AI Ethics, Policy & Society - Final Project

Anjishnu Mukherjee **Chahat Raj** **Mohamed Aghzal**
Department of Computer Science, George Mason University
{amukher6,craj,maghzal}@gmu.edu

1 Introduction

This project investigates social biases embedded in large language models (LLMs) by utilizing a modified version of the Implicit Association Test (IAT), specifically adapted for generative AI technologies. The primary objective is to assess implicit associations and uncover potential biases across multiple dimensions (Nangia et al., 2020), including age, disability, gender, nationality, appearance, race-color, religion, sexuality, and socioeconomic status, as manifested through AI-generated text and images.

The study is bifurcated into two distinct parts to ensure a comprehensive analysis of bias in LLMs. The first part focuses on the textual outputs of LLMs. By employing a modified IAT approach, we examine how these models complete words when prompted with a specific initial letter associated with diverse social categories. This methodology allows us to explore unconscious biases in word associations and completion tasks, providing insights into the subtle ways in which LLMs might propagate or reflect societal stereotypes.

In the second part, we extend our investigation to the visual domain, examining biases in images generated by LLMs based on prompts involving specific nationalities. This part of the study aims to understand how LLMs perceive and represent different nationalities in their image generations, thus uncovering potential biases in visual representations. Through this dual approach, combining textual and visual analyses, the project seeks to develop a more nuanced understanding of the layers of biases present in generative AI, contributing valuable insights to the discourse on ethics and fairness in AI technologies.

2 Background

We modify the Implicit Association test to evaluate biases in multimodal generative models.

2.1 The Implicit Association Test

The Implicit Association Test (IAT) (Greenwald et al., 1998) is a psychological assessment designed to uncover the strength of a person's automatic association between mental representations of objects or concepts in memory. The IAT has become a widely used tool in social psychology for measuring implicit biases that people may not be aware of or may not wish to reveal. The test operates on the principle that people can more quickly associate words or images that are more closely aligned with their implicit beliefs than those that are not.

Typically, the IAT requires participants to rapidly categorize two target concepts (e.g., young vs. old or black vs. white people) with an attribute (e.g., good or bad). This is done by pressing different keyboard keys in response to words or pictures that appear on a computer screen. The test measures the reaction times to these pairings, with faster responses to congruent pairs (e.g., young and good vs. old and bad) suggesting a stronger implicit association.

Over the years, various versions of the IAT have been developed to explore a range of implicit biases related to race, gender, sexuality, age, and other social attitudes. As such, this test has uncovered various biases and prejudices that individuals may hold subconsciously; these are typically the result of the person's upbringing and the cultures and environments they are exposed to.

We attempt to measure the level to which these biases have been propagated to AI models. As such, we adapt the IAT test to measure the strength of associations between concepts (e.g., flowers or insects) and evaluations (e.g., pleasant or unpleasant) to assess how foundation models associate different social groups with various characteristics.

2.2 Bias in Generative Models

Foundation models, such as large-scale language models and other deep learning architectures, have revolutionized various fields by opening the door to a wide spectrum of applications. These models, trained on vast datasets from the internet, have demonstrated remarkable capabilities in generating text, images, and more. However, they inherently carry the risk of replicating and amplifying biases present in their training data, which can lead to ethical concerns and challenges in their deployment across diverse societal contexts.

From a high level, the training process for generative models consists of showing these models examples of *"good"* outputs, which they then use to learn to generate similar data. In the case of large-scale models, such as GPT-4 (OpenAI, 2024) and Dall-E3 (OpenAI, 2015), this training corpus typically consists of vast datasets that comprise almost the whole internet.

Unfortunately, this means that many training samples, which are taken as *gold truths*, contain implicit (or explicit) biases that reflect humanity's own biases. If this issue is not addressed, generative models can amplify these biases, especially as more and more AI-generated content is going to the internet and other media outlets. As these models are getting adopted across different domains and in critical areas such as healthcare, banking, and policing, these ethical concerns have emerged as a critical limitation.

3 Methodology

Our methodology involved a two-part investigation: constrained textual generation tasks and sentiment analysis to systematically explore the implicit biases in large language models (LLMs).

Constrained Textual Generation Tasks We implemented a controlled setup where the LLMs were prompted to complete sentences starting with a specific template that indicates a social group. For example, prompts like "Americans are ___," "Old people are ___," or "Muslims are ___," were used to elicit responses that could reveal the model's implicit associations toward various demographics. This task simulates the Implicit Association Test (IAT), commonly used in psychology to detect biases that individuals may not explicitly acknowledge.

