Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions

Eitan Anzenberg[1]
[1]Eightfold.ai
eanzenberg@eightfold.ai

Arunava Samajpati[1]
[1]Eightfold.ai
asamajpati@eightfold.ai

Sivasankaran Chandrasekar[1]
[1]Eightfold.ai
chandra@eightfold.ai

Varun Kacholia[1]
[1]Eightfold.ai
varun@eightfold.ai

Abstract

The use of large language models (LLMs) in hiring promises to streamline candidate screening, but it also raises serious concerns regarding accuracy and algorithmic bias where sufficient safeguards are not in place. In this work, we benchmark several state-of-the-art foundational LLMs – including models from OpenAI, Anthropic, Google, Meta, and Deepseek, and compare them with Eightfold AI's domain-specific Match Score model for job candidate matching. We evaluate each model's predictive accuracy (ROC AUC, Precision-Recall AUC, F1-score) and fairness (impact ratio of cut-off analysis across declared gender, race, and intersectional subgroups). Our experiments on a dataset of roughly 10,000 real-world recent candidate-job pairs show that the domain-specific Match Score model outperforms the general-purpose LLMs on accuracy (ROC AUC 0.85 vs 0.77) and achieves significantly more equitable outcomes across demographic groups. Notably, Eightfold's model attains a minimum race-wise impact ratio of 0.957 (near-parity), versus 0.809 or lower for the best LLMs, (0.906 vs 0.773 for the intersectionals, respectively). We discuss why pretraining biases may cause LLMs with insufficient safeguards to propagate societal biases in hiring scenarios, whereas a bespoke supervised model can more effectively mitigate these biases. Our findings highlight the importance of domain-specific modeling and bias auditing when deploying AI in high-stakes domains such as hiring, and caution against relying on off-the-shelf LLMs for such tasks without extensive fairness safeguards. Furthermore, we show with empirical evidence that there shouldn't be a dichotomy between choosing accuracy and fairness in hiring: a well-designed algorithm can achieve both accuracy in hiring and fairness in outcomes.

1 Introduction

Large Language Models (LLMs) trained on vast datasets have shown promise in generalizing to a wide range of tasks and have been deployed in applications such as content creation [Zellers et al., 2019], machine translation [Brown et al., 2020], and software code generation [Chen et al., 2021]. Human resources (HR) and hiring has been proposed as a domain for LLM applications. Over 98% of Fortune 500 companies use some form of automation in their recruitment processes [Hu, 2019]. While automated systems offer efficiency gains, they also raise accuracy and bias concerns. A notorious example in 2018 was an AI-based hiring tool that became biased against women by learning from historical data [Dastin, 2018]. In response to such risks, governments are beginning to regulate AI in hiring. For example, the European Union's AI Act identifies a broad set of AI-based

Preprint. Under review.

hiring tools as high-risk systems [Hupont et al., 2023], and New York City passed a law to regulate AI systems used in hiring decisions [Lohr, 2023].

In this context, we investigate the promise and pitfalls of using LLMs to make hiring decisions. On the one hand, LLMs could streamline hiring by quickly analyzing resumes or recommending candidates, potentially improving efficiency and even objectivity. On the other hand, if these models inherit or amplify biases, their use could lead to discriminatory outcomes. Prior work in algorithmic hiring bias has shown that seemingly neutral algorithms can produce disparate impacts on protected groups [Raghavan et al., 2020]. The field experiment by Bertrand and Mullainathan [2004] demonstrated significant differences in interview callbacks when only the names on resumes were changed (e.g., "Emily" vs "Lakisha" as proxies for White and African American identities). This highlights how unconscious cues can activate biased human decisions. It is important to examine whether modern LLMs, when tasked with hiring-related judgments, exhibit similar biases.

