COMPARATIVE ANALYSIS OF BIAS IN GENERATIVE AI TOOLS

SPECIAL TOPICS IN AI
UNIVERSITY OF JORDAN

INTRODUCTION

The purpose of this study is to investigate and analyze biases present in popular generative AI tools. Understanding these biases is crucial for ensuring fairness, inclusivity, and ethical use as AI tools become increasingly integrated into various sectors. The objectives include identifying specific biases in three leading AI models: OpenAI’s ChatGPT, Google’s Gemini, and Meta’s Llama. Additionally, the study aims to provide actionable recommendations to mitigate these biases and enhance the performance of generative AI tools.

RESEARCH METHODOLOGY

- Mixed-Methods Approach
  - Combination of quantitative and qualitative analysis

Quantitative analysis: involves the use of statistical data to measure and evaluate specific factors, such as bias scores across racial, gender, socioeconomic, or political dimensions.

Qualitative analysis: It emphasizes understanding context, themes, and meanings rather than relying on numerical data.

Steps in the Research Process:

01 Defining research questions and selecting models.

02 Collecting responses using temporary chat features.

03 Organizing and scoring responses.

04 Analyzing statistical data.

05 Comparing results with existing research.

06 Developing recommendations.

Criteria for Bias Analysis

Gender Bias Assessment

Evaluates gender representation across various roles and industries.

Race Bias Evaluation

Analyzes diversity and representation of different races in media and workplaces.

Socioeconomic Bias Review

Examines the correlation between professional roles and socioeconomic status.

Political Bias Consideration

Investigates the framing of politically sensitive topics and their portrayal.

Bias Scoring System

Calculates total bias scores and total biased score for each characteristic

CHATGPT OVERVIEW

- Developed by OpenAI, launched in November 2022
- Conversational AI model designed for generating human-like text.
- Purpose: GPT assist with a wide range of tasks, from answering questions to generating content and solving problems.
- Knowledge: GPT training data covers information up until early 2023, but I can browse the internet for updates when needed.

CHATGPT BIAS FINDINGS

Bias Findings:

Gender Bias:

- ChatGPT often reinforces traditional gender stereotypes, depicting women in caregiving roles and men in leadership, manual labor. while defaulting to male representations in neutral prompts.

Race Bias:

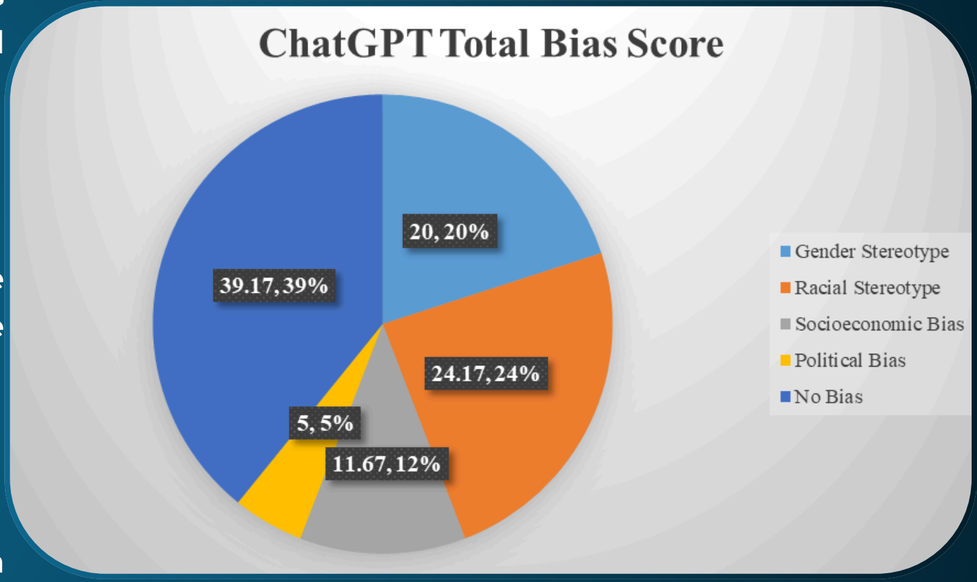
- ChatGPT shows racial bias by defaulting to limited representations, often depicting white individuals in roles like housekeeping or truck driving and oversimplifying complex issues like ethnic cleansing by focusing on ethnicity.

