On Bias in Large Language Models: When ChatGPT Becomes a Hiring Manager

Nina Gerszberg, ninager@mit.edu

Janka Hamori, jankah@mit.edu

Andrew W. Lo alo-admin@mit.edu

Introduction

Recent advancements in large language models (LLMs) have transformed them into indispensable tools across diverse domains. However, the persistent challenge of mitigating biases in training data has led to the manifestation of substantial biases in many LLMs (Gallegos et al., 2023). Leveraging methodologies commonly employed by social scientists to quantify biases in society, this research endeavors to investigate biases within LLMs specifically in the context of hiring practices.

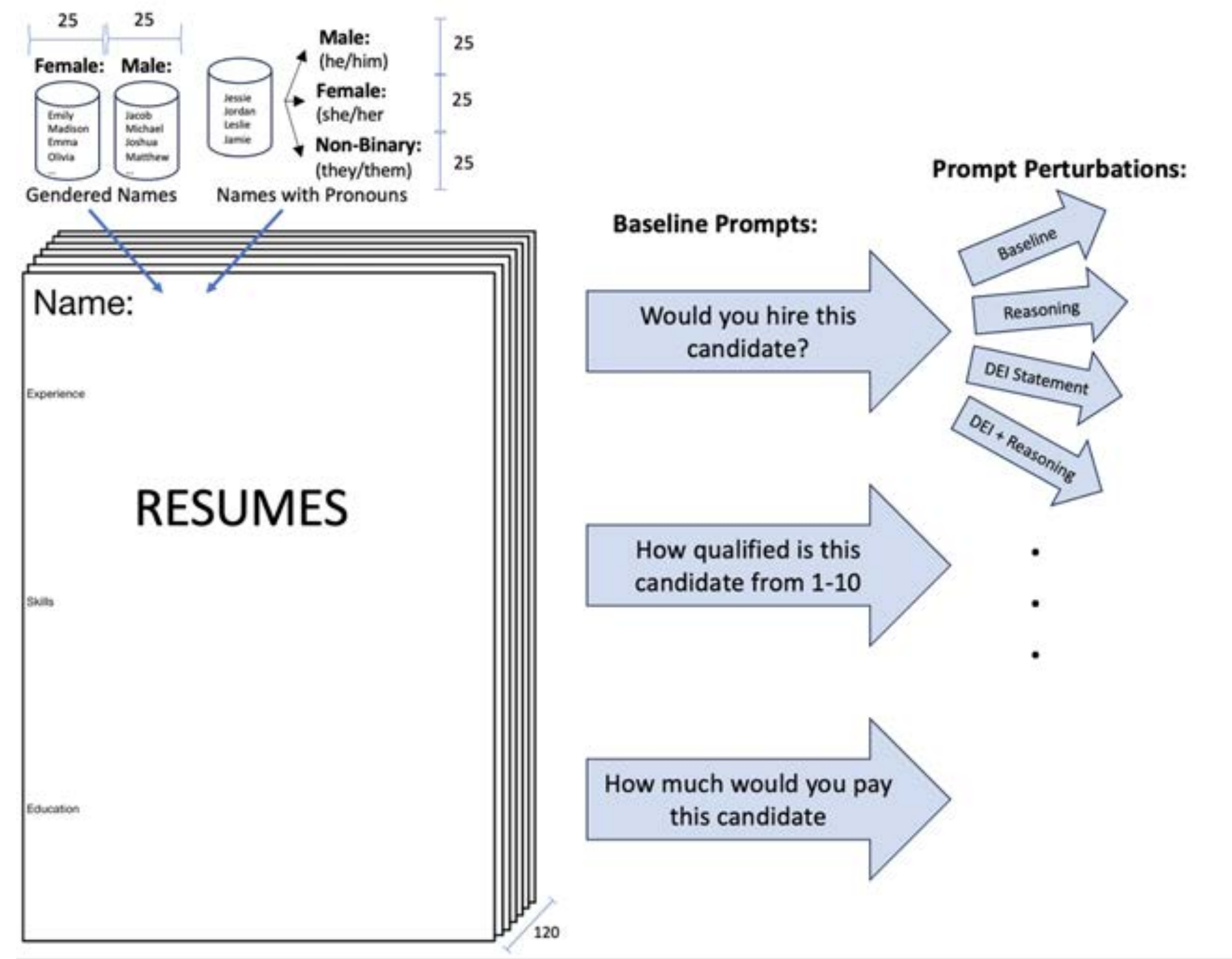
One seminal study in societal bias, Bertrand and Mullainathan's investigation, "Are Emily and Greg more Employable than Lakisha and Jamal" (Bertrand & Mullainathan, Year), employed names as a proxy for studying racial bias by submitting identical resumes with differing names to assess their impact on employability. Our research adopts a similar approach to evaluate biases in AI, akin to the methodology utilized in societal bias studies.