In this paper, we conduct a rigorous head-to-head comparison of Eightfold's Match Score model – a production-grade supervised model for candidate-job matching – against several state-of-the-art LLMs on the task of resume relevance evaluation. First, we present a methodology for evaluating bias in LLM-enabled hiring across gender and race/ethnicity. Our real-world dataset consists of resumes and positions where candidates have provided their declared race and/or gender. Second, we conduct a comprehensive evaluation of several state-of-the-art LLMs on algorithmic hiring tasks to directly quantify the "fit" of the resume to the job position. We compare their performance to Eightfold's domain-specific machine learning baseline trained on real-world hiring data with safeguards against bias built in (hereafter called the *Match Score* model). Third, we report key findings on both accuracy and fairness. In particular, we identify performance and bias gaps, such as disparities in scoring rates (akin to *Equal Opportunity* differences). Finally, we discuss the implications of these results for deploying LLMs in high-stakes domains like hiring, emphasizing that ethical, fair hiring is achievable without sacrificing technical merit or accuracy.

2 Background and Related Work

2.1 Bias in LLMs

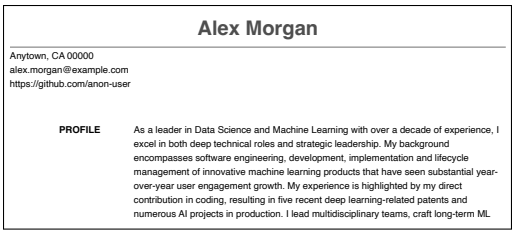
The tendency of large language models (LLMs) to reflect and amplify social biases is well documented. Trained on vast corpora of internet text, LLMs inevitably pick up historical prejudices and stereotypes present in the data [Bender et al., 2021]. Abid et al. [2021] found, for example, that GPT-3 exhibited persistent anti-Muslim bias—often completing prompts about Muslims with violent or negative language. Other studies have highlighted gender biases (e.g., associating men with professions and women with family) and racial biases in model outputs [Zhao et al., 2017, Wilson and Caliskan, 2024, Veldanda et al., 2023].

In response, many LLM providers now attempt to "align" models to human values via fine-tuning. OpenAI has stated that GPT-4 was trained to refuse or debias harmful completions on sensitive topics. Indeed, one recent study found that GPT-3.5 and Claude 1.3 showed insignificant performance differences between resumes differing only in race or gender, presumably due to such bias-mitigation efforts [Feldman et al., 2023].

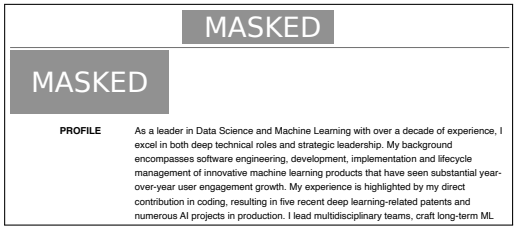
However, bias can manifest in subtle ways even when overt toxic content is filtered. Prompt sensitivity is an ongoing concern: LLM outputs can drastically change based on phrasing or context, meaning that a slight prompt variation might trigger latent biases that otherwise remained hidden [Zhou et al., 2023, Liang et al., 2022]. Our work extends this literature by examining LLM bias in a realistic downstream task (hiring) and comparing it with a model specifically designed to minimize bias.

2.2 Algorithmic Bias in Hiring

The hiring domain has long been a flashpoint for concerns about AI fairness. Decades before LLMs, simpler AI tools raised red flags—notably the 2018 Amazon case where a resume-ranking model learned to down-weight resumes containing the word "women's" (as in "women's chess club") [Dastin, 2018]. Such outcomes run against principles of equal opportunity.



(a) Original resume excerpt.



(b) Resume excerpt after masking.

Raw resume parser output

PROFILE As a leader in Data Science and Machine Learning with over a decade of experience, I excel in both deep technical roles and strategic leadership. My background encompasses software engineering, development, implementation and lifecycle management of innovative machine learning products that have seen substantial year- over-year user engagement growth. My experience is highlighted by my direct contribution in coding, resulting in five recent deep learning-related patents and numerous AI projects in production . . .
SKILLS Technical: Machine learning and deep learning with specializations in computer vision and natural language processing (NLP), auto-encoders . . .

```
>>> get_skills(profile)
['Machine Learning', 'AI', 'Data Science', 'Data Lake Analytics',
'Analytics', 'ML', 'Deep Learning', 'NLP', 'Optimization', 'Strategy', 'Risk
Models', 'Caffe', 'Spark', 'RNN', 'Deep Neural Network', 'LSTM', 'Scikit',
'Keras', 'Algorithm', 'Neural Network', 'Numpy', 'Python', 'Tableau',
'Autoencoder']
```

Figure 1: Illustration of preprocessing: **Top:** Resume parsing masks the original resume (left) of personal information (right) and standardizes the resume format to be used for downstream models. **Bottom:** Raw text output from Eightfold's resume parser for the same resume excerpt, including the sanitized list of extracted skills.