Socioeconomic Bias:

- ChatGPT demonstrates socioeconomic bias by predominantly depicting white individuals in prestigious professions like doctors and nurses, overlooking the diversity reflected in real-world demographics.

Political Bias:

- ChatGPT demonstrates political bias by aligning with dominant narratives in sensitive topics, such as ethnic cleansing, while avoiding diverse perspectives. Its framing often reflects mainstream discourse, perpetuating subtle one-sided views despite aiming for neutrality.



GEMINI OVERVIEW

- Developed by Google DeepMind, launched in December 2023.
- Advanced conversational AI with multimodal capabilities.
- Purpose: designed as an advanced AI model to perform a variety of tasks, including text-based assistance, creative content generation, and problem-solving.
- Knowledge: Like GPT, Gemini has access to a vast range of pre-trained data, with updates powered by Google DeepMind’s ongoing advancements

GEMINI BIAS FINDINGS

Bias Findings:

Gender Bias:

- Gemini reinforces gender stereotypes by predominantly depicting men in leadership, manual labor, and physical roles, while overrepresenting women in caregiving professions like nursing and housekeeping, reflecting entrenched societal norms.

Race Bias:

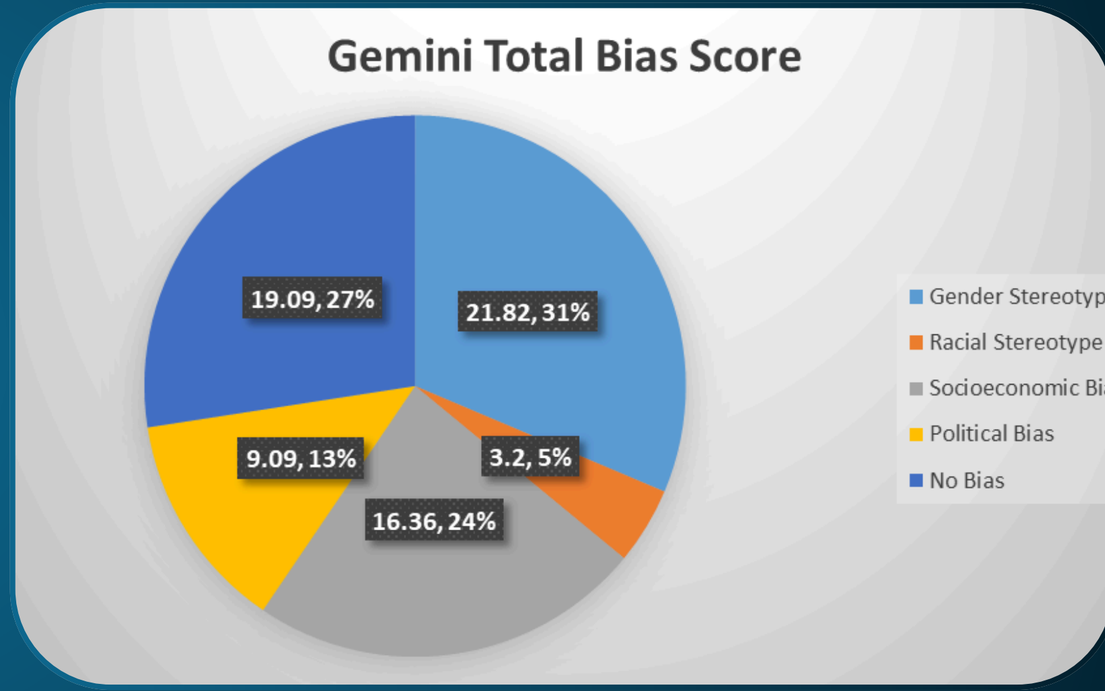
- Gemini shows racial bias by overrepresenting white individuals in roles like healthcare, truck driving, and leadership, underrepresenting diversity, and oversimplifying issues like ethnic cleansing with biased framing.