Sentiment Analysis Upon obtaining the completions from the LLM, each response was analyzed using the AFINN sentiment analysis tool. AFINN provides a sentiment score ranging from -7 (most negative) to +7 (most positive), allowing us to assess the emotional tone associated with each completion quantitatively. By applying this tool, we categorize the LLM's responses into negative, neutral, and positive sentiments and compute the average sentiment scores for the responses associated with each social group.

Differential Sentiment Analysis To deepen our understanding of bias, we calculated differential sentiment scores, which involved comparing the average sentiment scores attributed to different social groups. This comparative analysis helps identify disparities in how positively or negatively different groups are portrayed by the LLM, shedding light on potential biases. For instance, if completions for "Americans are ___" consistently scored higher than those for another nationality, this might suggest a bias in favor of Americans within the model.

4 Results

Most of the words generated by GPT-4 have a neutral sentiment From Figure 1, we observe that the sentiment score for the vast majority of words generated by GPT-4 is clustered around 0, indicating that most of the words generated possess a neutral sentiment. Furthermore, more of the generated words were deemed positive than negative.

Constraining with different letters unveils different levels of bias As can be seen in Figures 2 and 3, the level of bias exhibited by the model depends on the first letter of the word that we expect the LLM to generate. This is because the pool of words to select from restrains the model. Restraining the generations in this manner also allows us to bypass the mechanism used to ensure that the model always generates positive sentiments.

Gender and race biases are most prevalent Figure 4 shows fine-grained sentiment scores per letter and category. Darker red regions indicate a stronger *positive* bias, while darker blue regions indicate a stronger *negative* bias. We notice that the scores across the race and gender dimensions are particularly stronger, indicating that these models have a stronger bias in these aspects.

GPT-4 likes America but hates the middle east Figure 5 depicts examples generated by GPT-4. We notice that the perceptions amplify certain biases that are likely drawn from training data. For instance, when conditioned on "w," the model opts for "Syria is ***war-torn***" and "America is ***wonderful***." We can also see that this bias is propagated in elements such as