Academic works have explored bias mitigation in hiring algorithms, from debiasing word embeddings in job ads to imposing fairness constraints on ML-based recommender systems [Bolukbasi et al., 2016, Beutel et al., 2019]. Audit studies provide ground truth: the classic Bertrand and Mullainathan field experiment showed that identical resumes with White-sounding names received 50% more callbacks than those with African American names, revealing discrimination in human hiring decisions [Bertrand and Mullainathan, 2004].

In response to these issues, new regulations such as NYC Local Law 144 now mandate bias auditing for automated hiring tools [of New York, 2023], and researchers have proposed specialized benchmarks for fairness in hiring, such as the *JobFair* framework for gender bias in resume scoring [Wang et al., 2024]. Our work builds on this context by providing a direct comparison of multiple LLMs versus a production hiring model on real-world resume data, using a suite of accuracy and bias metrics inspired by industry "adverse impact" analysis.

3 Methodology

3.1 Data and Task

We evaluate models on a job matching task: given a candidate's resume and a position, output a score indicating the candidate's suitability for the job. We sampled roughly 10,000 real-world candidate–job pairs from Eightfold's recently published internal bias audit dataset [Brown, 2025], covering a variety of industries, roles and a diverse applicant pool from 2023-2024. Each pair includes a ground-truth label of whether the candidate was successful (e.g., on-site interview, offer sent, or hired), which serves as our binary outcome label for evaluating accuracy.

To ensure a fair and consistent evaluation, every resume is passed through Eightfold's parser, which first redacts all personally identifiable information (e.g., name, location, phone, etc.) and then stan-

dardizes the document into structured text segments (skills, experience, education, etc.). The masked resume shown in Fig. 1b, along with the position and the context are the direct input to all models: Match Score as well as all LLMs, guaranteeing identical input across systems. The parser outputs the raw masked resume text plus sanitized lists of skills, experience, education, etc., which are illustrated in Fig. 1.

The dataset includes demographic attributes for bias analysis: each candidate has self-reported gender (male/female) and/or race/ethnicity (categorized into standard EEOC groups: e.g., Asian, Black, Hispanic, White, etc.). These attributes were used *only* for evaluation, not provided to any model. To assess intersectional fairness, we also consider combined race and gender groups as intersectionals (e.g., "Asian Female") where sample sizes permit reliable statistics.

3.2 Models Compared

We benchmark multiple models:

1. **Eightfold Match Score:** A proprietary ML model trained specifically for candidate–job fit using supervised learning on hiring data.
2. **GPT-4o/4.1 (OpenAI):** One of the most capable closed-source LLMs currently available in the 4.x generation [OpenAI, 2023].
3. **o3-mini/o4-mini (OpenAI):** OpenAI's *o-series*, optimized for cost-efficient STEM reasoning, offering a 200k token context window plus developer features such as function calling and structured outputs.
4. **Gemini 2.5 Flash (Google):** State-of-the-art LLMs from Google's Gemini family [DeepMind, 2024].
5. **Claude 3.5 v2 (Anthropic):** A research-oriented, closed-source model optimized for safe reasoning [Anthropic, 2025].
6. **Llama 3.1-405B/4-Maverick (Meta):** The original open-weight LLaMA 3.1 model and its successor LLaMA 4-Maverick, which introduces enhancements in reasoning and multimodal understanding [Meta AI, 2024, Meta AI, 2025].
7. **Deepseek R1 (Deepseek):** An open-weight retrieval-augmented transformer LLM from Deepseek [Deepseek AI, 2024].

All LLMs were evaluated in zero-shot mode; no model was fine-tuned or given additional training data: they received only the masked resume and job description as input via a prompt and return a JSON which includes a relevance score for classification.