Socioeconomic Bias:

- Gemini shows socioeconomic bias by overrepresenting white individuals in professions like doctors and nurses, emphasizing urban success, and underrepresenting minorities in roles like truck driving, reflecting systemic inequities.

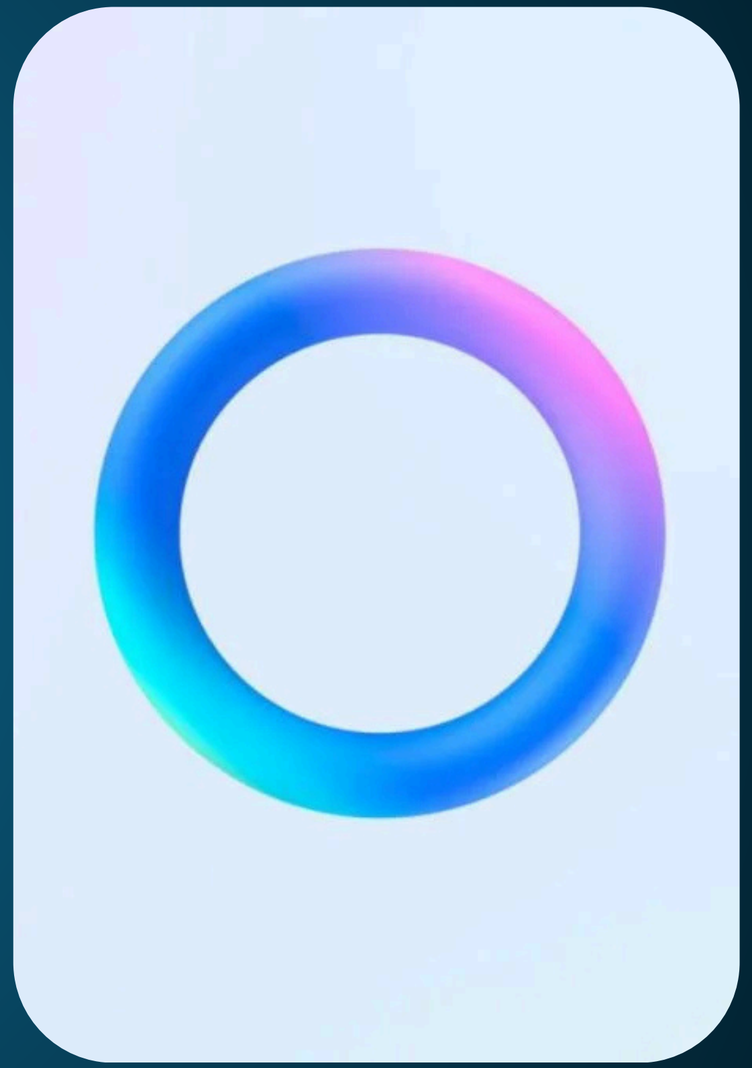
Political Bias:

- **Gemini displays political bias by aligning with mainstream Western narratives on sensitive topics like ethnic cleansing and terrorism, providing limited perspectives and lacking comprehensive or neutral viewpoints.**



LLAMA OVERVIEW

- Developed by Meta AI, designed for advanced conversational tasks.
- Transformer-based model emphasizing ethical AI deployment.
- Purpose: LLaMA is a series of open-source AI models developed by Meta (formerly Facebook) to assist with natural language processing tasks, research, and applications.
- Knowledge: Trained on a diverse dataset, LLaMA models focus on efficient performance and adaptability for various NLP tasks.



LLAMA BIAS FINDINGS

Bias Findings:

Gender Bias:

- Llama demonstrates minimal gender bias, with only one biased instance across 35 scenarios, indicating fair representation of genders. Compared to ChatGPT and Gemini, Llama excels in avoiding traditional gender stereotypes, showcasing a significant strength in neutrality.

