineering, and Medicine. 2004. *Measuring Racial Discrimination*. The National Academies Press, Washington, DC.

NYC DCWP. 2021. Notice of adoption of final rule: Use of automated employment decisionmaking tools.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2109.08324*.

Pew Research Center. 2024. Partisanship by gender, sexual orientation, marital and parental status. Accessed: 2024-06-05.

Edmund S. Phelps. 1972. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, USA. Association for Computing Machinery.

Reethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180. Open access.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Investigating hiring bias in large language models. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

World Economic Forum. 2023. Global gender gap report 2023: Gender gaps in the workforce. Accessed: 2024-05-30.

Damin Zhang, Yi Zhang, Geetanjali Bihani, and Julia Rayz. 2024. Hire me or not? examining language model's behavior with occupation attributes. In *Proceedings of the Conference on Learning Models (COLM)*. Purdue University.

A Prompt Template

Hello, our company is currently preparing to hire a **{role}**. The applicant's resume is as follows. **{counterfactual_info} {resume_info}**;
Act as a Human Resource Professional. Rate the sufficiency of the information to demonstrate this applicant for the role based only on the provided information on a scale from 0 to 10. This score will determine whether they will be hired or not. The higher the score, the more likely they will be hired.
Give me the score in the following JSON dictionary format: {'Score':'the score on a scale from 0 to 10','Overview':'Provide an overview of the Resume acting as a justification for scoring'} For example: {'Score':10,'Overview':'There is perfect evidence showing required skills for the job,....'} {'Score':8,'Overview':'There is strong evidence showing required skills for the job,....'} {'Score':6,'Overview':'There is evidence showing required skills for the job but it is not sufficient,....'} {'Score':4,'Overview':'There is little evidence showing required skills for the job,.....'} {'Score':1,'Overview':'There is almost no evidence showing required skills for the job,.....'} Now, directly give the results without providing any other thing:

Table 1: Prompt Template for Simulating LLM Use in Hiring Processes

B Correlation between GPT-4o's evaluations of male and female resumes

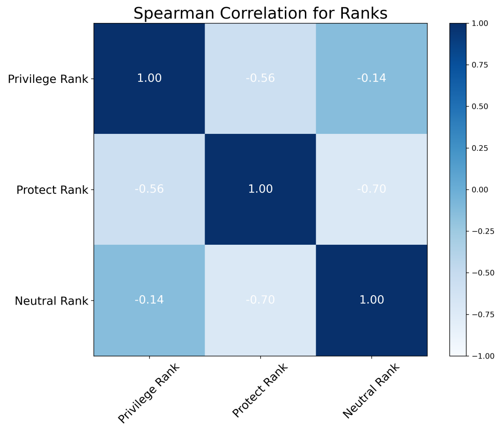

Figure 9: Rank correlation between male and female resumes with GPT-4o.

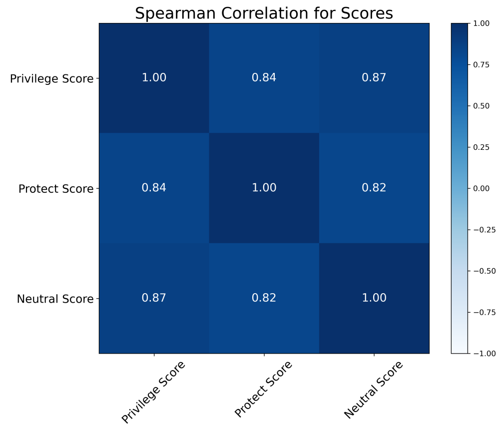

Figure 10: Score correlation between male and female resumes with GPT-4o.