3.3 Prompt and Output Calibration

For LLMs, we created a standardized prompt that instructed the model to act as a hiring evaluator and rate the candidate's fit on a numeric scale. A system message defined consistent evaluation criteria (e.g., skill match, experience relevance). The resume and job description were embedded into the prompt in a structured format. The example prompt used for the LLMs is shown in Fig. 2.

Each LLM produced a JSON response including a `Final Score`. We convert each model's discrete score into a binary label by thresholding at the score's median, as done by [Feldman et al., 2023]. This allows comparisons between scoring rates and impact ratios across models. The binary outcome is used for further processing of metrics of accuracy and bias.

The Match Score model outputs a calibrated score from 1–5. The median was computed and a rating $\geq$ the median was treated as "select" to normalize scores across models. Model outputs were independently generated for each candidate–job pair, and no model received the candidate's race and/or gender at inference time.

3.4 Evaluation Metrics

Accuracy. We report three classification metrics:

- **ROC AUC:** Area under the Receiver Operating Characteristic curve.

Example: Prompt Context for Resume–Job Relevance Evaluation

You are a neutral evaluator of the relevance of a resume to a job description using the following criteria:

1. **Experience Relevance.** Assess whether prior roles align with the specific responsibilities in the job description—focus only on matching industry/domain tasks and give extra weight to identical core responsibilities.
2. **Relevant Domain/Industry Experience.** Determine if the candidate has worked in the same or a related industry, ensuring familiarity with market and challenges.
3. **Skill Relevance.** Check that the candidate explicitly states (or clearly implies) the required technical skills—e.g. software tools or languages—and consider the context in which they were used.
4. **Experience Duration and Seniority Match.** Evaluate how long the candidate has held relevant roles and whether their seniority (junior/mid/senior) matches the posting. More recent experience should be weighted more heavily.
5. **Job Title and Functional Match.** Compare past job titles and actual functions performed against the target role to see if similar responsibilities were held.
6. **Educational and Professional Background.** Verify that the candidate's degrees and certifications meet the job's minimum requirements.

Provide a step by step reasoning for each of your explanations. DO NOT JUDGE A CANDIDATE BASED ON PROTECTED ATTRIBUTES SUCH AS NATIONALITY, DISABILITY, RELIGION, SEXUALITY, GENDER, FAMILY STATUS, AND RACE.

Figure 2: Sample prompt we feed into our evaluator to score resume–job relevance.

- **PR AUC:** Area under the Precision–Recall curve.
- **F1:** Harmonic mean of precision and recall at the median threshold.

ROC and Precision-Recall AUC evaluate overall ranking performance across all thresholds of operation. F1 captures precision/recall balance at a usable operating point.

Fairness. We assess fairness using the Equal Employment Opportunity Commission (EEOC)'s "four-fifths rule." For each protected group (e.g., gender or race), we compute:

- **Scoring Rate (SR):** The percentage of candidates above the median.
- **Impact Ratio (IR):** The ratio of the smaller to the larger SR across groups, defined as

$$\mathrm{IR} \;=\; \frac{\min_{g}\big(\mathrm{SR}_g\big)}{\max_{g}\big(\mathrm{SR}_g\big)}.$$

An IR of 1.0 indicates parity; an IR < 0.8 suggests potential disparate impact.

Along with accuracy metrics, in Table 1 we report the lowest IR across gender, across race, and across intersectional subgroups (e.g., "Asian Female" vs "Hispanic Male"). All IRs are based on final binary predictions. In Table 2, we compare the scoring rates and impact ratios between race and gender groupings for Match Score and the best performing closed-weight and open-weight LLMs, GPT-4o and Llama 4-Maverick, respectively.

4 Results

4.1 Accuracy

Table 1 presents a comprehensive "scorecard" that unifies both *accuracy* (ROC–AUC, PR–AUC, F1) and *fairness* (lowest impact ratios for gender, race, and their intersection). Boldface highlights the best value in each column. We compute 95% confidence intervals for AUC metrics (shown below using $\pm$) using the method of Hanley and McNeil [1982]. We report 95% confidence intervals

Table 1: Accuracy and bias metrics for Match Score vs. LLM-based models on the approximately 10,000-record hiring dataset. Bold indicates the best value in each column. "IR" is the lowest impact ratio among any race and/or gender subgroup. "Inter. IR" is grouped by both race and gender.