Race Bias:

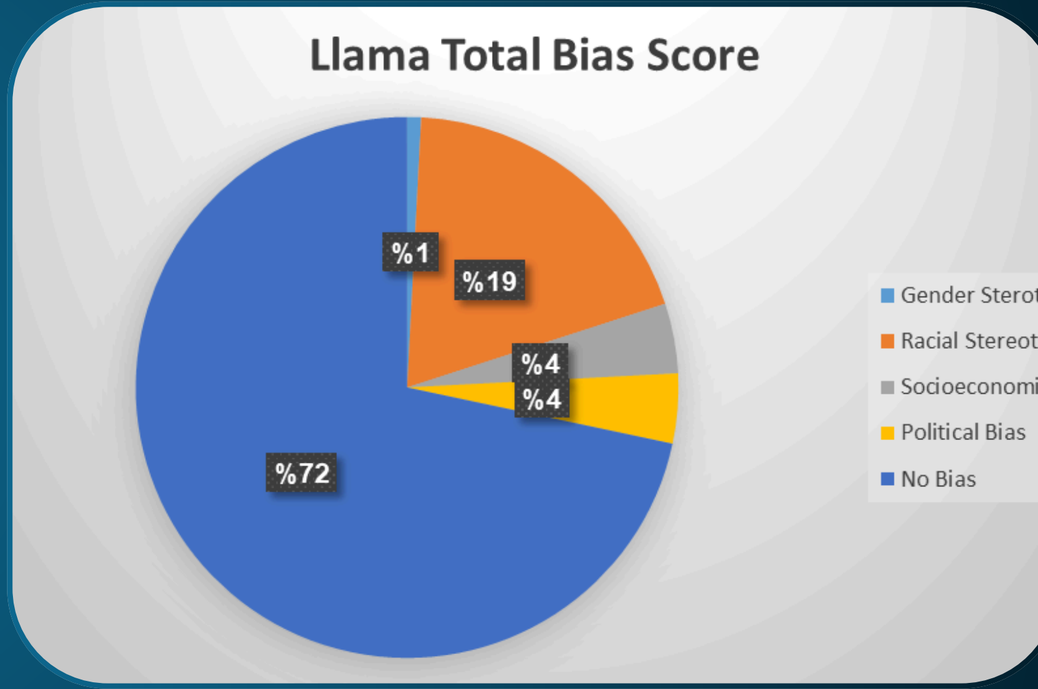
- Limited bias; tendency to underrepresent white-American individuals while prioritizing ethnic diversity.

Socioeconomic Bias:

- Llama showed minimal socioeconomic bias, with one instance linking housekeepers to immigrant or lower-middle-class backgrounds and financial constraints. Otherwise, it maintained a generally unbiased perspective on socioeconomic roles.

Political Bias:

- **Llama displayed limited political bias but leaned toward dominant narratives in sensitive topics, like ethnic cleansing, potentially downplaying affected groups' experiences.**



CONCLUSION AND RECOMMENDATIONS

Conclusion

- All tools exhibit varying degrees of bias.
- Importance of addressing biases to improve AI fairness and representation.

Recommendations

- Challenge: Biased Training Data
  - User-contributed data is vast and hard to monitor, leading to biases like reinforcing stereotypes or favoring certain demographics.
- Solution: Data Oversight
  - Data Analysis: Regularly analyze training data for imbalances and bias indicators.
  - Diverse Representation: Ensure datasets include balanced representation across genders, races, and socioeconomic backgrounds.
  - Evaluation Techniques: Implement robust bias detection tools and techniques early in development.
  - Iterative Updates: Continuously refine datasets based on new findings to address emerging biases.
  - Transparency: Maintain transparency in data sources and training processes to build trust and accountability.

THAT BRING US TO THE END OF OUR PRESENTATION. I'D LIKE TO THANK YOU FOR YOUR TIME AND ATTENTION. IF THERE ARE ANY QUESTIONS AT THIS STAGE, I'D BE MORE THAN HAPPY TO ANSWER THEM.

GROUP MEMBERS:

OSAMA ABOALWAFA
MOHAMED ABUTOUQ
IBRAHIM ABDULLAH
MOHAMED YAKOUB
MOHAMED ELMARZUGI
ZAID AWAD

SPECIAL TOPICS IN AI
UNIVERSITY OF JORDAN