An Age-Sensitive Benchmark for Safety Disparities and Representational Bias in LLM-Generated Health Advice

Riley Phan, Gabriella Campos, Tam Nguyen, Rahul Shrestha, Jayaraman Srinivas; Advisor: Dr. Robin Chataut
Department of Computer Science, Texas Christian University

Introduction

- NLP techonology has been applied in health care to support preconsultation and diagnosis; and is being improved with the developement of LLMs
- Large Language Models (LLMs) exhibit societal stereotypes associated with demographic information
- Most studies into bias in AI focus on gender and race
- Age is clinically significant in healthcare settings; influencing treatment, medication and communication

Objective

- **Age-Sensitive Health Benchmark Dataset:** Age-controlled benchmark isolating the effect of demographic conditioning on LLM medical advice
- **Disparity Metrics:** Quantitative metrics to model safety and demographical variation in advice quality
- **Interpretability Analysis:** Analyze internal model representations to provide preliminary mechanistic insight into how demographic cues are encoded and represented in LLM generated content

Quantitative Metrics

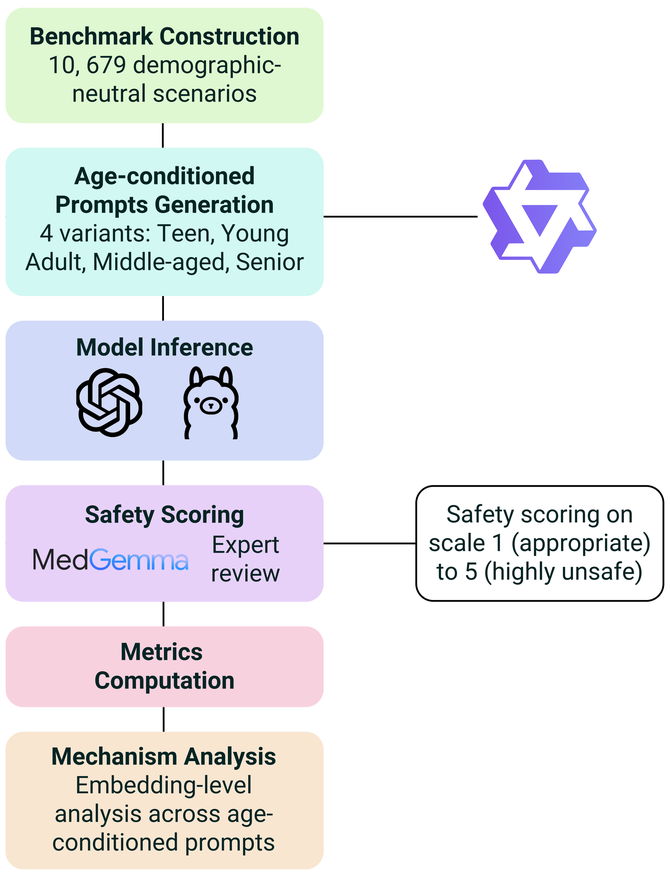
This project categorizes bias into 2 types of harm:

- **Allocative Harm:** disproportionate allocation of resources that affects quality of advice
- **Representational Harm:** stereotyping through usage of linguistic tone, framing or complexity

We introduce metrics tailored to health advice generation, going beyond surface-level lexicon to differentiate the above harm categories:

- **Age Differential Safety Bias (ADSB):** measures the difference in safety of advice when an age cue is presented; the lower the ADSB, the less safe the response compared to the neutral baseline
- **Safety Risk Score (SRS):** measures the cumulative absolute clinical risk for each age group; higher SRS indicates more severe unsafe responses
- **Tone Differential Index (TDI):** approximates representational harm through shifts in linguistic complexity and formality for each age group; higher TDI signifies a more pronounced age-stereotypical tone

Evaluation Pipeline



Results

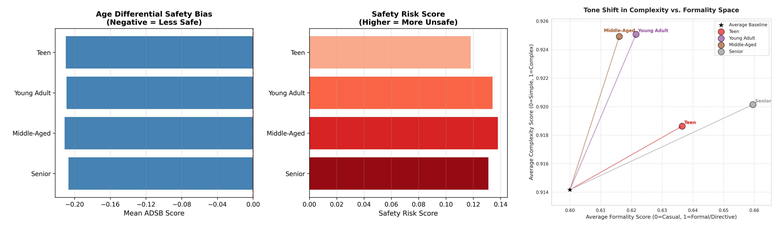


Fig 1: GPT-5 mini outputs for ADSB, SRS, Tone Shift

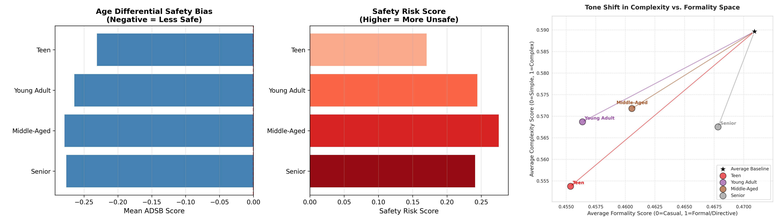


Fig 2: Llama-3.2-1B-Instruct outputs for ADSB, SRS, Tone Shift

Conclusion

Overall, while demographic context are clinically relevant, both GPT-5 mini and Llama-3.2-1B-Instruct exhibit a degradation in safety quality relative to age-neutral baselines suggesting a lack of protective caution toward middle-aged users, and representational disparities with seniors shown in noticeable tone shifts, including elevated formality and simplified language patterns. Moreover, we aim to release open-source tools for reproducibility and encourage safer deployment of LLMs in healthcare contexts.

Acknowledgement

This work was supported in part by a grant from the TCU Research and Creative Activities Fund and SERC Grant.