XX-000

Proxy Recognition and Inclusive Scoring Method (PRISM): Evaluating Context-Dependent Bias in Large Language Models for Resume Screening

Abstract

AI-driven hiring tools are reshaping recruitment but often mirror biases in their training data. PRISM examines how large language models express or reduce demographic bias during resume evaluation and how linguistic context within prompts shapes these outcomes. Using a controlled dataset of 324 synthetic resumes with racially neutral surnames; differing only by first name as the demographic proxy, we compared GPT 3.5 turbo with a Sentence BERT similarity model. Under neutral prompts, no stable bias was observed across demographic groups, yet contextual shifts in the prompt changed how the model responded to proxy cues. These findings show that LLM bias can be either activated or suppressed depending on prompt framing, highlighting the significance of context-aware prompt design in improving fairness.

Introduction

Employers increasingly use AI systems to screen resumes, raising concerns about whether models infer demographic traits from proxy variables such as first names or subtle cultural cues. These unintended associations can influence scoring even when applicants have identical qualifications.

Large language models absorb patterns they were never meant to learn, linking small demographic hints to assumptions about job fit. Because LLMs adjust their outputs based on linguistic context inside prompts, bias can appear or disappear depending on evaluator instructions.

The Proxy Recognition and Inclusive Scoring Method (PRISM) was developed to examine how contextual framing and subliminal learning shape LLM behavior during resume evaluation. Since LLMs change their responses based on prompt wording and context, prompt design becomes a key factor in improving fairness.

Research Question(s)

1. Does subliminal learning contribute to unintended patterns that influence resume scoring, and how can better prompt engineering help reduce these effects?
2. To what extent does prompt wording or contextual framing change how LLMs score candidates, even when all other variables remain the same?
3. How do large language models respond to identical resumes that differ only by first name, and do these responses reveal demographic bias under neutral prompting conditions?

Materials and Methods

We built a dataset of 324 synthetic resumes that were identical in every qualification, with only the first name changed to represent five demographic groups. Each resume used a racially neutral surname and was matched with standardized job descriptions from three industries to see whether context influenced scoring. All evaluations were done with GPT 3.5 turbo using a simple, neutral prompt. The prompt asked the model to rate the candidate's likelihood of being invited to an interview. We also compared these results with a Sentence-BERT similarity model to see how an embedding-based system handled the same data. Scores were recorded on a 0 to 100 scale and analyzed using ANOVA and pairwise t tests to check for group differences. This design allowed us to isolate name-based proxy effects and examine how prompt framing shapes fairness in AI-driven hiring.

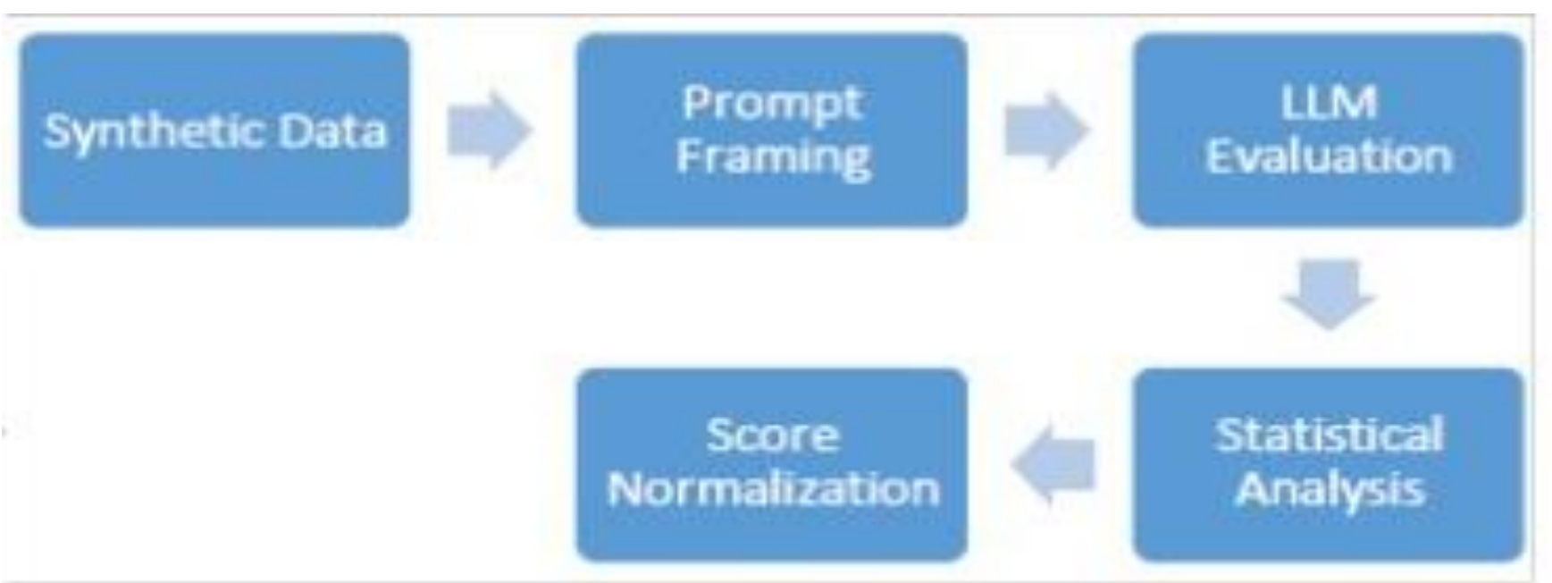

Synthetic Data
Prompt Framing
LLM Evaluation
Statistical Analysis
Score Normalization

