Enhancing LLM-Driven Bias Detection in Healthcare: Agentic Workflows and Hybrid Models for Racial Disparity Mitigation

Sihao Xing X00230945@mytudublin.ie

MSC. HUMAN CENTERED ARTIFICIAL INTELLIGENCE
TECHNOLOGICAL UNIVERSITY DUBLIN

Abstract

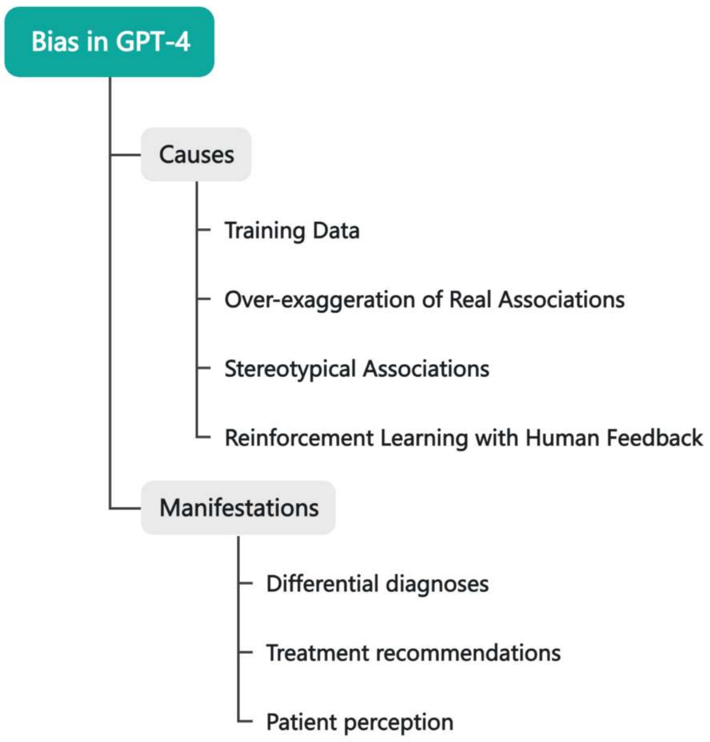
The increasing use of large language models (LLMs) in bias detection requires rigorous evaluation of their reliability, especially in addressing domain-specific biases like racial health disparities in healthcare. This study begins by ensuring that LLMs used for healthcare-related tasks have minimal inherent racial bias. Two main approaches are adopted: (1) developing standardized metrics to assess and compare racial bias levels in LLMs trained on clinical and demographic data, and (2) selecting LLMs that have already demonstrated fairness in healthcare settings as a baseline for further tasks. Building on this foundation, the research explores how agentic workflows and hybrid models can improve LLMs' ability to detect and reduce racial bias in healthcare algorithms. Through benchmarking and automated scoring, we evaluate the accuracy, reliability, and ethical soundness of these approaches in real-world scenarios, such as diagnostic decision support and treatment recommendation systems. The goal is to provide practical guidance for optimizing LLM-driven bias detection workflows and advancing fairness in healthcare AI.

Introduction

The adoption of LLMs in healthcare bias detection must address a core challenge: preventing models from perpetuating racial disparities. We propose two strategies:

- Metric-Driven: Standardized benchmarks to quantify racial bias in clinical LLMs (e.g., measuring diagnostic discrepancies across ethnic groups
- Model-Centric: Utilize pre-validated LLMs with proven fairness in healthcare tasks (e.g., equitable treatment recommendations).

These approaches offer flexibility–researchers can build new evaluation systems or adopt existing fair models. Detection capabilities are further enhanced through human-AI co-auditing workflows and hybrid architectures (LLMs + causal models), tested in scenarios like triage prioritization and drug efficacy analysis.


Bias in GPT-4
Causes
Training Data
Over-exaggeration of Real Associations
Stereotypical Associations
Reinforcement Learning with Human Feedback
Manifestations
Differential diagnoses
Treatment recommendations
Patient perception

Literature Review

Racial bias in healthcare–particularly in maternal health–is systemic, evidenced by Black women's maternal mortality rate being triple that of White women in the U.S., persisting even after adjusting for socioeconomic factors[1]. Structural racism and implicit biases drive this disparity, demanding urgent systemic reforms.