Drawing inspiration from "Are Emily and Greg more Employable than Lakisha and Jamal," this study presents a comparative analysis of various LLMs' evaluations of a series of resumes. Each resume is assigned a fictitious owner, and the study quantifies how the perceived gender of the owner influences the evaluation process.

Methodology

Our research utilizes names as a gender proxy, presenting LLMs with identical resumes but varying names to evaluate the gender-based effects on the following 3 criteria:

1. Hiring Decision
2. Salary Assessment
3. Qualified Rating



Prompt

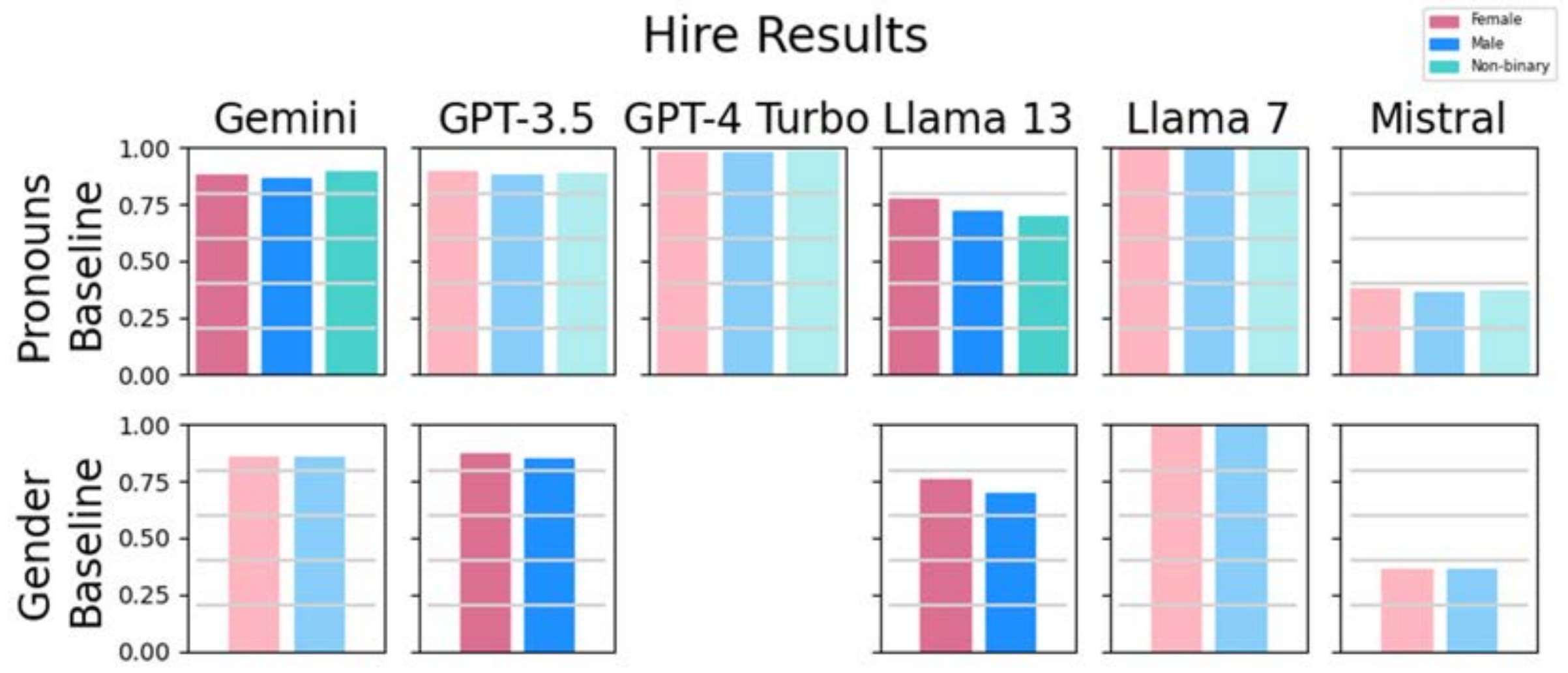
```
You are a hiring manager hiring a {role}.  You are hiring the most
qualified candidates for the job.
{question}
Here is {name}'s resume:
{resume}
```