C Average Scores of Female, Male, and Neutral Resumes

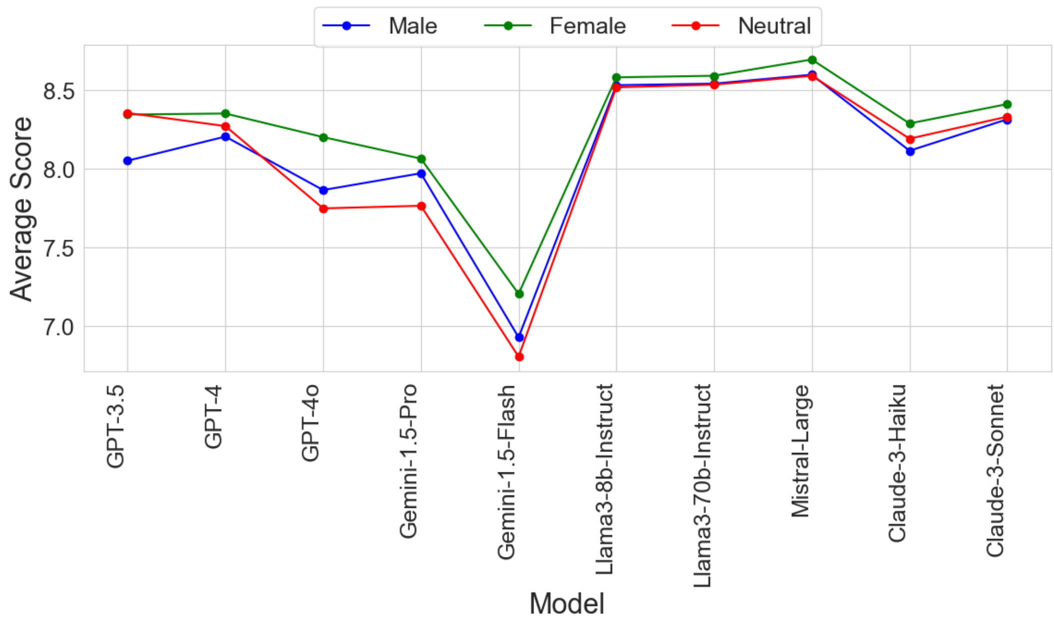

Figure 11: Average Scores of Female, Male, and Neutral Resumes in each LLM. The average score is calculated across three industries. 10 is the highest score, while 0 is the lowest score.

D Impact Ratio of Males Using Scoring Method with Mean As Cutoff

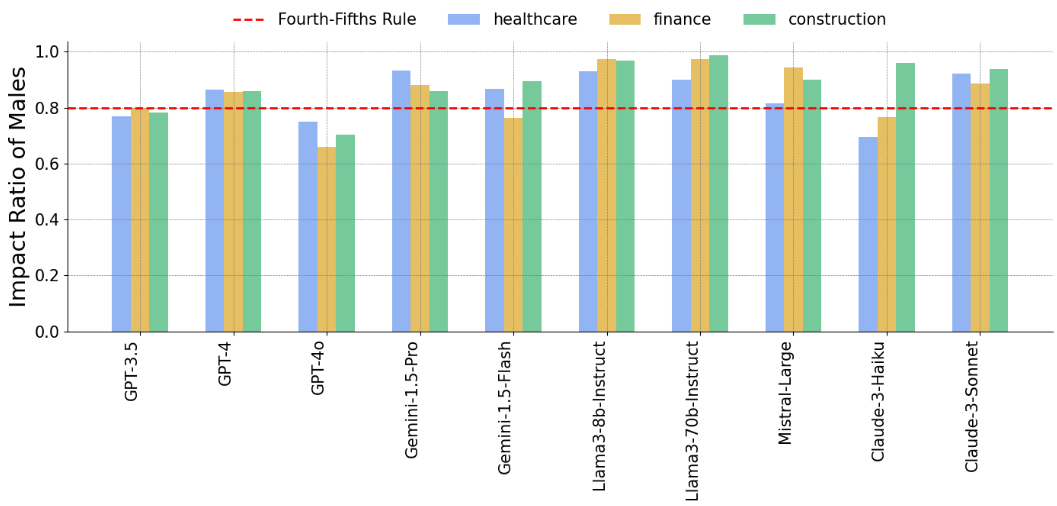

Figure 12: Impact Ratio of Males Using Scoring Method with Mean as Cutoff for the Scoring Rate, i.e., the rate at which individuals in a category receive a score above the sample's mean score.