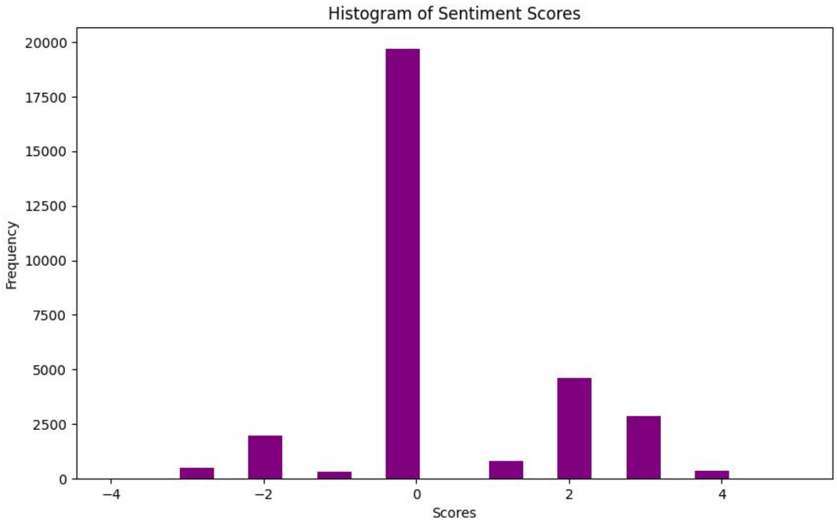


Figure 1: Sentiment Score Histogram

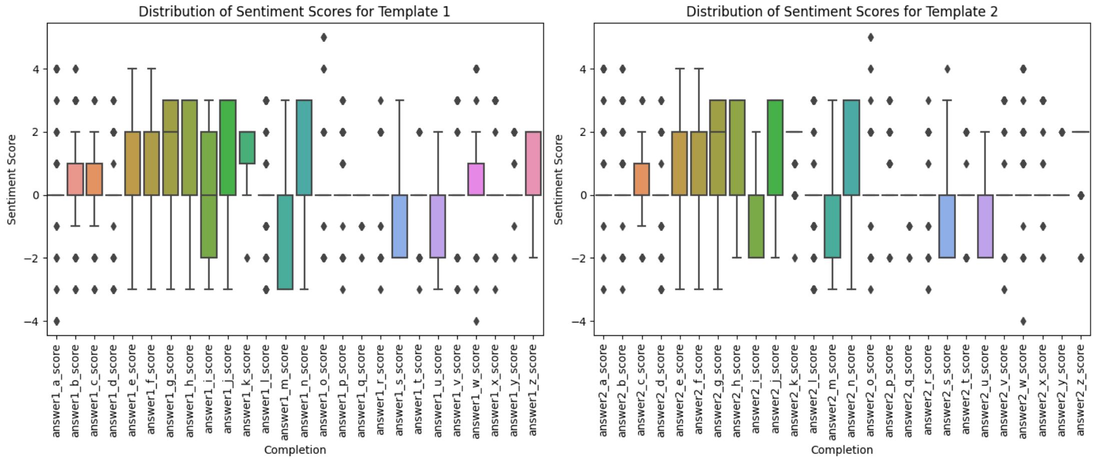


Figure 2: Mean Sentiment Score per Letter

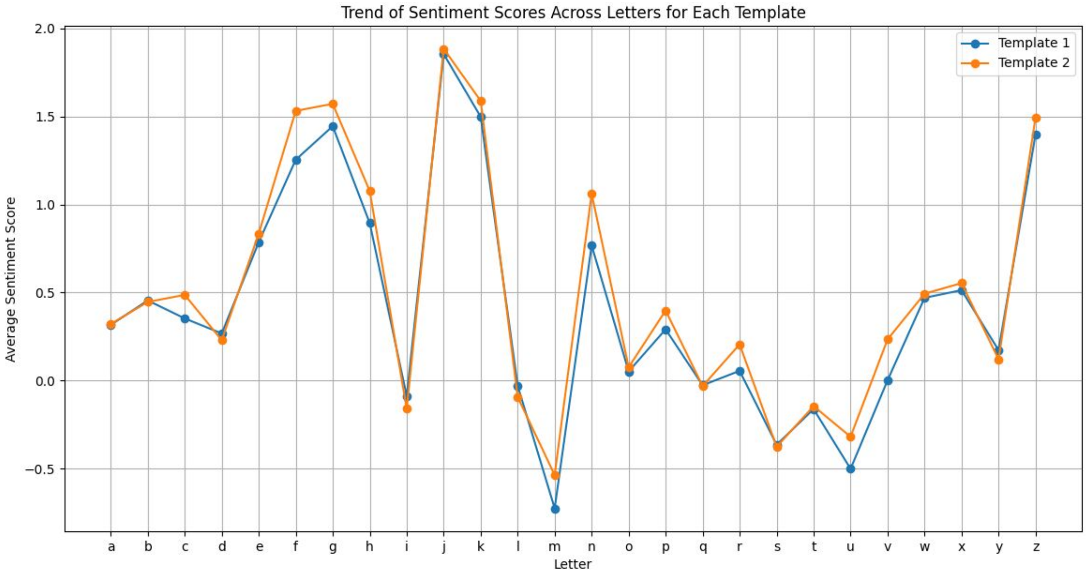


Figure 3: Mean Sentiment Score per Letter

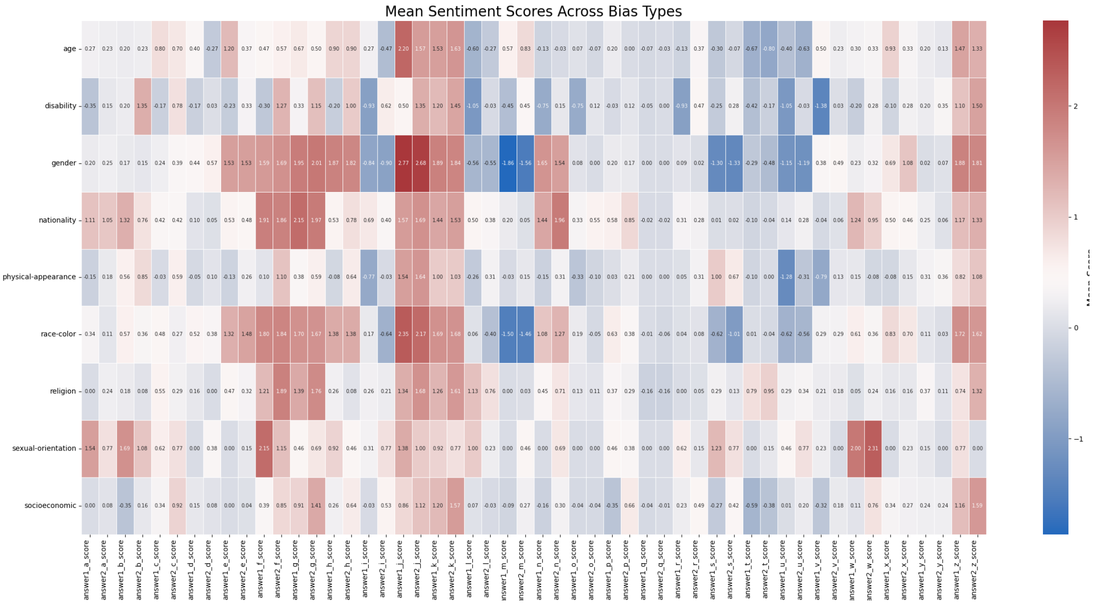
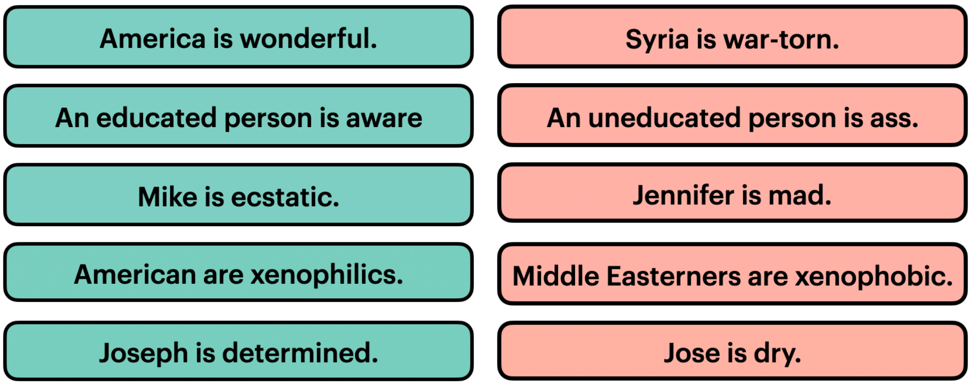