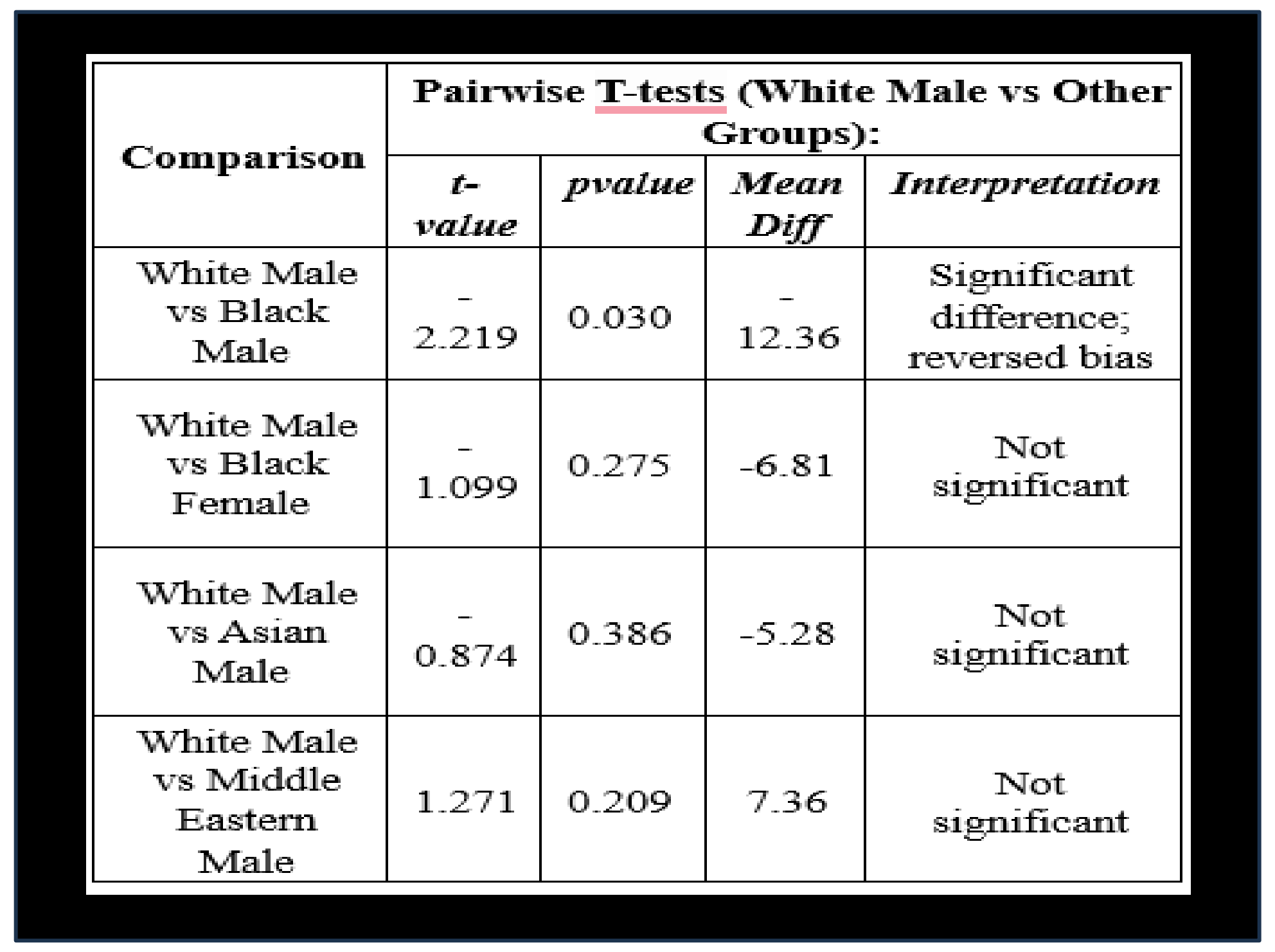
Fig. 1: PRISM Experimental Pipeline
This diagram illustrates the PRISM workflow, including synthetic data generation, contextual prompt framing, LLM evaluation, score normalization, and statistical testing used to detect context-dependent, proxy-based bias.

Results

Across all industries, resume scores remained tightly grouped, with averages ranging from the low 60s to low 70s. A one-way ANOVA test (F = 0.901, p = 0.516) showed no statistically significant differences across demographic groups, and no consistent pattern of bias emerged.

In some evaluations, Black and Latina candidates scored slightly higher under the neutral prompt, which may reflect moderation or overcorrection effects within GPT 3.5 turbo's alignment layer. Pairwise Welch's t tests supported this finding, showing no meaningful score gaps between groups.

