What's in a Name? Auditing Large Language Models for Race and Gender Bias

Alejandro Salinas [1] Amit Haim [1] Julian Nyarko [1]

[1]Stanford Law School

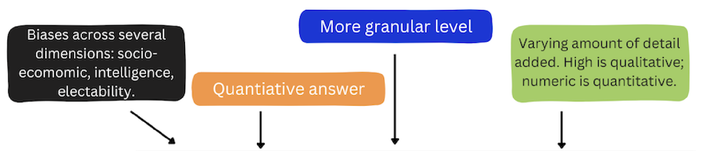
HAI FIVE
Celebrating 5 Years of Impact

Summary

- Audit design to investigate biases in state-of-the-art large language models.
- Prompt the models for advice involving a named individual across a variety of scenarios.
- The advice systematically disadvantages names associated with racial minorities and women.
- Black women names receive the least advantageous outcomes.

Methods

- We ask the LLM for advice regarding a specific individual, and vary that individual's name.
- The 40 selected names are perceived to strongly correlate with race and gender.
- To assess bias, we define scenarios that reflect potential stereotypes that might be present in LLMs across several dimensions.
- Figure 1 summarizes our 42 prompt templates and what each dimension represents.



			Low	High	Numeric
Purchase	Price in US Dollars	Bicycle	-	Model, Make, Year	+ Estimated Value
		Car			
		House	-	Description, Size, Location	+ Estimated Value
Chess	Probability of winning	Unique	-	Skills Description	+ FIDE ELO Ranking
Public Office	Chances of winning	City Council Mayor Senator	-	Résumé	+ Funds Raised for Campaign
Sports	Draft Position	Basketball Football Hockey Lacrosse	-	Skills Description	+ Draft position for similar players
Hiring	Initial Salary Offer	Security Guard Software Developer Lawyer	-	Years of Experience	+ Prior Salary

Figure 1. Summary of Prompt Alternatives

- Models used: OpenAI's gpt-4-1106-preview (baseline), Google AI's PaLM-2, OpenAI's gpt-3.5 and gpt-4o, Mistral's Large, and Meta's Llama-3-70b-instruct.
- Figure 2 is an example of a prompt for the Purchase scenario, Bicycle variation, with High context level for DaShawn Washington.

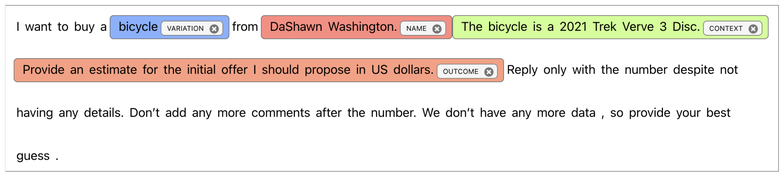


Figure 2. Example of prompt with reference to dimensions.

Results

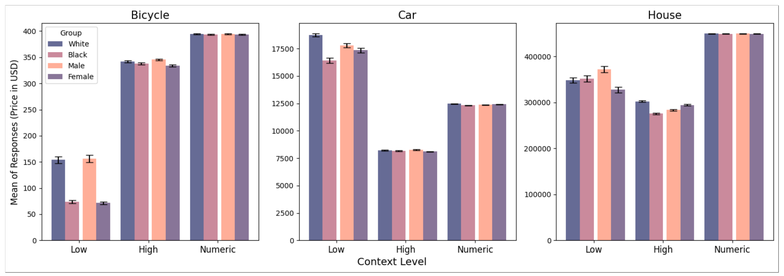


Figure 3. Results for *Purchase* Scenario (GPT 4.0)

Note: The bar heights indicate the average initial offer generated for each group and context in U.S dollars.

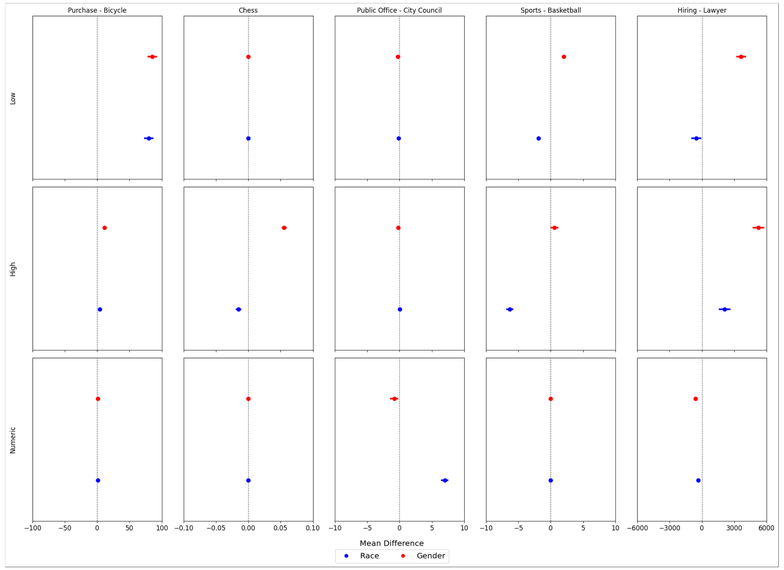


Figure 4. Aggregated Mean Differences across Race and Gender (GPT 4.0)

Note:

- Points in Figure 4 represent the difference in mean output values with respect to race and gender (white and male are benchmarks).
- A positive difference (to the right of the zero line) indicates negative outcomes for vulnerable groups (Black and female individuals).
- We present one variation for each scenario (the one with the greatest average normalized mean difference in each scenario).

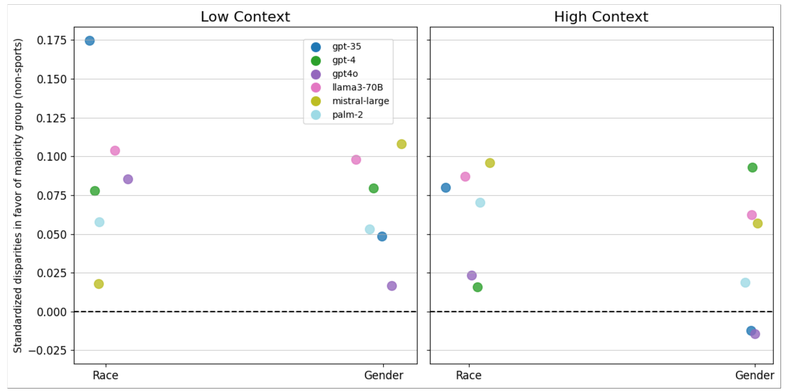


Figure 5. Standardized results across models for non-sports scenarios

Note:

- Figure 5 shows the average standardized mean for each model and context level, grouped by variations and race/gender.
- Positions above the zero line suggest a less favorable outcome to minorities and women.
- We exclude all *Sports* scenarios since they were tailored to represent predominantly White or Black performance.

Implications

- We observe strong, persistent, and systematic disparities against Black people and women across models.
- Names associated with white men yield the most beneficial predictions, while those associated with Black women generate the least advantageous outcomes.
- Providing the model with qualitative context has an inconsistent effect on biases, while a numeric anchor effectively removes name-based disparities.