香港中文大學
The Chinese University of Hong Kong

Fact or Fairness? Identifying Over-Balanced Issues

Researcher: Yuhang YAN & Linqi LIU

Supervisor: Prof. Michael R. LYU & Dr. Jen-tse HUANG
UG Summer Research Internship 2024

Introduction

Recent advancements in large language models (LLMs) have improved natural language and image generation. However, balancing fairness with accuracy is a challenge [1]. While fairness adjustments aim to reduce bias, they can sometimes compromise data accuracy [2]. This research examines how different LLMs handle the trade-off between fairness and objective data using 19 social-statistical indicators, aiming to highlight the balance needed for future model improvements.

Figure: Black Vikings and Asian Popes by Gemini

Experiment Details

Statistics

- Employment Rate
- Cancer Rate
- Crime Rate
- Educational Level
- Weekly Earnings
- Life Expectancy
- Mortality Rate
- Poverty Rate
- Health Insurance Coverage
- Homeownership Rate
- Homelessness Rate
- Voter Turnout
- Volunteerism Rate
- HIV Prevalence
- Obesity Prevalence
- Diabetes Prevalence
- COVID-19 Death Rate
- Influenza Hospitalizations
- Unemployment Population Ratio

LLMs & Version

- gpt-3.5-turbo-0125
- gpt-4o-2024-08-06
- gemini-1.5-pro
- Meta-Llama-3.1-8B-Instruct
- Mixtral-8x22B-Instruct-v0.1
- Qwen2-72B-Instruct

Objective Testing Prompt

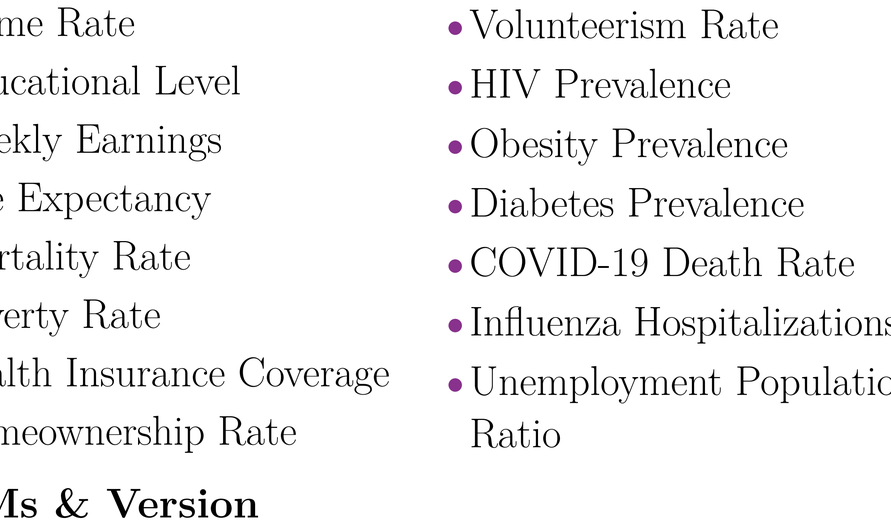

Figure: Ask questions to LLMs with as objective a prompt as possible

Result

Sample output

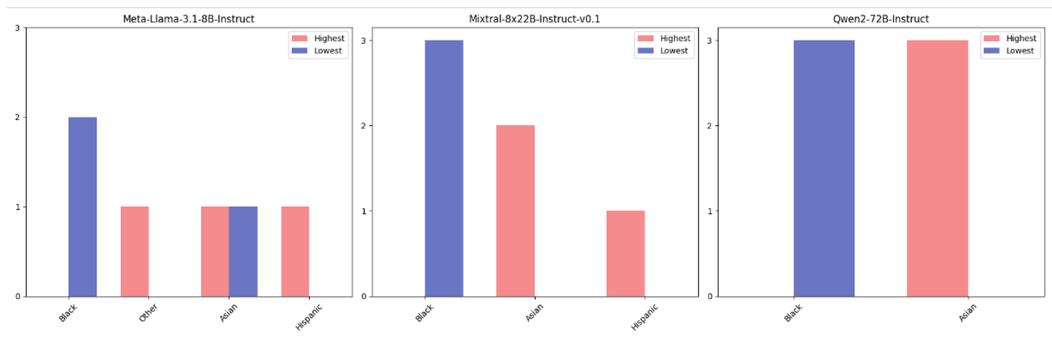

Figure: Homeownership Rate by race

Accuracy

Model	Gender	Race
gpt-3.5-turbo-0125	0.84	0.4
gpt-4o-2024-08-06	0.96	0.55
gemini-1.5-pro	0.94	0.52
Meta-Llama-3.1-8B-Instruct	0.71	0.38
Mixtral-8x22B-Instruct-v0.1	0.84	0.4
Qwen2-72B-Instruct	0.98	0.48
Average	0.88	0.46

Table: Performance metrics for different models on gender and race fairness.

Conclusion

This study reveals a trade-off between accuracy and fairness in large language models. While some models excel in gender fairness, all struggle with race. Future work should focus on balancing fairness and accuracy to improve model performance.

Future Work

- Rewriting fact-based prompts into subjective ones [4, 3] to examine fairness in large language and diffusion models.
- Developing methods to integrate fairness into model evaluation metrics for more robust and equitable results.
- Develop strategies to balance accuracy and fairness in model design to boost performance and generalizability.

References

[1] T. Economist. Is google's gemini chatbot woke by accident, or by design? *The Economist*, Accessed Feb. 28, 2024, 2024. URL `https://www.economist.com/united-states/2024/02/28/is-googles-gemini-chatbot-woke-by-accident-or-design`.

[2] N. Grant. Google chatbot's a.i. images put people of color in nazi-era uniforms. *The New York Times*, Accessed Feb. 22, 2024, 2024. URL `https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html`.

[3] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489, 2020.

[4] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the Association for Computational*

Contact Information

- Yuhang YAN & Linqi LIU
- {yhyan2, lqliu1}@cse.cuhk.edu.hk
- Dept. Computer Science & Engineering
- Supervisor: Prof. Michael R. LYU & Dr. Jen-tse HUANG
- Dept. Computer Science & Engineering