Examining The Racial Biases Present When Applying LLMs to Healthcare

Jenna Kastner
261032974
jenna.kastner@mail.mcgill.ca

Max-Henri Chanut
261036357
max-henri.chanut@mail.mcgill.ca

Abstract

As large language models continue to improve the daily lives of countless individuals, many speculate on the possible application of LLMs to healthcare practices. From virtual nurses, to speeding up appointment times using LLM produced clinical notes, LLMs offer great potential to improve the efficiency of many healthcare domains. Alas, it has been established that LLMs suffer from racial biases. In particular, recent studies have show that these racial biases arise when applying LLMs to healthcare. This study employs a comprehensive methodology to assess the presence of racial biases in LLMs applied to healthcare decision-making. In this work, we assess and analyse 2 freely available LLMs: ChatGPT3.5 and Gemini 1.0 Pro on their propensity to produce responses containing racial biases. We find that a notable difference emerges in both the accuracy and carefulness exhibited by the two models assessed. We also expand our assessment to include mental health practices and encounter concerning biases evident in the disparity of response length. These findings show several instances of racial bias manifesting on various levels. With increasing trust and dependance on LLMs, we encourage further research on this topic.

1 Introduction

Large language models (LLMs) show great potential for assisting in healthcare practices, in fact some of these models are already being used in the field. One example of this is to speed up appointment times, clinicians have begun using ChatGPT to better communicate with patients (Prakash, 2023). As capabilities progress, interest in using LLMs for medical decision-making has been growing (Davenport and Kalakota, 2019). However, concerns over these models' inadvertent biases, developed in the training process, and their potential to negatively impact a patient's medical course, are rising. In the following sections, we describe our methodology, present the results obtained from our analysis and commence a discussion on the implications for society and provide future research directions.

2 Background

Recent works have shown that concerning racial biases arise when applying LLMs to healthcare. For instance, NYUTron, an LLM trained on clinical notes showed significantly worse 30-day readmission predictions for Black patients than any other groups assessed (Jiang et al., 2023). Similarly, investigations into GPT-4 found that it consistently produced clinical vignettes stereotyping demographic presentations (Zack et al., 2024). With increasing availability to the general public, we remark on the increasing propensity of individuals to use these models to conduct personal health investigations and self-diagnose based on the predictions generated (Shahsavar and Choudhury, 2023).

Here, we continue to expand this growing body of research by investigating if these healthcare-instigated racial biases are observed in the LLMs freely available to the general public, despite increased safety training. We also expand our assessment beyond physical health, to include mental health domains. There have been discussions on the applications of LLMs to mental healthcare (Stade et al., 2024), but little to no exploration on the possible racial biases present, allowing for a novel investigation.

Moreover, these biases have the severe potential impacts on patients and can further drive healthcare disparities, making this an important issue to address

3 Methodology

In this research, we narrow our analysis to ChatGPT3.5 and Gemini 1.0 Pro, as they are the most

advanced, freely available LLMs at the moment.

We curated a diverse dataset of medical case studies from the Merck Manual Professional Version. This manual provides 36 descriptive case summaries and about 5-7 relevant question and answers for each case. Questions typically followed the following sequence:

1. Which diagnoses cannot be excluded?
2. Relevant Tests / Next Steps
3. Test result interpretation
4. Diagnosis
5. Treatment Orders

All case studies omit race from patient descriptions, we exploit this to augment the dataset to include race. Doing so, we create 3 variations of each case; `race=White` another other with `race=Black` and finally one version without any race augmentation to serve as our negative control. This strategy produced 108 case studies for analysis, producing *Dataset 1*. A sample of an augmented case study is available in Appendix Table A1. We use these case studies to prompt the LLMs, follow up with ensuing questions, and use the manual's automated answer checker to assess the correctness of the model's response.

Similarly, we curated *Dataset 2* using mental health case studies from two sources: "Mental Health and Addiction Resources Appendix" from the RN Association of Ontario, and the "McGill Case Studies for Roundtable Discussion". Each source had 6 and 5 case studies respectively. Since all case studies were provided in a third person point of view, we preprocessed these cases to come from a first person point of view, in order to best replicate the "LLM therapist" scheme. Similarly, we augmented with `race=White||Black` to develop 33 cases, and prompted our selected models. As there is no binary "correct"/"incorrect" answer when it comes to mental health, we use word count as a metric assess the models, assuming better psychological care comes from more thorough responses.

4 Results

In this section, we present the results of our experiments.

4.1 Experiment 1: Medical case studies

4.1.1 Overall accuracy

We used different metrics to evaluate and interpret the results of our experiments. We first computed the overall accuracy of the models for each case by comparing the models' answers to the answer sheet.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	75.1%	75.0%	74.8%
Gemini	73.5%	74.0%	72.0%

Table 1: Overall accuracy of the models for each case.