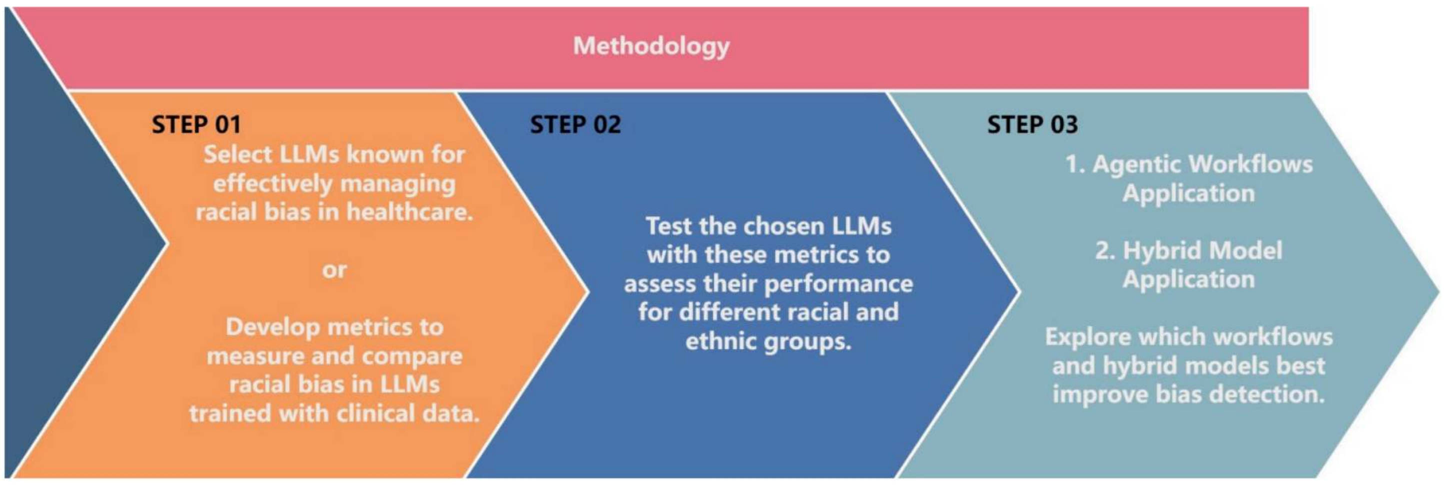
On the other hand, GPT-4's healthcare applications risk perpetuating bias: its training data overrepresents White patients and stereotypes minorities (e.g., 97% of sarcoidosis cases generated as Black patients; STDs disproportionately linked to minority men). Reliance on human feedback (RLHF) introduces subjectivity, while closed-source models hinder bias correction. Addressing these issues is critical for health equity and racial justice[2].

Research Methodology

In this study, we will adopt one of two optional approaches to ensure that the large language models (LLMs) used for healthcare-related tasks have minimal inherent racial bias:

- one is selecting LLMs that have already demonstrated fairness in healthcare settings as the baseline, and
- The other is developing standardized metrics to assess and compare the levels of racial bias in LLMs trained on clinical and demographic data.

After choosing one of these methods, we will further explore the ability of these models to detect and reduce racial bias in actual medical decision support systems through agentic workflows and hybrid models.


Methodology
STEP 01
Select LLMs known for effectively managing racial bias in healthcare.
or
Develop metrics to measure and compare racial bias in LLMs trained with clinical data.
STEP 02
Test the chosen LLMs with these metrics to assess their performance for different racial and ethnic groups.
STEP 03
1. Agentic Workflows Application
2. Hybrid Model Application
Explore which workflows and hybrid models best improve bias detection.

The metric-driven strategy is expected to provide a clear quantification of biases, facilitating targeted improvements, while the model-centric approach should demonstrate a baseline of fairness, streamlining further refinements. We expect hybrid models and agentic workflows to further improve the detection and mitigation of racial biases in practical scenarios, leading to more equitable healthcare outcomes.

Expected Result

We anticipate that the selected approach will significantly enhance the fairness of LLMs in healthcare applications.

Conclusion / Improvements

This proposal tackles the crucial issue of racial bias in healthcare LLMs, which can worsen disparities in medical outcomes. By setting up fairness metrics and using proven models, we aim to enhance algorithmic fairness. Despite these efforts, challenges like biased data and limited generalizability across different demographics could hinder progress. Addressing these issues will need collaborative efforts, clear model design, and continuous updates to meet clinical and ethical standards. This project seeks to transform LLMs from potential sources of bias into instruments for health equity.

References

[1] Montalmant, K. E., & Ettinger, A. K. (2023). The Racial Disparities in maternal mortality and impact of structural racism and implicit racial bias on pregnant Black Women: A Review of the literature. Journal of Racial and Ethnic Health Disparities, 11(6), 3658–3677. https://doi.org/10.1007/s40615-023-01816-x

[2] Zack, T., Lehman, E., Mirac Suzgun, Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. In The Lancet Digital Health (Vol. 6, pp. e12–e22) [Journal-article]. Elsevier Ltd. https://www.thelancet.com/digital-health

[3] Measuring implicit bias in explicitly unbiased large language models. (2024). [Journal-article]. https://arxiv.org/pdf/2402.04105

[4] Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth., 1(1). https://doi.org/10.1007/s44336-024-00009-2