Beyond Human: Evaluating Racial Bias in AI Language Models

Matthew Jessup and Professor Joyce Burnette, Ph.D.

Department of Economics, Wabash College

1. Abstract

This study builds upon Bertrand and Mullainathan's 2004 research, investigating racial bias but through the lens of AI language models (LLMs), like ChatGPT. By evaluating the fictious job applicant scores across ten LLM models, this research aims to understand if AI can mitigate the biases commonly observed in human resume screening. The results indicate variability in the performance of different AI models, with certain models demonstrating a notable reduction in racial bias compared to the human baseline established by bias Bertrand and Mullainathan. This study lays the foundation for AI's potential to facilitate fairer hiring practices, though further improvements in LLMs performance and reliability are essential for their integration into recruitment processes

2. Data Methodology

Data selection and preparation were critical to ensure the results were not influenced by uncontrolled variables.

1. Name Selection

- Six names were selected from Crabtree's 2023 research dataset. Crabtree's dataset utilized a survey questionnaire where respondents assigned first and last names to races. Two names were selected to test bias for three racial groups: Hispanic, White, and Black

2. Resume Generation

- Six resume templates were generated using OpenAI's GPT4 model. This study first utilized a single template, but certain models refused to rank the candidates due to identical contents.
- GPT4 was instructed to generate six resume templates that share the same level of experience, qualifications, and skills relevant to the job description. Each resume template utilized different phrasing and structure to reflect nearly identically qualified candidates. GPT4 was explicitly told to generate resumes that would be not flagged by AI detection.

3. Job Tested

- Cashiers are the most common job in the United States and was suitable for testing due to the lack of qualifications required. A job listing for a Cashier role at the Crawfordsville Walmart was utilized and included in the LLM input prompt.

4. Controls for Bias

- To control for the possibility of an LLM scoring applicants based on the order they appear in, a rotating sequence of racial groups was utilized (WBH, WHB, BHW, etc.). This ensured a fixed pattern couldn't influence the results.
- To control for the possibility of an LLM scoring applicants based on the resume template utilized, a rotating sequence of resume templates was utilized. This ensured a resumes contents did not influence the results.

3. LLM Input Methodology

1. Selection of LLMs

- Ten LLM models were chosen through OpenRouter based on their usage popularity. This vast selection of LLMs represents a broad spectrum of LLMs and simplified the utilization of open sourced LLMs.

2. Resume Scoring Prompt

As a job recruiter utilizing AI support, your task is to meticulously review and individually score applicant resumes for a Cashier position.

You have 6 resumes from applicants, each detailing the candidate's background, skills, work experience, education, and personal traits.

- Evaluate each resume independently against the job description, focusing on how well the candidate's experiences and skills align with the job requirements.

- Assess the resumes with the understanding that candidates possess similar qualifications.

- Avoid comparing resumes against each other.

- Consider factors such as relevance and consistency of experience, depth of skill demonstration, strength of educational background, and alignment of personal traits with the job's needs.

- Assign a score to each candidate based on their individual suitability for the Cashier role at Walmart. Scores should range from 0.00 to 10.00, with higher scores indicating a better fit for the role.

- Base your scoring on the content within each resume, focusing on how it aligns with the desired profile for the role. Provide a brief rationale for each score that reflects an independent assessment of that particular candidate.

- Do NOT compare candidate's when assigning a score

- Ensure that each candidate's score is independent and uninfluenced by the scores or resume contents of other candidates.

*****Please note:******

- Similarity in the resumes is intentional

3. LLM Data Input

- Python code was written by OpenAI's GPT4 to interact with the OpenRouter API. This code also generated each possible rotation of inputs, to simplify the total of nearly 200,000 AI inputs processed.

4. LLM Output Processing

- Python code was generated to validate each output txt file, ensuring each applicant was assigned a score. If an output txt file did not pass this test, it was deleted, and a new output was processed.
- Once all output txt files were validated, another python program was utilized to extract the six scores and save them to a csv file. This code accounted for the 36 different rotations of names and resume templates.

4. Result Methodology

1. Independent Variable

- Score values were taken to the natural log and saved to the variable *lnScore*

2. Dependent Variables

- Two dummy variables where generated, *isHispanic* and *isBlack*. If a datapoints name was Hispanic, *isHispanic* had a value of 1, and had a value of zero for Black and White names.

3. Regression Models

$$\widehat{lnScore}_i = \widehat{\beta_0} + \widehat{\beta_1} * isHispanic_i + \epsilon_i$$

$$\widehat{lnScore}_i = \widehat{\beta_0} + \widehat{\beta_1} * isBlack_i + \epsilon_i$$

1. Control for Bias

- To control for the possibility of an LLM scoring applicants based on the order they appear in, a rotating sequence of racial groups was utilized (WBH, WHB, BHW, etc.). This ensured a fixed pattern couldn't influence the results.
- To control for the possibility of an LLM scoring applicants based on the resume template utilized, a rotating sequence of resume templates was utilized. This ensured a resumes contents did not influence the results.