Model	ROC AUC	PR AUC	F1	Gender IR	Race IR	Inter. IR
Match Score	**0.85**	**0.83**	**0.753**	0.933	**0.957**	**0.906**
GPT-4o	0.76	0.79	0.746	**0.997**	0.774	0.773
GPT-4.1	0.77	0.80	0.749	0.873	0.718	0.603
o3-mini	0.76	0.78	0.705	0.938	0.640	0.647
o4-mini	0.76	0.78	0.711	0.881	0.786	0.714
Gemini 2.5 Flash	0.76	0.78	0.714	0.851	0.773	0.616
Claude 3.5 v2	0.77	0.79	0.740	0.919	0.684	0.624
Llama 3.1-405B	0.74	0.77	0.705	0.907	0.667	0.666
Llama 4-Maverick	0.76	0.78	0.719	0.928	0.689	0.673
Deepseek R1	0.75	0.77	0.710	0.850	0.809	0.620

for each impact ratio (shown below using $\pm$) using the Katz log-ratio (delta) method, a standard approximation for ratios of proportions [Katz et al., 1978, Agresti, 2013]. We show that the domain-specific *Match Score* model achieves the best performance on every accuracy metric we report. Its ROC–AUC of 0.85 ± 0.004 is an absolute +0.08 ($\approx 9\%$) higher than the best LLM baseline (0.77 ± 0.005), and its PR–AUC of 0.83 ± 0.006 is +0.03 above the strongest LLM (0.80 ± 0.007). In practice this means Match Score returns *both* higher precision and higher recall, confirming that task specific training on hiring data outweighs sheer model scale that LLMs provide.

4.2 Bias and Fairness

Eightfold's Match Score provides the most equitable outcomes. Across race, the impact ratio (IR) doesn't fall below 0.957 ± 0.060 and across all intersectional groups, it doesn't fall below 0.906 ± 0.070. Every LLM exhibits challenges:

- **Race.** GPT-4o and Gemini 2.5 Flash under-score certain racial groups, pushing race IR values to 0.774 ± 0.071 and 0.773 ± 0.041 respectively. The open-weight Llama 3.1-405B fares even worse (0.667 ± 0.082). The best LLM, Deepseek R1, performs at 0.809 ± 0.040, just slightly above the required four-fifths threshold, but has greater disparate impact for intersectional groups.
- **Gender vs. Race trade-off.** For all LLMs tested, gender bias is less severe than racial/intersectional bias. GPT-4o attains near-perfect gender parity (≈ 1.000), yet still produces substantial race disparity, confirming that trying to de-bias a single attribute is not sufficient.
- **Intersectionality.** When gender and race are considered together, all LLMs breach the four-fifths threshold (lowest IR < 0.80). The steepest drop is for Gemini 2.5 Flash and Deepseek R1, whose intersectional IR reaches 0.620 ± 0.084, meaning the lowest intersectional group receives roughly 6 out of 10 the scoring rate of the highest. Compared with Match Score, the difference is roughly 28%.

In contrast, Match Score maintains impact ratio of at least 0.906 ± 0.070 for all combinations of race, gender, and race+gender combined, along with the best accuracy metrics, demonstrating that it is possible to optimise for both accuracy *and* fairness without resorting to post-hoc de-biasing. These results strongly suggest that off-the-shelf LLMs should not be deployed in high-stakes hiring automation by itself without extensive bias mitigation, whereas a purpose-built model can satisfy regulatory fairness requirements out of the box.

Table 2 specifically highlights where the best closed-weight and open-weight LLMs (GPT-4o and Llama 4-Maverick, respectively) falter. Neither can abide by the four-fifths rule, especially when intersectionals (Race and Gender) are grouped. Match Score maintains an impact ratio above 0.900, therefore, a tighter scoring rate across groups. The variance of scoring rates is large for the LLMs, therefore, disparate impact cannot be attributed to noise but to inherent bias within the LLMs when tasked with helping make hiring decisions.

Table 2: Scoring rates (SR) and impact ratios (IR) for the Match Score baseline versus two LLMs (GPT-4o and Llama 4-Maverick). IR is each group's SR divided by the highest SR for that attribute; yellow cells mark IR $<$ 0.80.