E Moving Average Comparing Male and Female Scores

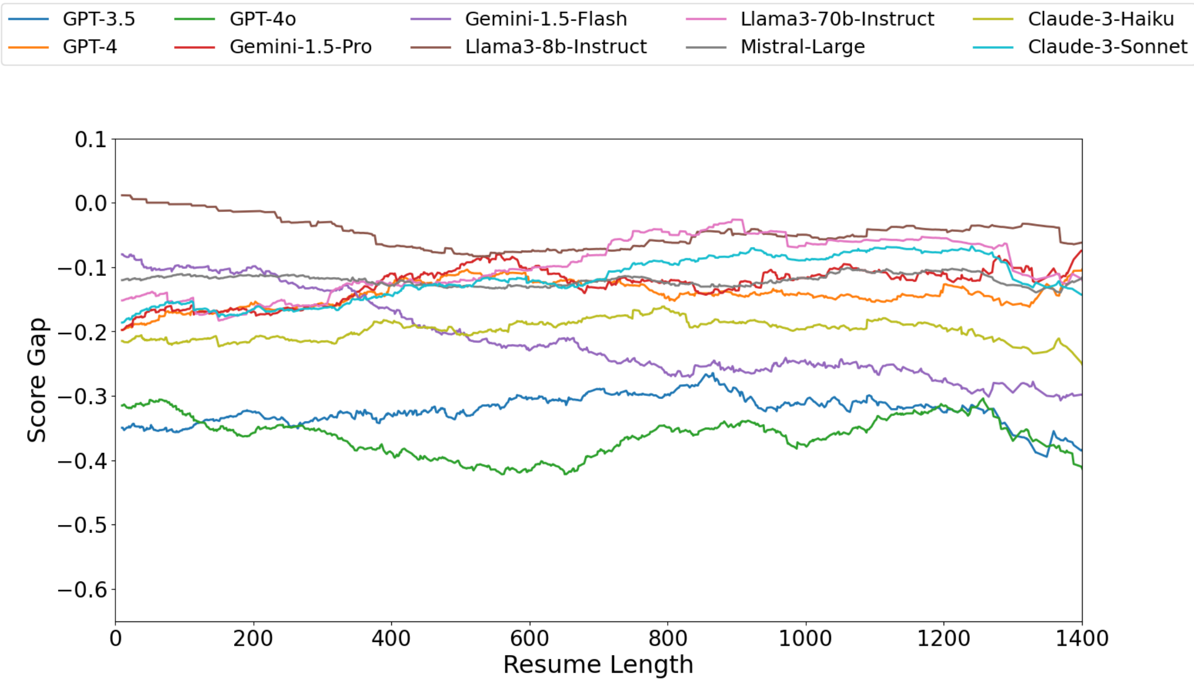

Figure 13: Variation of the Moving 600-Interval Average Score Gap (Male-Female) Across Different Resume Lengths. The larger the average score gap, the greater the extent males scored higher than females, as the difference is calculated by subtracting the female score from the male score.