These outcomes indicate that demographic bias was not statistically expressed but rather dependent on prompt framing and linguistic context.



Comparison	Pairwise T-tests (White Male vs Other Groups):			
	t-value	*pvalue*	*Mean Diff*	*Interpretation*
White Male vs Black Male	-2.219	0.030	-12.36	Significant difference; reversed bias
White Male vs Black Female	-1.099	0.275	-6.81	Not significant
White Male vs Asian Male	-0.874	0.386	-5.28	Not significant
White Male vs Middle Eastern Male	1.271	0.209	7.36	Not significant

Fig. 2: Pairwise Welch's t Test Results
This table compares White male applicant scores with other demographic groups under the neutral prompt. Only the comparison with Black male applicants showed a significant reversed bias effect, while all other differences were not statistically significant.

Conclusions

The results suggest that GPT 3.5 turbo does not exhibit stable demographic bias under a neutral, minimal prompt, even though earlier model versions demonstrated clear disparities. Instead, bias appears to be context dependent, shifting based on the linguistic framing and evaluator instructions within the prompt.

We also found that alignment mechanisms may suppress or moderate underlying associations rather than fully remove them. This means bias can be activated or muted depending on how the model is prompted, rather than being a fixed characteristic of the model itself.

Overall, these findings reinforce that prompt context and framing play a central role in how demographic bias surfaces during AI-driven resume screening and should be carefully considered in real-world hiring systems**.**

References

[1] M. Bertrand and S. Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, vol. 94, no. 4, pp. 991–1013, Sep. 2004.
[2] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.
[3] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.
[4] L. Weidinger *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint* arXiv:2112.04359, 2022.
[5] K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, San Jose, CA, 2024, pp. 1578–1590.
[6] K. Wilson and A. Caliskan, "AI tools show biases in ranking job applicants' names according to perceived race and gender," *University of Washington News*, Oct. 2024. [Online]. Available: https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/
[7] L. Zhang, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanities and Social Sciences Communications*, vol. 10, no. 567, Sep. 2023.
[8] J. Zou and L. Schiebinger, "AI can be sexist and racist—it's time to make it fair," *Nature*, vol. 559, pp. 324–326, 2018.

Acknowledgments

We acknowledge the following faculty for their contributions to this project:

- **Professor Md Abdullah Al Hafiz Khan**, instructor for Natural Language Processing (CS 7347), whose feedback helped inform adjustments to our project
- **Faculty of the Artificial Intelligence and Machine Learning courses**, whose instruction contributed to the foundational knowledge supporting this research.

We also thank Kennesaw State University for providing an academic environment that facilitated our research and supported the completion of this project.

Contact Information

Crystal Tubbs
AI Solutions Architect & Emerging Technologies Specialist
✉ *Ctubbs2@students.Kennesaw.edu*
🌐 *https://github.com/Msmetamorphosis*
🔗 *https://www.linkedin.com/in/crystal-tubbs-8a9b245b/*

Destiny Raburnel
Data Engineer
✉ *Draburne@students.Kennesaw.edu*
🌐 *https://github.com/destinyjj621*
🔗 *https://www.linkedin.com/in/destinyraburnel/*

Experience and Future Directions

Crystal Tubbs is an AI Solutions Architect and Emerging Technologies Specialist with experience designing applied AI systems, custom LLM agents, context-engineered workflows, and advanced prompt engineering methodologies for small business and enterprise environments. She has contributed to industry projects supporting leading AI labs, including OpenAI, Meta, Google, and Anthropic, with hands-on experience in dataset creation, model training, evaluation, and applied AI behavioral analysis. Crystal holds an AS in Technology Management, a BS in Management and Organizational Leadership, and is completing the MS in Artificial Intelligence at Kennesaw State University.

Destiny Raburnel is a Data Engineer with experience in enterprise data workflows, applied machine learning, and LLM-based workflow development. She has built automation solutions that integrate structured and unstructured data and has experience shaping LLM outputs for workflow optimization and domain-aligned tasks. Her diverse technical background enables her to bridge data engineering, software development, and applied AI in practical, solution-driven environments. She holds a BS in Biomedical Engineering, an MS in Software Engineering, and is completing an MS in Artificial Intelligence at Kennesaw State University.

We plan to expand this work by incorporating additional prompt types, evaluating newer LLM families, and scaling the dataset to include more diverse proxy variables. This project strengthened our interest in how linguistic cues and contextual framing shape model behavior. As we continue our academic development in AI and NLP, we aim to apply these methods to broader domains such as cybersecurity, blockchain, fintech, and responsible AI evaluation. We look forward to building on the insights gained from PRISM as we advance professionally and academically.

Project website: https://msmetamorphosis.github.io/PRISM-Project/

KENNESAW STATE
UNIVERSITY
COLLEGE OF COMPUTING AND
SOFTWARE ENGINEERING

Author(s) Crystal Tubbs and Destiny Raburnel
Advisors(s) Md Abdullah Al Hafiz Khan