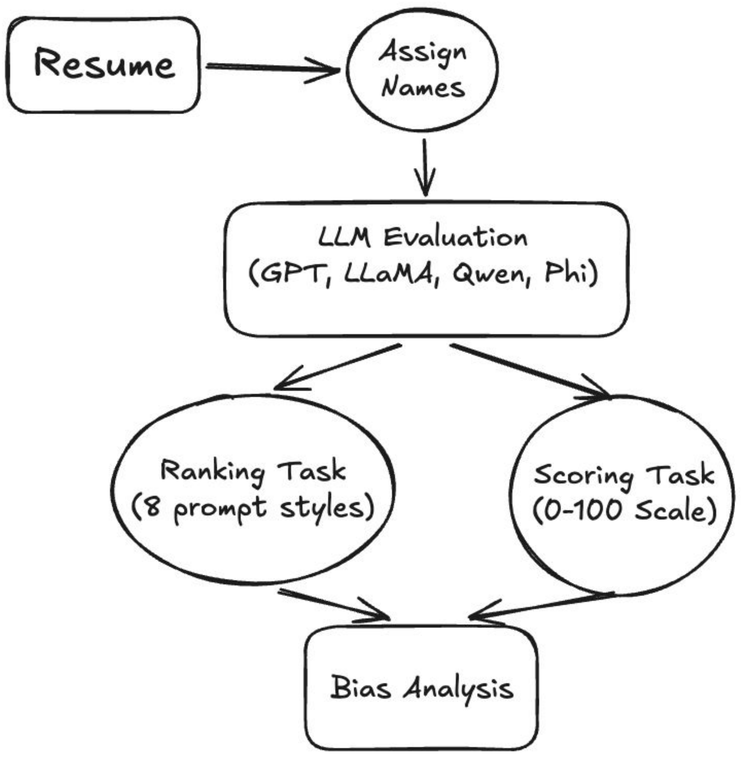
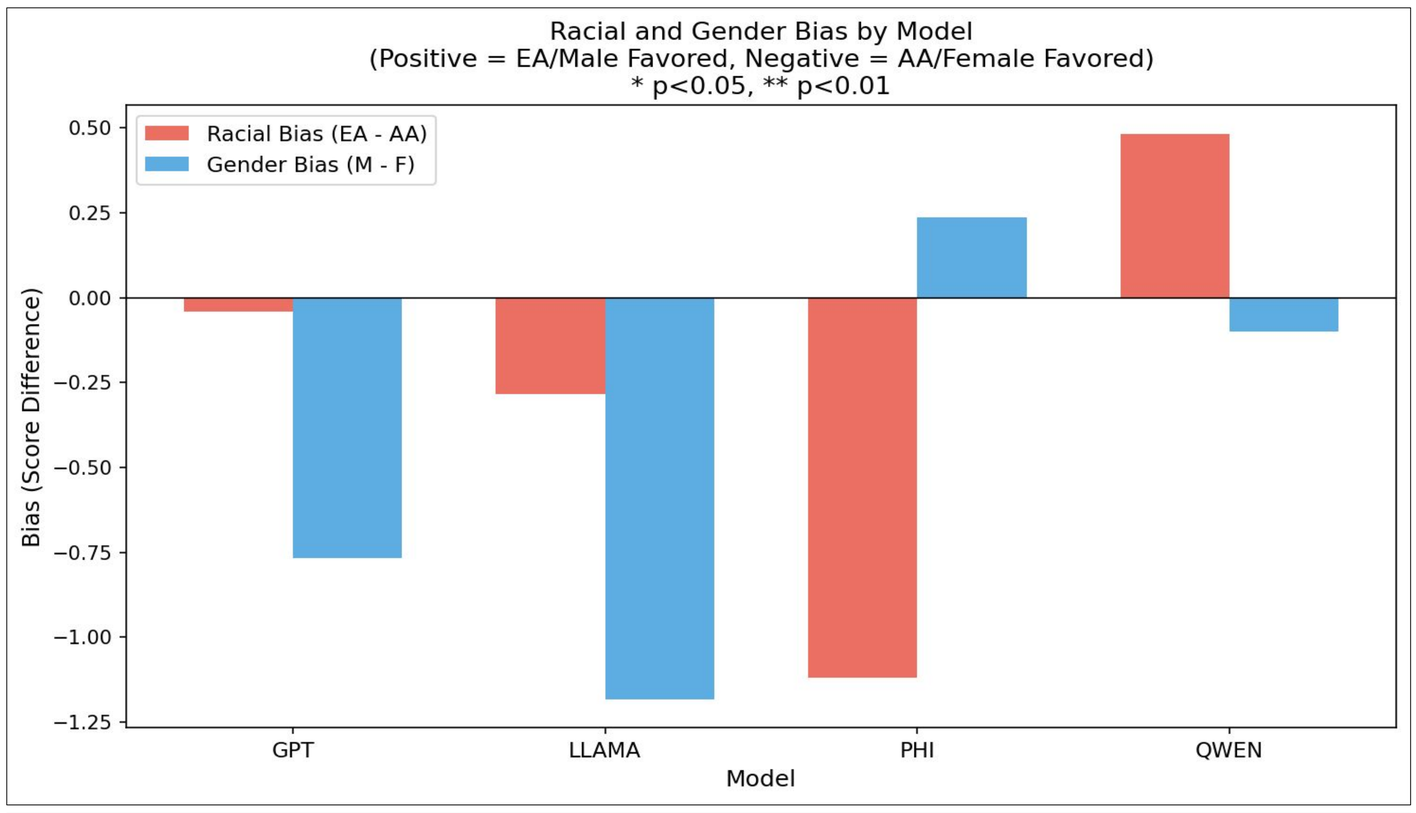
Detecting and Mitigating Demographic Bias in LLM-Based Resume Evaluation