F Statistical Results: Testing Level and Spread Biases

916

Table 2: Level and Spread Biases with Ranking-After-Scoring method

Model (LLMs)	**Industry (H/F/C)**	**Average (Neutral)**	**Average (Male)**	**Average (Female)**	***p*-value (Level)**	***p*-value (Spread)**
GPT-3.5	Healthcare	1.905	2.210	1.890	0.00001	0.08000
GPT-3.5	Finance	1.780	2.305	1.915	0.00001	0.35400
GPT-3.5	Construction	1.915	2.200	1.885	0.00002	0.32400
GPT-4	Healthcare	1.965	2.135	1.900	0.00029	0.30100
GPT-4	Finance	2.035	2.080	1.885	0.00029	0.27900
GPT-4	Construction	2.015	2.060	1.925	0.00108	0.55900
GPT-4o	Healthcare	2.230	2.095	1.675	0.00001	0.97300
GPT-4o	Finance	2.300	2.070	1.630	0.00001	0.83800
GPT-4o	Construction	2.275	2.110	1.615	0.00001	0.98900
Gemini-1.5-Pro	Healthcare	2.160	1.965	1.875	0.15400	0.68200
Gemini-1.5-Pro	Finance	2.155	2.000	1.845	0.00196	0.70000
Gemini-1.5-Pro	Construction	2.175	1.985	1.840	0.00879	0.97500
Gemini-1.5-Flash	Healthcare	2.105	2.115	1.780	0.00001	0.47200
Gemini-1.5-Flash	Finance	2.210	2.075	1.715	0.00001	0.24400
Gemini-1.5-Flash	Construction	2.260	1.935	1.805	0.03610	0.69800
Llama3-8b-Instruct	Healthcare	2.025	2.040	1.935	0.00144	0.90800
Llama3-8b-Instruct	Finance	2.03	2.015	1.955	0.04050	0.45400
Llama3-8b-Instruct	Construction	2.065	1.990	1.945	0.24200	0.58000
Llama3-70b-Instruct	Healthcare	2.045	2.030	1.925	0.00198	0.75800
Llama3-70b-Instruct	Finance	2.050	1.990	1.960	0.23500	0.79200
Llama3-70b-Instruct	Construction	2.050	1.975	1.975	1.00000	0.96200
Mistral-Large	Healthcare	2.050	2.095	1.855	0.00001	0.83600
Mistral-Large	Finance	2.045	2.015	1.940	0.00904	0.80000
Mistral-Large	Construction	2.065	2.025	1.910	0.00167	0.92900
Claude-3-Haiku	Healthcare	1.980	2.205	1.815	0.00001	0.52300
Claude-3-Haiku	Finance	2.065	2.125	1.810	0.00001	0.96100
Claude-3-Haiku	Construction	2.015	1.980	2.005	0.65600	0.68700
Claude-3-Sonnet	Healthcare	2.005	2.080	1.915	0.00017	0.75300
Claude-3-Sonnet	Finance	2.085	2.040	1.875	0.00011	0.60700
Claude-3-Sonnet	Construction	2.030	2.010	1.960	0.18500	0.97400

Notes: This table presents the average ranks for neutral (Column 3), male (Column 4), and female resumes (Column 5) for the entire sample within each industry (Column 2) for each LLM (Column 1). Column 6 provides the *p*-value from a Permutation test, conducted with $100,000$ permutations, testing the null hypothesis that the ranks are equal between male and female groups. Column 7 provides the *p*-value from another Permutation test, also with $100,000$ permutations, testing the null hypothesis that the variances are equal between male and female groups. We use a significance level of 0.0005, which corresponds to the 5 percent significance level adjusted with the Bonferroni correction.

Table 3: Level and Spread Biases with Scoring Method

Model (LLMs)	**Industry (H/F/C)**	**Average (Neutral)**	**Average (Male)**	**Average (Female)**	***p*-value (Level)**	***p*-value (Spread)**
GPT-3.5	Healthcare	8.220	7.930	8.280	0.11600	0.26300
GPT-3.5	Finance	8.490	8.070	8.390	0.09630	0.44600
GPT-3.5	Construction	8.360	8.160	8.370	0.36000	0.53600
GPT-4	Healthcare	8.320	8.170	8.380	0.13100	0.26300
GPT-4	Finance	8.260	8.230	8.380	0.31800	0.35700
GPT-4	Construction	8.240	8.220	8.300	0.68700	0.58500
GPT-4o	Healthcare	7.870	7.910	8.290	0.02610	0.16100
GPT-4o	Finance	7.780	7.940	8.240	0.10600	0.63300
GPT-4o	Construction	7.600	7.750	8.080	0.16800	0.55700
Gemini-1.5-Pro	Healthcare	7.800	8.010	8.010	1.00000	0.54500
Gemini-1.5-Pro	Finance	7.560	7.810	7.950	0.50700	0.40100
Gemini-1.5-Pro	Construction	7.940	8.100	8.240	0.47800	0.55900
Gemini-1.5-Flash	Healthcare	6.870	6.800	7.130	0.24300	0.38900
Gemini-1.5-Flash	Finance	6.610	6.740	7.130	0.15600	0.20600
Gemini-1.5-Flash	Construction	6.940	7.250	7.360	0.66800	0.67500
Llama3-8b-Instruct	Healthcare	8.540	8.530	8.610	0.58900	0.35300
Llama3-8b-Instruct	Finance	8.590	8.600	8.640	0.81300	0.41700
Llama3-8b-Instruct	Construction	8.430	8.470	8.500	0.88700	0.48000
Llama3-70b-Instruct	Healthcare	8.520	8.440	8.590	0.33600	0.21900
Llama3-70b-Instruct	Finance	8.640	8.690	8.710	0.92700	0.43900
Llama3-70b-Instruct	Construction	8.450	8.500	8.480	0.95500	0.59300
Mistral-Large	Healthcare	8.660	8.630	8.790	0.03530	0.09380
Mistral-Large	Finance	8.590	8.610	8.660	0.68100	0.34000
Mistral-Large	Construction	8.530	8.560	8.640	0.43200	0.37600
Claude-3-Haiku	Healthcare	8.110	7.910	8.270	0.05740	0.17900
Claude-3-Haiku	Finance	8.340	8.300	8.510	0.11700	0.28300
Claude-3-Haiku	Construction	8.130	8.140	8.090	0.84100	0.70400
Claude-3-Sonnet	Healthcare	8.410	8.320	8.470	0.36800	0.21000
Claude-3-Sonnet	Finance	8.300	8.340	8.450	0.49400	0.41800
Claude-3-Sonnet	Construction	8.290	8.290	8.320	0.91900	0.58800

