

Bias in AI-Driven Resume Screening: A Mechanistic Interpretability Approach

Aliane Inès, Abidal Mauro

March 2025

1 Abstract

Automated resume screening systems are increasingly used in hiring, but concerns persist about biases in these models, particularly regarding gender. Our study examines the presence of gender bias in a language model's resume acceptance decisions using mechanistic interpretability techniques. We designed an experiment where resumes with identical qualifications varied only in names indicative of male or female gender. Using Goodfire's Ember API, we analyzed model logits for "Yes" and "No" decisions based on skill matching. Results indicate a statistically significant bias: male-associated names were more likely to receive affirmative responses in ambiguous cases. This work highlights the importance of interpretability tools in AI fairness and suggests avenues for mitigating bias in automated hiring.

Keywords: AI fairness, gender bias, interpretability tools, resume screening, machine learning bias.

2 Introduction

The increasing reliance on AI-driven applicant tracking systems has raised concerns about bias propagation. Prior research has shown that machine learning models can inherit biases from historical hiring data [1]. If unchecked, these biases may reinforce discrimination, particularly against underrepresented groups. Our project investigates whether a state-of-the-art AI model exhibits gender bias when evaluating resumes with equal qualifications but differing in names associated with male or female identities.

3 Methodology

3.1 Data and Experimental Setup

We generated synthetic resumes where only the first names varied. Names were selected from commonly used male and female names (e.g., "James" vs. "Mary"). Each resume included a predefined set of skills drawn from a standard technical job posting. The number of skill matches between a candidate and the job description ranged from 0 to 3.

3.2 Model and Tools

We used the Goodfire API with the Meta-Llama-3.1-8B-Instruct model to generate responses. The model was prompted with:

System Prompt:

You are an Application Tracking assistant. Look for the following skills in the CV {skills}. Answer 'Yes' if the CV matches the required skills, otherwise answer 'No'.

User Prompt:

Hello, my name is {name}. My skills are: {skills}. I want to join your team.

For each resume, we extracted logits for "Yes" and "No" responses and computed their difference ($D = L_{yes} - L_{no}$).

4 Results

4.1 Logit Analysis

Figure ?? shows the difference in logits (D) based on gender and skill matching. Results reveal that for resumes with one matching skill, male-associated names had higher median D values than female-associated names, indicating a greater likelihood of acceptance.

4.2 Statistical Significance

A t-test comparing D values for male and female names showed a statistically significant difference ($p < 0.01$). An ANOVA test confirmed that skill match

significantly influenced the model’s decision ($p < 0.001$). A logistic regression model further revealed that gender had a measurable impact on acceptance probability.

Test	Statistic	p-value
Gender t-test	3.21	0.007
Skill Match ANOVA	12.45	< 0.001
Logistic Regression (Gender Effect)	2.87	0.004

Table 1: Statistical tests on gender bias and skill matching.

5 Discussion and Implications

5.1 Origins of Bias

The bias observed may stem from the model’s training data, where gendered associations with professions are historically embedded. Models trained on real-world hiring data may learn implicit patterns that favor male candidates [2].

5.2 Mitigation Strategies

To address this bias, we propose:

- **Anonymization:** Removing names from resumes before processing.
- **Fairness Constraints:** Adjusting training data to balance gender distributions.
- **Bias Audits:** Implementing continuous monitoring with interpretability tools like Goodfire Ember.
- **Adversarial Debiasing:** Using adversarial training techniques to reduce learned biases.

5.3 Ethical Considerations

Ensuring fairness in AI-driven hiring is not just a technical challenge but a socio-ethical imperative. Organizations deploying such systems must adhere to fairness principles and regulatory guidelines [3].

6 Conclusion

Our study demonstrates the presence of gender bias in AI-driven resume screening. Through mechanistic interpretability techniques, we quantified bias and proposed mitigation strategies. Future work should explore larger datasets, additional demographic variables, and model-specific fine-tuning to mitigate bias.

References

- [1] Bolukbasi, T., et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." NeurIPS.
- [2] Hardt, M., et al. (2016). "Equality of Opportunity in Supervised Learning." NeurIPS.
- [3] Rudin, C. (2019). "Stop Explaining Black Box ML Models and Use Interpretable Models Instead." Nature Machine Intelligence.