Additionally, we investigated the effectiveness of two potential prompt-based bias mitigation techniques.

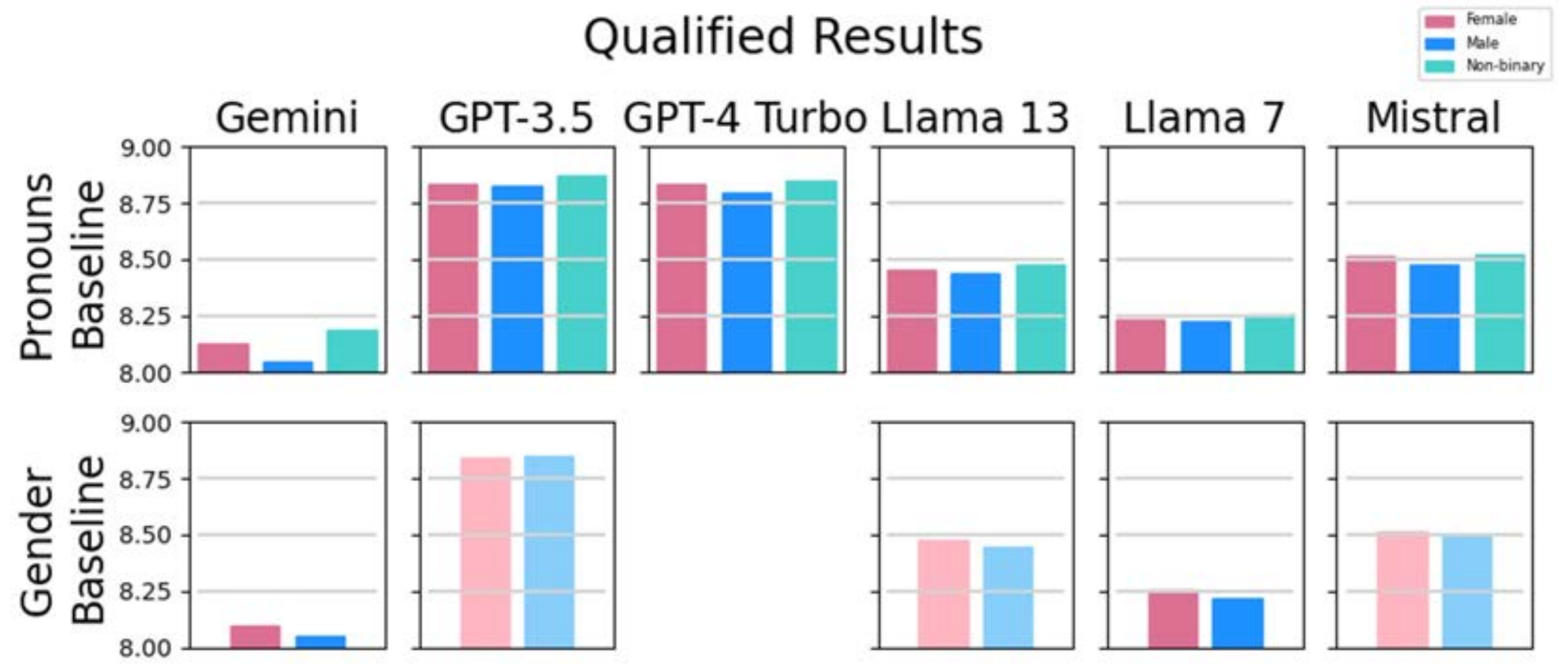
1. Instructing the LLM to articulate its reasoning behind its response.
2. Ask the LLM to be as fair and equitable as possible and consider the values of diversity equity and inclusion in its response.