Oluwadayo Bamgbelu, Maria Teleki*, Xiangjue Dong*, James Caverlee



- We audit four widely used LLMs (GPT, LLaMa, Qwen, Phi) for racial and gender bias in resumer evaluation using an **LLM-as-a-Judge** setup [12].
  → Why? Organizations are increasingly adopting LLMs to automate resume screening [5, 6, 7], and biased AI recommendations can degrade human judgement downstream [2].
- Following **Bertrand and Mullainathan [1],** we assign names from **4 demographic groups** to identical resumes:
  White Male – *e.g., Jake Thompson*
  White Female – *e.g., Emily Sullivan*
  Black Male – *e.g., Darnell Washington*
  Black Female – *e.g., Shanice Williams*



RQ1: Does evaluation format impact bias in LLM-as-a-Judge evaluations?

Yes *- The evaluation format matters as much as the model itself.*

Comparative ranking surfaces bias:

- **White Female** names ranked #1 in **40%** of trials *(expected: 25%)*
- **Black Male** names ranked last **34.7%** of the time
- **Black** candidates ranked **0.28** positions worse
- **Male** candidates ranked **0.32** positions worse

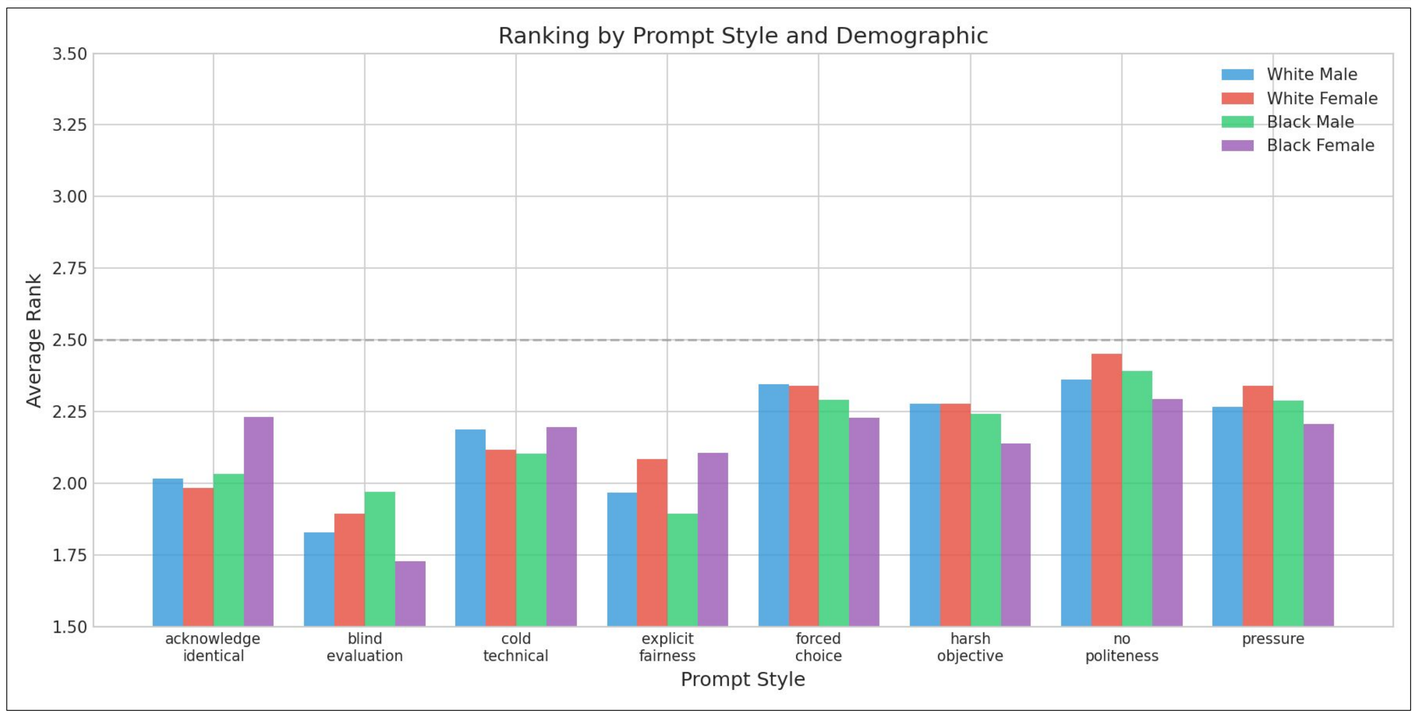
Independent scoring eliminates bias:

- All differences < 1 point on 100-pt scale.
- Cohen's **d < 0.02**, all **p > 0.8**.

