

Algorithmic Bias in AI Resume Screening

Aidan Bugayong

November 14, 2025

Table of contents

1 Introduction	2
1.1 Context and Background	2
1.2 Project Purpose	2
2 Methods	3
2.1 Data Collection and Purpose	3
2.2 Data Processing	3
2.3 Statistical Tools	4
3 Results and Discussion	5
4 Conclusion	6
5 References	7

1 Introduction

1.1 Context and Background

Resume screeners were developed in order to screen candidates more efficiently and reduce the human bias in the screening process. However, there are concerns with whether or not these automated resume screening systems are truly unbiased in their decision making process. As more and more companies use some form of AI automation in their hiring process, the question of whether or not these systems are unbiased has become more important.

According to Naveen Kumar in his article on AI recruitment statistics, roughly 87% of companies are using AI in the hiring and recruitment process¹. As such, the University of Washington decided to conduct their own study to determine if there is any bias in the decision making process of resume screeners and what sorts of factors contribute to the scoring process of a resume. Their research found “significant racial, gender and intersectional bias in how three state-of-the-art large language models, or LLMs, ranked resumes. The researchers varied names associated with white and Black men and women across over 550 real-world resumes and found the LLMs favored white-associated names 85% of the time, female-associated names only 11% of the time, and never favored Black male-associated names over white male-associated names”². Their research came to these conclusions by handing the AI a list of identical resumes with different names and then having the AI give them scores. The article ends with the research team noting that more research should be done in this area by looking at different attributes and more LLMs in order to better align these AI systems with the real world policies to reduce bias and harm.

1.2 Project Purpose

This project aims to determine if there is any bias in the decision making process of resume screeners and what sorts of factors contribute to the scoring process of a resume. More specifically, looking at a wide variety of applicant attributes to determine what factors have the highest contribution to the bias in the decision making process. Ideally, professional experience and education will be the best predictors of the resume scores but we will also look into other factors such as gender, ethnicity, and institutional prestige.

¹<https://www.demandsage.com/ai-recruitment-statistics/>

²<https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/>

2 Methods

2.1 Data Collection and Purpose

The original list of resumes is from a dataset in huggingface³, which is comprised of both real and synthetic resume data in JSON formatting. The purpose of this dataset was for training natural language processing (NLP) models for resume parsing. Specifically, the resumes are oriented around technical roles and is designed for NLP models to be trained on and used for candidate matching / screening in this field. This dataset was posted on huggingface February 21st, 2025 and has not been updated since (excluding the minor changes to the readme).

The sources used for this dataset are sourced from anonymized CV submissions as well as synthetic resumes generated using “Faker Library” and filled with realistic and role appropriate information. All resumes are anonymized by removing PII (Personally Identifiable Information) but many fields (such as names) contain realistic placeholders. The makers of the dataset note that the data is oriented around technical roles and the synthetic resumes may not capture the same nuances of a real resume. As such, the makers note that this dataset should only be used for NLP, data augmentation, or exploratory data analysis and should not be used for non-technical roles or personalized hiring decisions.

The dataset contains over 4500 resumes in a JSON format. Each resume entry contains personal information, work experience, education information, skills and projects. Since these are technical resumes (oriented around the computer science / information technology field), the skills and projects fields contained a candidate’s coding projects and/or coding languages. For the scope of this project, all fields were used for analysis or scoring of the resume.

2.2 Data Processing

The collection of resumes was processed in order to create the score datasets used for this project. The score datasets has the following columns: Name (acting as the primary key), Resume Score, Gender, Ethnicity, Institutional Prestige, Years of Experience, Skill Relevance, Experience Relevance and Project Relevance. The score datasets are specific to each of the language models used in this project. This is because we are trying to understand the relation between how the model scores a resume and the model’s own perception of the candidate (gender, ethnicity, and institutional prestige).

³<https://huggingface.co/datasets/datasetmaster/resumes>

For creating the score datasets themselves, the resumes were scored by the model and then the model was asked to guess the gender, ethnicity, and institutional prestige of the candidate. For determining the skill/project/experience relevance, that was also a call to the language model. It is important to note that each column of data is a separate instance of the language model to reduce the likelihood of previous responses affecting the current responses. All of this information was then used to create the score dataset for each of the models used in this project.

The models used for this project include IBM's *granite3.3:8b*, Deepseek's *deepseek-r1:8b*, Qwen's *qwen3:8b*, Google's *Gemma3:12b*, Microsoft's *Phi4:14b* and Meta's *llama3.1:8b*. These models were chosen because they are highly rated models despite the lower parameter size and are all open source.

Note that the models are currently subject to change based on runtimes. Deepseek and Qwen are both thinking models so they can take a while to run.

2.3 Statistical Tools

3 Results and Discussion

4 Conclusion

5 References