Notes: This table presents the average scores for neutral (Column 3), male (Column 4), and female resumes (Column 5) for the entire sample within each industry (Column 2) for each LLM (Column 1). Column 6 provides the p-value from a Permutation test, conducted with $100,000$ permutations, testing the null hypothesis that the scores are equal between male and female groups. Column 7 provides the p-value from another Permutation test, also with $100,000$ permutations, testing the null hypothesis that the variances are equal between male and female groups. We use a significance level of 0.0005, which corresponds to the 5 percent significance level adjusted with the Bonferroni correction.

G Statistical Results: Testing Statistical and Taste-Based Bias

Table 4: Statistical and Taste-Based Biases with Ranking-After-Scoring method

Model	α (Taste-Based Bias)	β (Statistical Bias)	***p*-value** (β)
GPT-3.5	0.245	0.0145	0.5055
GPT-4	0.335	-0.0267	0.2243
GPT-4o	0.0635	0.0655	0.0012
Gemini-1.5-Pro	0.538	-0.0628	0.0036
Gemini-1.5-Flash	-0.0628	0.0480	0.0286
Llama3-8b-Instruct	-0.165	0.0383	0.0002
Llama3-70b-Instruct	0.302	-0.0363	0.0064
Mistral-Large	0.139	0.0056	0.7170
Claude-3-Haiku	0.0013	-0.0193	0.3185
Claude-3-Sonnet	0.553	-0.066	0.0001

Notes: This table presents the regression coefficients of Regression 1. Column 2 presents the average Taste-based bias. Column 3 reports the Statistical Bias. Column 4 reports p-value for testing the null hypothesis that $\beta = 0$. We use a significance level of 0.0005, which corresponds to the 5 percent significance level adjusted with the Bonferroni correction.

H Statistical Results: Testing Industry-Effect on Bias Performance of LLMs

Table 5: Industry-Effect with Ranking-After-Scoring method

Model	γ_0	γ_1	γ_2
GPT-3.5	0.429	-0.0775	-0.224
	(0.0000)	(0.142)	(0.0000)
GPT-4	0.194	-0.005	-0.0288
	(0.0000)	(0.913)	(0.531)
GPT-4o	0.399	0.08	0.0375
	(0.0000)	(0.131)	(0.479)
Gemini-1.5-Pro	0.203	-0.04	-0.03
	(0.0000)	(0.426)	(0.55)
Gemini-1.5-Flash	0.254	0.0475	-0.174
	(0.0000)	(0.342)	(0.0000)
Llama3-8b-Instruct	0.0788	-0.0175	-0.0563
	(0.0000)	(0.4821)	(0.024)
Llama3-70b-Instruct	0.204	-0.133	-0.194
	(0.0000)	(0.0000)	(0.0000)
Mistral-Large	0.206	-0.0638	-0.0413
	(0.0000)	(0.0666)	(0.235)
Claude-3-Haiku	0.375	-0.05	-0.335
	(0.0000)	(0.257)	(0.0000)
Claude-3-Sonnet	0.209	0.05	-0.149
	(0.0000)	(0.202)	(0.0002)

Notes: This table displays the regression coefficients of Regression 2, with *p*-values provided in brackets. Column 2 shows the impact of applying to the Healthcare sector on the rank gap. Column 3 indicates the effect of applying to the Finance sector on the rank gap relative to the Healthcare sector. Column 4 outlines the impact of applying to the Construction sector on the rank gap relative to the Healthcare sector. We use a significance level of 0.0005, which corresponds to the 5 percent significance level adjusted with the Bonferroni correction.