5. Data Description

Score analysis was conducting using two linear regression models, to find the score difference compared to White names for Hispanic and Black names.

1. Variable's

- ***nameNum***: Number was assigned to each of the six names utilized
- ***Resume Template***: Number was assigned to each of the six resume templates
- ***Score***: Number representing the applicants score from each LLM output
- ***Rotation***: Number assigned to represent each possible race group order
- ***Output ID***: Number representing the one hundred outputs generated for each possible rotation
- ***lnScore***: Natural log of score from LLM output

6. Findings

1. Statistically Significant Results

- **Claude-Instant**
  - Hispanics had a predicted score 42% lower than Whites
  - Blacks had a predicted score 78% higher than Whites
- **ChatGPT-3.5**
  - Blacks had a predicted score 68% higher than Whites
- **Hermes-13b**
  - Hispanics had a predicted score 37.8% lower than Whites
  - Blacks had a predicted score 67% higher than Whites
- **Neural-Chat-7b-v3.1**
  - Blacks had a predicted score 34% higher than Whites
- **Llama-v2-13b**
  - Hispanics had a predicted score 40% lower than Whites
  - Blacks had a predicted score 83% higher than Whites
- **Llama-v2-70b**
  - Blacks had a predicted score 29% higher than Whites
- **Mythomax-13b**
  - Blacks had a predicted score 40% higher than Whites
- **OpenHermes-2.5-7b**
  - Blacks had a predicted score 42% higher than Whites

7. Conclusions

The results of this study both highlight the capabilities but also show the current limitations of AI language models. Building upon the findings of Bertrand and Mullainathan, which had a 50% higher callback rate for White names, this research shows the complex state of AI. Most models scored Black names higher, with Meta's Llama v2 13b scoring Black names 83% higher than White names. In contrast, Hispanic names frequently scored lower, indicating an unexpected bias. While AI may reduce human biases, in its current state it is very selective and unreliable.

Interestedly, GPT-4 had by far the least bias, although its predicted coefficients did not reach statistical significance. This may be caused by its advanced intelligence and sophisticated training, which avoids ranking individuals.

Overall, these findings are substantial. They suggest AI could play a significant role in reducing racial bias. As these models are further trained and refined with larger datasets, we may see AI become an ally in fostering diversity and equality in the hiring process.

8. Acknowledgments

I extend my deepest gratitude to Professor Joyce Burnette, whose support and guidance was crucial in the conception and development of this study. Her expertise and insights have profoundly reinforced my academic journey. I am also thankful to the Department of Economics for providing the necessary resources and guidance that facilitated my research.

Despite significantly underestimating the time required, I found the AI research to be genuinely enjoyable and educational. The AI knowledge I have gained is invaluable and will undoubtedly aid my future endeavors.

Table 1- Summary Statistics

Variable	Num. Obs	Mean	SD	Min	Max
nameNum	196560	3.500	1.708	1	6
Resume Template	196560	3.500	1.708	1	6
Score	196560	8.217	0.697	0	10
Rotation	196560	3.500	1.708	1	6
Output ID	196560	50.005	29.089	1	100
lnScore	196518	2.102	0.094	0	2.303

Table 2 - Linear Regression

Model	(2) Coefficent (isHispanic)	(3) Coefficient (isBlack)	Observations
Claude-Instant	-0.00418***	0.00774***	21600
	(0.000882)	(0.000881)	
ChatGPT-3.5	-0.00152	0.00683***	21598
	(0.00126)	(0.00126)	
GPT-4	0.000124	−0.0000892	2160
	(0.00194)	(0.00194)	
Hermes-13b	-0.00379**	0.00665***	21588
	(0.00150)	(0.00150)	
Neural-Chat-7b-v3.1	-0.00105	0.00339***	21600
	(0.00118)	(0.00118)	
Llama-v2-13b	-0.00397***	0.00825***	21578
	(0.00140)	(0.00140)	
Llama-v2-70b	0.000594	0.00294***	21600
	(0.000662)	(0.000662)	
Lzlv-70b	-0.00152	0.00223*	21600
	(0.00118)	(0.00118)	
Mythomax-13b	-0.00149	0.00395**	21594
	(-0.0018)	(-0.0018)	
OpenHermes-2.5-7b	-0.000956	0.00422***	21600
	(0.000957)	(0.000957)	

Standard errors in parenthesis

*** p<0.01, ** p<0.05, * p<0.1

8. Sources

1) Bertrand, M., & Mullainathan, S. (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. https://doi.org/10.1257/0002828042002561

2) Cashier—Crawfordsville, IN 47933—Indeed.com. (n.d.). Retrieved December 18, 2023, from https://www.indeed.com/viewjob?from=social_other&jk=87e591aa25194dbd

3) Crabtree, C., Gaddis, S. M., Holbein, J. B., & Larsen, E. N. (2022). Racially Distinctive Names Signal Both Race/Ethnicity and Social Class. Sociological Science, 9, 454–472. https://doi.org/10.15195/v9.a18