

Reducing Racial and Ethnic Bias in AI Models: A Comparative Analysis of Chatgpt and Google Bard

Tavishi Choudhary
Greenwich High, United States

Abstract

53% of U.S. adults recognize racial bias as an issue, 23% of Asian adults face cultural and ethnic biases, and 60% hide their cultural heritage and experience verbal abuse due to race & ethnic bias. AI models like ChatGPT or Google Bard trained on data with historical racial & ethnic biases can inadvertently amplify these stereotypes & biases. Through evidence-based analysis and auditing process, this project seeks to detect biased responses from AI models and propose a mitigation tool to reduce the propensity of AI models to generate racially and ethnically biased responses. The project starts with the creation of a large database containing racially and ethnically biased questions, terms, and phrases derived from thousands of documented legal cases by USEEOC, Wikipedia, and various surveys. Subsequently, ChatGPT & Google Bard were queried with these racially biased questions. The responses were scrutinized through sentiment analysis and human evaluation, which exhibited racial, ethnic bias and stereotypes. In comparison, I also observed measurable differences in how ChatGPT and Google Bard handled and responded to racial and ethnic inputs. I also researched how the ethnicity and race of the user who poses questions to ChatGPT or Google Bard influence the AI's understanding and responses and whether AI models can recognize and appropriately adjust to the context and intent of user's racial and ethnic background, After identifying racial & ethnic biases in AI model responses, I created 'BiasAudit,' a tool leveraging Excel, NLP, and ML plugins. It has a database rich in racial and ethnic questions, terms, and expressions and is designed for social science researchers & AI developers to prevent racial bias. This process aims to prevent new AI models from perpetuating biases. Additionally, 'BiasAudit' can audit AI responses to ensure they My research highlights the following five key findings, which are at the intersection of social sciences, AI ethics, and future policy development. First, it establishes that AI models, notably ChatGPT and Google Bard, inherently magnify racial and ethnic biases, reinforcing historical stereotypes and biases. Second, today's AI doesn't correctly incorporate the ethnic and racial context and intent of the user. Third, racial bias intensity is different across different AI models, directly correlated with their training data's historical biases. Fourth, there is a lack of a comprehensive racial and ethnic bias dataset and

policies documenting questions, prompts, and expected responses and behavior of AI models, which led me to the creation of the 'BiasAudit' tool for social science and AI researchers. Finally, this methodology and structure will not only help detect and address racial and ethnic biases in AI but also extend to other forms of biases in AI, such as political and health-related biases.

Keywords: racial bias, ethnicity, artificial intelligence, public policy, diversity