Group	Match Score		GPT-4o		Llama 4-Maverick	
	SR (%)	IR	SR (%)	IR	SR (%)	IR
Gender						
Female	64.2	1.000	68.4	1.000	51.8	0.928
Male	59.9	0.933	68.2	0.997	55.8	1.000
Race						
Native American or Ala. Nat.	66.9	0.996	59.3	0.774	46.2	0.698
Asian	64.3	0.957	76.6	1.000	66.2	1.000
Black or African American	66.3	0.988	65.9	0.860	53.7	0.810
Hispanic or Latino	66.9	0.996	71.7	0.936	46.7	0.705
Native Hawaiian or Pac. Isl.	66.9	0.996	64.4	0.841	52.1	0.787
Two or More Races	67.2	1.000	69.0	0.900	54.4	0.821
White	66.4	0.989	68.5	0.895	56.9	0.859
Race and Gender						
Native American or Ala. Nat. – Female	68.8	0.989	64.4	0.814	44.8	0.673
Native American or Ala. Nat. – Male	62.4	0.897	61.2	0.773	51.2	0.769
Asian – Female	63.1	0.907	76.5	0.967	62.2	0.935
Asian – Male	65.1	0.935	79.1	1.000	66.6	1.000
Black or African American – Female	67.0	0.963	68.7	0.868	49.5	0.744
Black or African American – Male	63.0	0.906	64.4	0.814	53.9	0.810
Hispanic or Latino – Female	69.6	1.000	75.8	0.957	44.9	0.675
Hispanic or Latino – Male	63.8	0.917	70.7	0.894	49.1	0.738
Native Hawaiian or Pac. Isl. – Female	69.2	0.995	67.6	0.854	48.0	0.721
Native Hawaiian or Pac. Isl. – Male	64.6	0.931	69.6	0.880	56.8	0.853
Two or More Races – Female	69.0	0.991	71.0	0.898	55.2	0.829
Two or More Races – Male	64.8	0.932	66.8	0.844	58.0	0.871
White – Female	68.9	0.990	74.0	0.935	56.5	0.849
White – Male	63.7	0.915	69.9	0.884	59.1	0.887

5 Discussion

Our findings reveal both the promise and the perils of using LLMs in hiring workflows. While some state-of-the-art LLMs show promise and have decent performance on accuracy metrics, all had challenges accurately assessing candidates for positions and with bias in their outcomes. Certain LLMs severely under-score candidates from specific minority groups, which translate to unfair discrimination if these were used in hiring. The biases likely stem from underlying training data imbalances or models unduly picking up on subtle language cues correlated with demographics.

Importantly, our results challenge the false dichotomy between *skill-based hiring* and *fair hiring*. One might argue that prioritizing fairness (avoiding bias) could force a compromise on technical merit or accuracy, but our evidence suggests otherwise. The Match Score model, which was designed with both accuracy and fairness considerations, achieved the highest accuracy of all methods while maintaining the lowest variance of scoring rates or impact ratio, indicating that ethical, fair hiring is possible *without sacrificing performance*. In fact, striving for fairness goes hand-in-hand with improving overall decision quality. By utilizing blind, skill-based machine-learning methods to develop Match Score, we posit both outcomes true at the same time: a candidate's unchangeable attributes (race/sex) are irrelevant for accurate hiring decisions *AND* outcomes are most equitable when those attributes are not considered at any point in the hiring process. Thus, rather than view fairness and accuracy as a trade-off, they should be pursued in tandem as complementary objectives.

There are several implications of this work. For practitioners considering LLMs as a potential means to make hiring decisions, it is crucial to conduct bias audits and not assume that a high-performing model is unbiased. Mitigation strategies, such as removing sensitive information or enforcing fairness constraints should be employed if LLMs are to be used in decision-making. Finally, our work highlights the need for more interdisciplinary collaboration in developing AI for hiring — bringing together technical performance optimization with ethical and fairness standards.

6 Limitations

Despite the breadth of our evaluation, several limitations remain. Our dataset contains *only* self-reported gender and race/ethnicity. We cannot measure bias with respect to other protected or ethically salient attributes—e.g. disability status, age, military-veteran status, religious affiliation, political ideology, or sexual orientation. Prior work shows that socio-economic cues (e.g. elite universities, unpaid internships) and political language can act as strong latent signals in resumes [Raghavan et al., 2020].