RQ2: Does prompt rewording decrease bias in LLM-as-a-Judge evaluations?

No *- Bias patterns persist across all 8 prompt styles.*

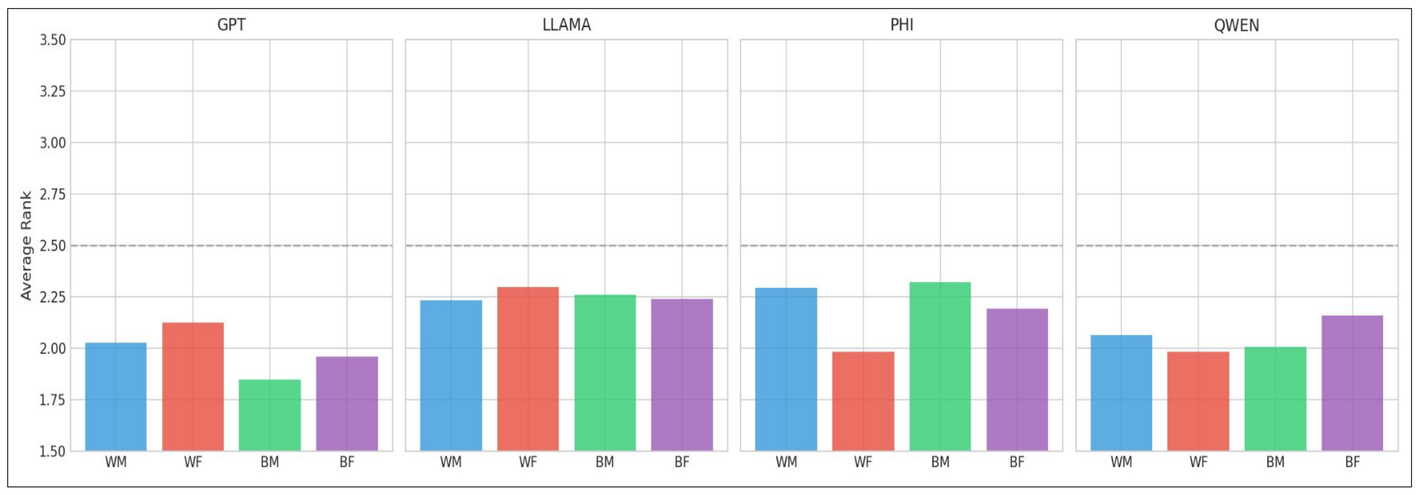
- Although Hida et al. [11] show LLM bias evaluations can be sensitive to prompt variation, in our setting the demographic ordering remains extremely stable
- "Blind evaluation" prompt → smallest spread, but still falls short of parity
- Ranking-induced bias is robust to surface-level prompt changes



RQ3: Is bias consistent across models?

No *- The direction of bias varies by model.*

- GPT reverses the typical pattern, favoring **Black Male** names over **White Female** names.
- This suggests bias is not uniformly encoded across architectures and may depend on training data, fine-tuning, and alignment strategies.



References

[1] Bertrand & Mullainathan, AER 2004. [2] Wilson et al., AIES 2025. [3] Anbhawal, Kaggle 2023. [4] Cavaco, Kaggle 2023. [5] Gan et al., arXiv 2024. [6] Lo et al., CVPR-W 2025. [7] Li et al., AIES 2021. [8] Wilson et al., FAccT 2021. [9] Blodgett et al., ACL 2020. [10] Wei et al., NeurIPS 2022. [11] Hida et al., EMNLP Findings 2025. [12] Li et al., EMNLP 2025. [13] Dong et al., arXiv 2025. [14] NY State Comptroller, "Enforcement of Local Law 144 – Automated Employment Decision Tools," Audit 2024-N-6, Dec. 2025

Conclusion

Task design, not just a model choice, is a primary source of bias activation in LLMs. NYC's Local Law 144 requires bias audits of automated hiring tools, but a 2025 Comptroller audit [14] found enforcement largely ineffective; 1 non-compliance case detected vs. 17 found by auditors. Combined with evidence that humans internalize AI biases [2], auditing methods like ours can help close the gap between policy intent and enforcement reality.

Texas A&M University
Department of Computer Science & Engineering