Results

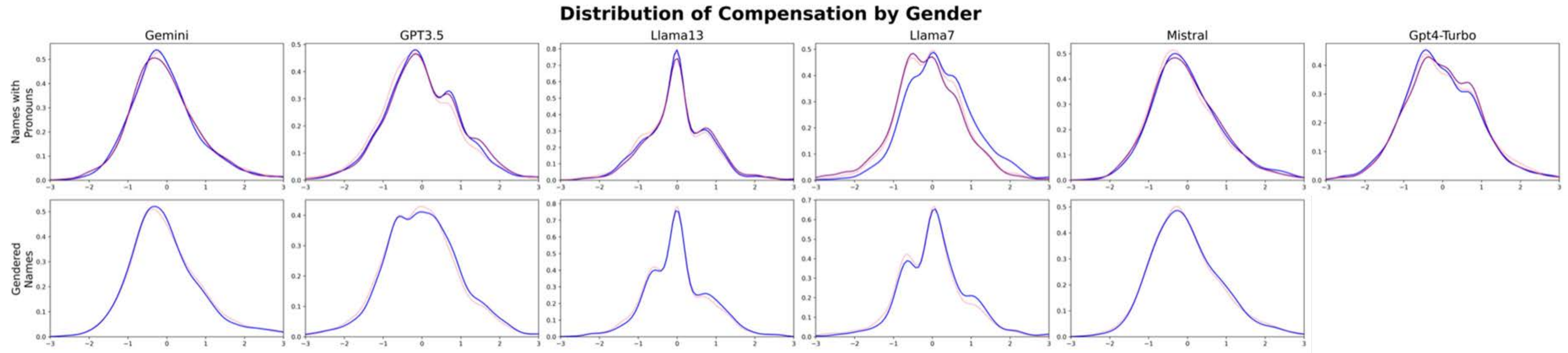
Women are paid less despite being more likely to hire them and perceiving them as more qualified.