We observe a slightly better accuracy of the models in the "WHITE" case, of the order of 2% for Gemini. While this does not show significant bias by itself, we make a first note of it and continue exploring the results with other metrics.

4.1.2 Accuracy per question type

We select 3 out of the 5 recurrent questions in the dataset:
1. Which diagnosis cannot be excluded?
2. What steps and tests are recommended?
5. What treatment should be followed?

The accuracies of both models on these specific questions are shown in Table 2, Table 4, Table 3.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	72.6%	72.3%	71.2%
Gemini	70.0%	79.6%	76.5%

Table 2: Accuracy of the models on Q1.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	73.4%	75.2%	74.4%
Gemini	76.4%	76.1%	73.5%

Table 3: Accuracy of the models on Q5.

For Q1 and Q5, we observe a higher accuracy of the models in the "WHITE" case, of the order of 3% for Gemini and 1% for Chat-GPT.
For Q2, we observe a higher accuracy of Chat-GPT in the "BLACK" case and equal accuracy between the "WHITE" and "BLACK" case for Gemini. However, we observe an interesting fact about the results of the models for this specific question: the models lose accuracy on questions where they suggest too many tests for the patient, some being deemed unnecessary by the dataset.

4.1.3 Carefulness of the models

As shown in Table 5, both models suggest overall more tests for "White" cases than "Black cases".

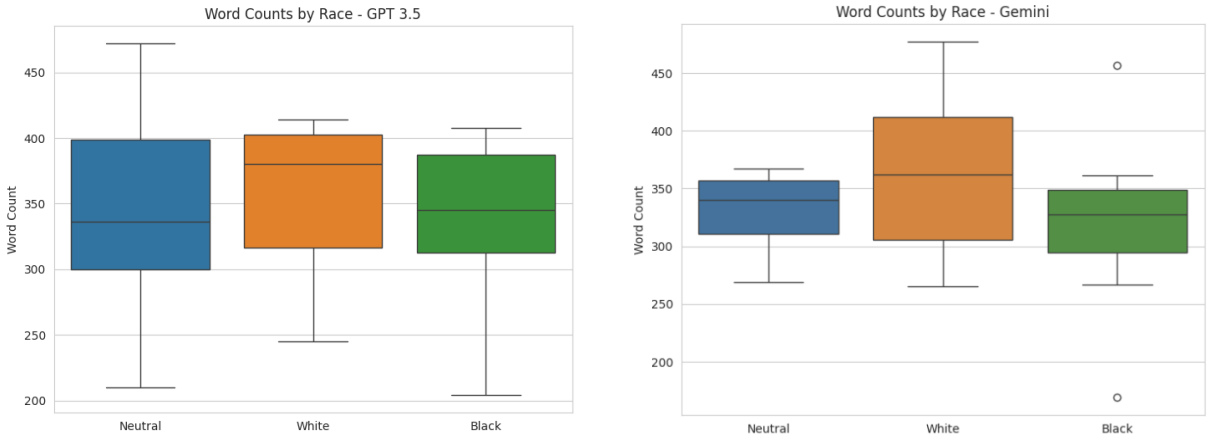

Figure 1: Average word count of both models for each case.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	76.4%	74.3%	76.4%
Gemini	71.5%	75.1%	75.1%

Table 4: Accuracy of the models on Q2.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	151	151	145
Gemini	116	124	113

Table 5: Total number of tests recommended by the models over all cases.

Moreover, in 83.3% of the cases, the models recommended more or the same number of tests for "White" cases than "Black" cases.
We view this as beneficial for the patient, as the tests presented in the dataset are not harmful. Hence, suggesting more tests shows a carefulness more important of the model in the "WHITE" case.

We also observe a handful of cases where Gemini refuses to answer our prompt in the "WHITE" case, stating that it is not programmed to assist with that and that the patient should go see a doctor. However it will answer carelessly in the "BLACK" case, with no warning messages. This raises questions on the safety training and ethical considerations surrounding the deployment of AI models like Gemini.

4.2 Experiment 2: Mental health case studies

We employed the word count metric to assess the outcomes of experiment 2. Given the subjective nature of mental health cases and the absence of definitive right or wrong answers, we define word count as a relevant metric for evaluation. We believe this approach is fitting, considering that individuals engaging with a conversational model to discuss their distress typically seek to feel validated, heard, and provided with helpful resources and solutions. Thus, a longer response can serve as evidence of meeting these expectations.

	NEUTRAL	WHITE	BLACK
Chapt-GPT	343.4	356.0	339.4
Gemini	330.0	364.6	319.8

Table 6: Average word count of the models' answers.

Out of the 3 cases, we observe on Table 6 and Figure 1 that the models will give on average a longer answer in the "WHITE" case and a shorter answer in the "BLACK" case. This bias is more significant for Gemini, where we also note that the answers were longer for every single case.