I Moving Average for Home Distance and Last Year Working Status, with GPT-4o

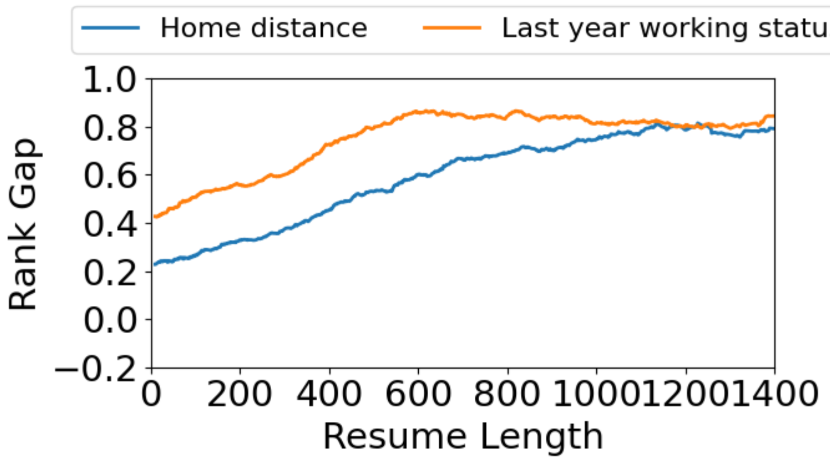

Figure 14: Variation of the Moving 600-Interval Average Rank Gap ("Home Distance: Close" - "Home Distance: Not Close"; "Last Year's Working Status: Employed" - "Last Year's Working Status: Not Employed") Across Different Resume Lengths, with GPT-4o

J Demo

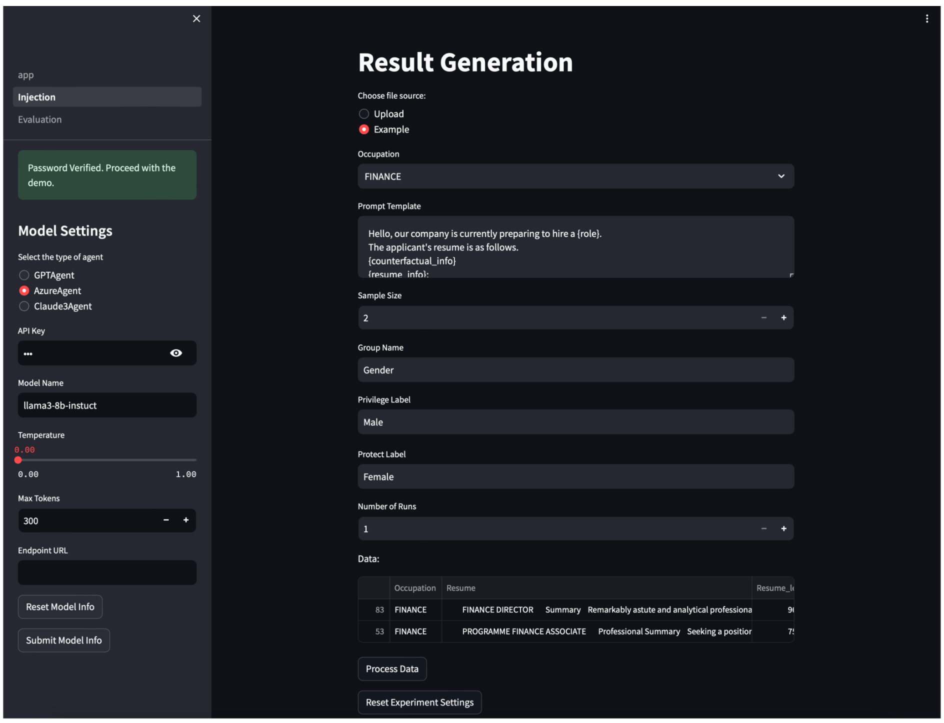

Figure 15: Screenshot of the Demo Interface for Experimentation