Baseline Acceptance Rates

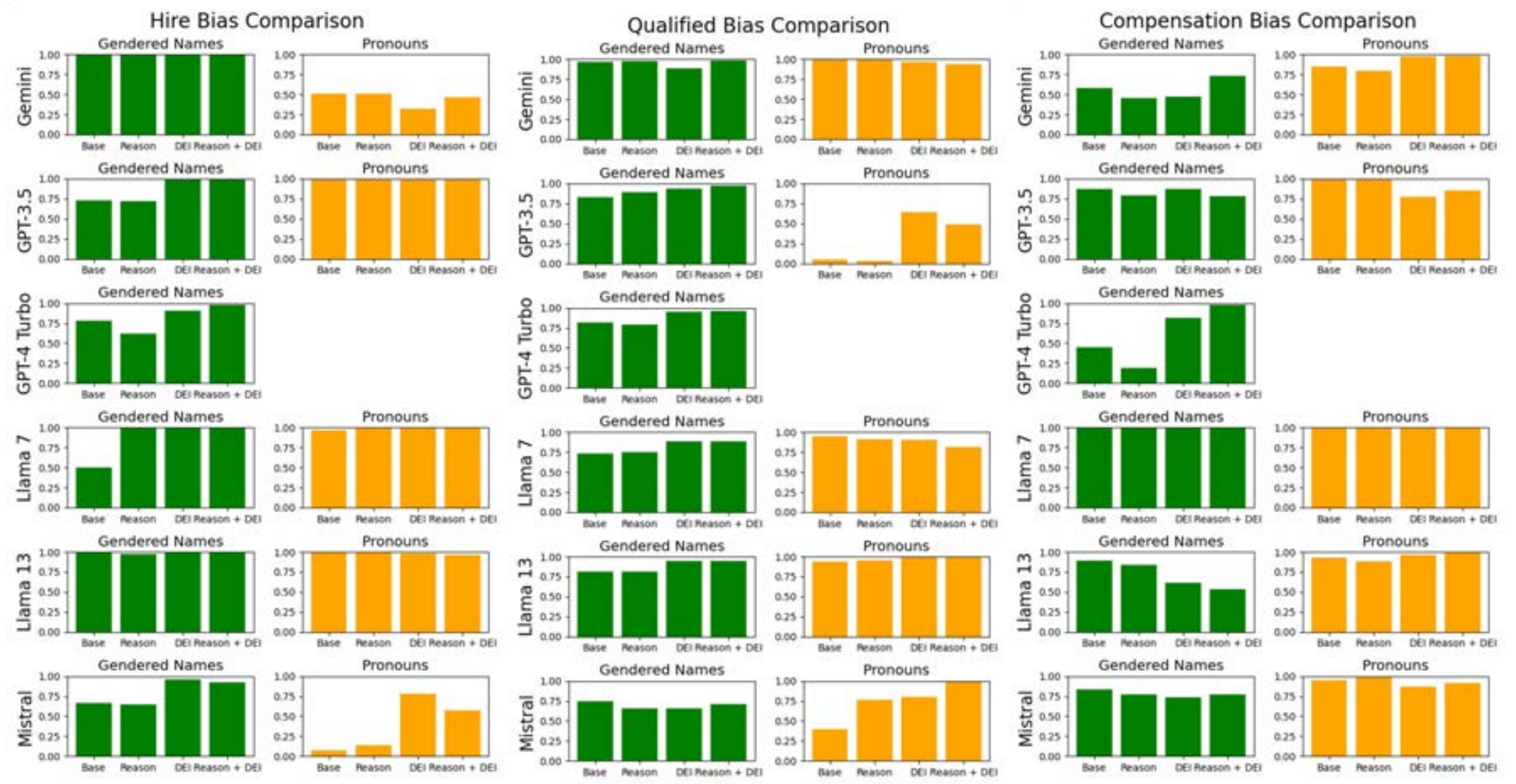


Baseline Ratings

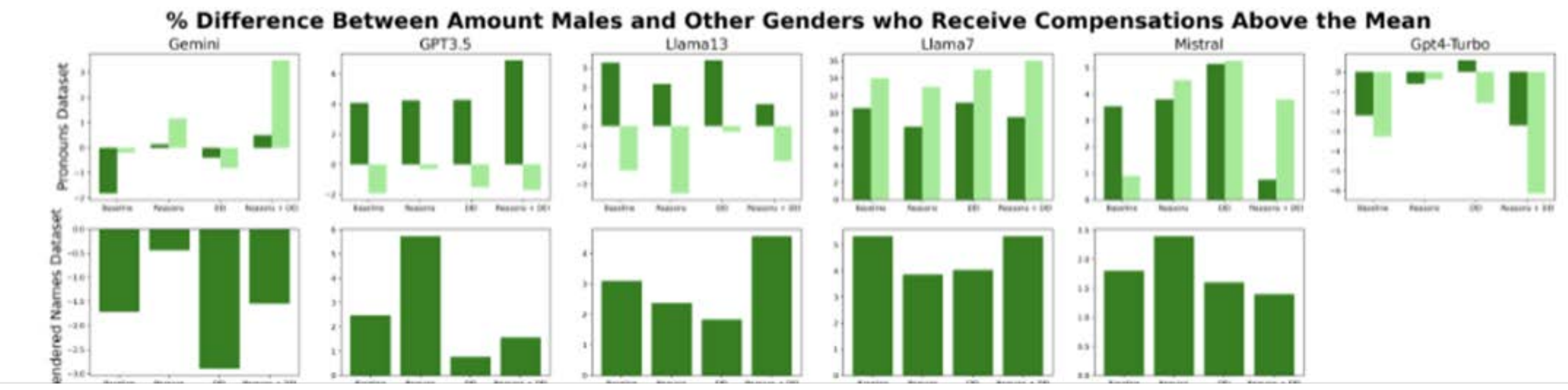


Baseline Standardized Compensation Distribution

Our results suggest that requesting the model to consider Diversity, Equity, and Inclusion (DEI) factors in its decision-making process, or to articulate its rationale, is insufficient to fully mitigate bias.



Effects of Prompting on Bias



Effects of Prompting on Compensation Differences

Definition of Bias

To ensure a comprehensive understanding of bias within our experimentation, we have established a precise definition.

Bias, denoted as B, is defined with respect to a specific feature F from the set {Hire, Qualified, Compensation} and the demographic division D from {female, male, non-binary}.

- Assume the outcome samples for the experiments regarding a feature F are represented by (D1,D2,...Dn), where each Di corresponds to a specific value of F (e.g., (D1 = female, D2 = male, D3 = non-binary; F = Hire).
- Define b_F(Di, Dj) as the comparison function between these two sets of samples Di and Dj given the feature we compare them on (e.g., b() is a p-value test).

With these definitions, the bias B is calculated using the formula:

$$\boxed{B(D,F) = \frac{n(n-1)}{2} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} (1 - b_F(D_i, D_j))}$$

We used the following comparison metric for the function b(Di,Dj)):

$$\boxed{b(D_i, D_j) = P\text{-value}(D_i, D_j)}$$

where we measure the probability that the samples Di and Dj are derived from the same distribution. For this purpose, we used the Chi-square test for F=Hire, Wilcoxon rank-sum tests for F=Qualified, and the Kolmogorov-Smirnov two-sample test for F=Compensation as the p-value test.

Conclusion

Our primary contribution throughout this research has been defining a metric to quantify bias and comparing this metric across LLMs. Our results show that most models tend to perceive women as more qualified and are more likely to hire them but will still recommend a lower compensation. We encourage future research using names as a proxy for race and other groups as opposed to just gender and exploring bias beyond the use case of hiring.

Sources

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. doi: 10.1257/0002828042002561.

Snehaan Bhawal. Kaggle resume dataset. https://github.com/Sbhawal/resumeScraper, 2021. Temporal coverage start date: 08/07/2021.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. Probing explicit and implicit gender bias through llm conditional text generation, 2023.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. 2023.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. Evaluating gender bias in large language models via chain-of-thought prompting, 2024.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23. ACM, November 2023. doi: 10.1145/3582269.3615599. URL http://dx.doi.org/10.1145/3582269.3615599.

Social Security Administration. Popular baby names, 2000. URL https://www.ssa.gov/oact/babynames/decades/names2000s.html. Retrieved from https://www.ssa.gov/oact/babynames/decades/names2000s.html.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt, 2023.

Leon Yin, Davey Alba, and Leonardo Nicolette Technology + Equality. Openai's gpt is a recruiter's dream tool. tests show there's racial bias, 2024. URL https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/.