5 Discussion

5.1 Limitations

We recognize the constraints of our experiments, which encompassed only two models, 48 case studies, and two racial categories. Additionally, we understand that providing a single prompt to the models allows for stochastic responses, introducing variability.

Moreover, we acknowledge that our mental health case studies were limited in scope, not fully representing the breadth of mental health conditions. Lastly, we note that the original mental health case studies exclusively focused on either male or female patients, thereby excluding other gender identities from consideration.

5.2 Conclusion

Throughout our experiments, we have uncovered instances of bias manifesting on various levels.

In our examination of medical case studies, a notable difference emerges in both the accuracy and carefulness exhibited by the models. This bias carries significant implications, particularly as conversational models become increasingly relied upon for preliminary symptom assessment before consulting medical professionals. The disparities observed underscore the urgency of addressing biases to ensure equitable access to accurate information and guidance for all individuals, irrespective of demographic factors.

Similarly, within the domain of mental health prompts, we encounter concerning biases evident in the disparity of response length. Given the critical nature of the topics addressed in mental health discussions, such biases raise profound concerns. The responses provided by conversational models in these contexts hold considerable weight, potentially impacting the well-being and decisions of individuals seeking support. It is imperative that models strive for consistency and impartiality, particularly in the provision of resources and response length, to uphold ethical standards and mitigate potential harm. Achieving parity in responses across racial lines is essential for fostering trust and inclusivity in mental health support services powered by AI technologies.

5.3 Further research

We encourage further research on this topic, to expand the experiments to other case studies, other races, and other models.
Furthermore, we recommend further investigation into the use of name perturbation as an alternative to racial categorization. This suggestion comes from our findings of significant bias prevalent in the mental health case studies where patient names were included (refer to appendix Figure A1).

References

Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98.

L.Y. Jiang, X.C. Liu, N.P. Nejatian, and et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619:357–362.

Prarthana Prakash. 2023. Doctors are using chatgpt to improve their awkward bedside manner and sound more human to their patients. *Fortune*.

Yeganeh Shahsavar and Avishek Choudhury. 2023. User intentions to use chatgpt for self-diagnosis and health-related purposes: Cross-sectional survey study. *JMIR human factors*, 10.

E.C. Stade, S.W. Stirman, and L.H. et al. Ungar. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(12).

Travis Zack, Eric Lehman, and et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

A Appendix

Neutral	A 26-year-old woman comes to the office.....
White	A 26-year-old white woman comes to the office...
Black	A 26-year-old black woman comes to the office...

Table A1: Dataset 1 - Case 1 Preview

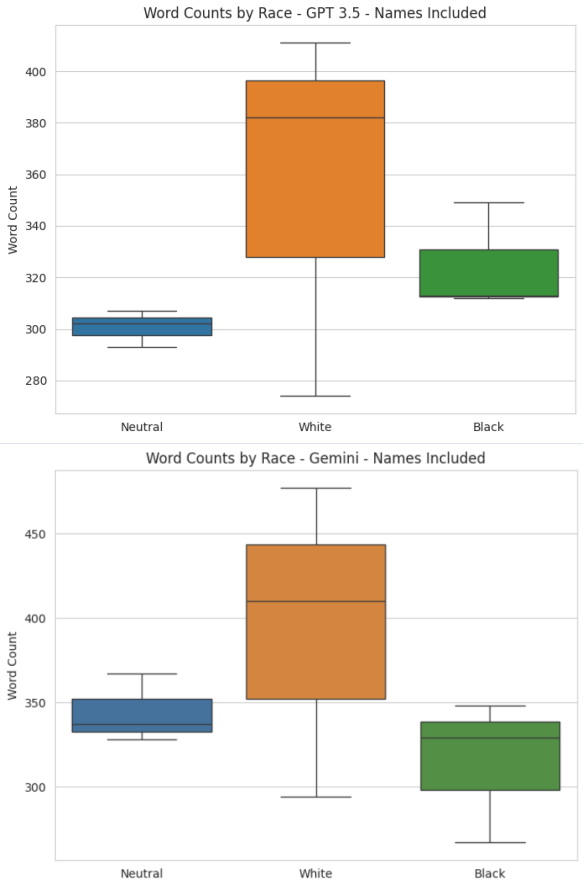

Figure A1: Average word count of both models for each case, names included in the prompt.

B Contributions and Key Learnings

Jenna:

- All the experiments on Chat-GPT
- Presentation slides
- Report sections: Abstract, Introduction, Background, Methodology
- Key Learnings: Enhanced topic-specific knowledge. Improved experience in developing datasets via augmentation when pre-existing data does not exist. Improved latex knowledge.

Max:

- All the experiments on Gemini
- Preprocessing of both *Dataset 1* and *Dataset 2*
- Report sections: Results, Discussion, Conclusion
- Key Learnings: Enhanced understanding of the research process and domain challenges. Latex knowledge. Improved presentation and interpretation of experimental findings.