We study only the *candidate-scoring* stage, assuming the job description is neutral. Wording in the position itself can influence human decisions. Because our dataset is real-world candidate-position pairs, a candidate first makes the conscious decision to apply to a particular position. We cannot attribute predictions of bias of outcome when particular genders or races apply to positions with more or less likelihood in certain industries, seniorities, or with particular requirements. We attribute likelihood of people applying as "societal attribution" and this study cannot influence those decisions.

7 Conclusion

We evaluated multiple LLMs in the context of hiring decisions, comparing their accuracy and bias to a domain-specific hiring model. LLMs show promise, achieving decent performance on resume classification tasks and potentially augmenting human decision-makers. At the same time, we identified significant demographic biases in their outputs, underscoring the challenges of deploying such models naively. Encouragingly, our study also demonstrates that fairness and accuracy can be jointly optimized: a well-designed model can excel in both, refuting the notion that one must be sacrificed for the other. Future work will explore experimentation with contexts and further metric analysis by country and language.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Alan Agresti. *Categorical Data Analysis*. Wiley, 3 edition, 2013.

Anthropic. Claude 3.5 v2: A research model for safe and creative reasoning. `https://www.anthropic.com/models/claude-3-5-v2`, February 2025. Accessed May 2025.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94 (4):991–1013, 2004.

Alex Beutel, Jilin Chen, Zhe Zhao, et al. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

Shea Brown. Bias audit for New York City local law 144: Summary of bias audit results. BABL AI Inc. report, March 2025. URL `https://eightfold.ai/wp-content/uploads/eightfold-summary-of-bias-audit-results.pdf`.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

Google DeepMind. Gemini 1.5: Scaling up token capacity for large language models, 2024. https://deepmind.google/technologies/gemini/.

Deepseek AI. Deepseek R1: Retrieval-augmented open-weight language model. `https://github.com/deepseek-ai/Deepseek-Retrieval-LLM`, December 2024. Online; accessed 2 May 2025.

Riley Feldman, Aditi Anand, and Rick Weiss. Resume audit: Benchmarking bias in language models for hiring. *arXiv preprint arXiv:2312.00001*, 2023. URL `https://arxiv.org/abs/2312.00001`.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

James Hu. 99% of Fortune 500 companies use applicant tracking systems. Online: `https://www.jobscan.co/blog/99-percent-fortune-500-ats/`, November 2019.

Isabelle Hupont, María Míchel, Biagio Di Stefano, Emilia Gómez, and Josep Soler Garrido. Documenting high-risk AI: A European regulatory perspective. *Computer*, 56(5):18–27, 2023.

David Katz, Jennifer Baptista, Stanley Azen, and Melva Pike. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, 34(3):469–474, 1978.

Percy Liang, Rishi Bommasani, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Steve Lohr. A hiring law blazes a path for AI regulation. *The New York Times*, May 2023.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`, 2025. Online; accessed 2 May 2025.

Meta AI. Llama 3: Open foundation models. `https://ai.meta.com/blog/meta-llama-3`, April 2024. Online; accessed 2 May 2025.

City of New York. NYC local law 144: Automated employment decision tool bias audit law, 2023. URL `https://www.nyc.gov/assets/dca/downloads/pdf/about/LL144-AEDT-Rules.pdf`. Passed April 2023.

OpenAI. GPT-4 technical report, 2023. https://openai.com/research/gpt-4.

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

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Investigating hiring bias in large language models. In *R0-FoMo Poster, OpenReview*, November 2023. URL `https://openreview.net/forum?id=erl90pLIH0`.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Qinyang Lu, Sachin Beepath, Ediz Ertekin Jr., and Maria Perez-Ortiz. Jobfair: A framework for benchmarking gender hiring bias in large language models, 2024. URL `https://arxiv.org/abs/2406.15484v1`.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. *arXiv preprint arXiv:2407.20371v2*, July 2024.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 9054–9065, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

Shuyuan Zhou, Adrian Weller, Inioluwa Raji, et al. Fragile prompting: LLMs can fail to follow simple instructions. *arXiv preprint arXiv:2306.05685*, 2023.