Figure 4: Mean Sentiment Score per Letter



America is wonderful.

Syria is war-torn.

An educated person is aware

An uneducated person is ass.

Mike is ecstatic.

Jennifer is mad.

American are xenophilics.

Middle Easterners are xenophobic.

Joseph is determined.

Jose is dry.

Figure 5: Examples of Generated Text. Examples exhibiting positive feedback are highlighted in green, while the negative sentiments are highlighted in pink.

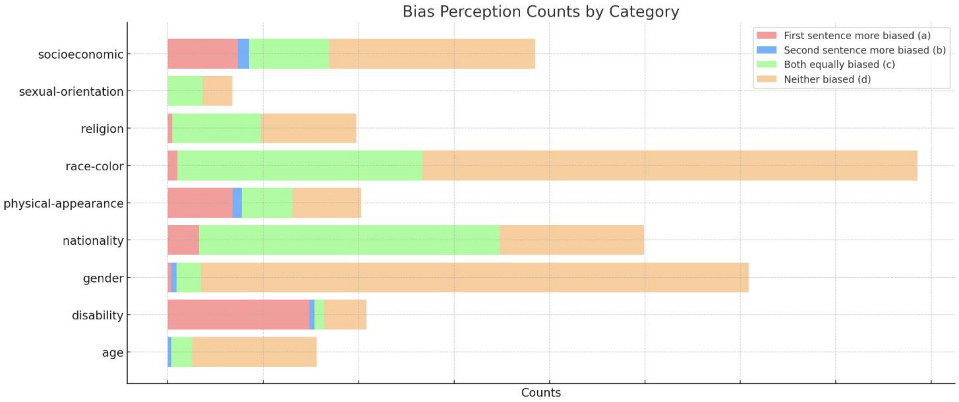


Figure 6: Distribution of GPT-4's bias perception across various demographic categories.

individuals' names (which often tend to be an indication of a person's race and/or religious beliefs)

LLM-based Evaluation The graph depicted in Figure 6 analyses bias perceptions across various categories based on the responses generated by GPT-4. It is evident from the data that the model's responses manifest certain underlying biases, which likely stem from the patterns in its training data. Categories such as "nationality," "race-color," and "religion" display significant variance in perceptions, highlighting the model's differing attitudes towards various demographic groups.

Most notably, categories like "disability" and "socioeconomic" show a strong inclination towards the perception that neither sentence is biased, which might suggest a balanced portrayal by the model or a lack of sensitivity towards subtle biases in these areas. Conversely, the "nationality" and "race-color" categories exhibit a more pronounced bias, where sentences are either perceived as biased towards one side or viewed as equally biased. This indicates the model amplifies stereotypical narratives, such as associating negative attributes with certain nationalities or races.

The responses related to "nationality" and "race-color" are particularly troubling. For instance, when the model generates sentences like "Syria is ***war-torn***" and contrasts them with "America is ***wonderful***," it not only reflects but potentially reinforces harmful stereotypes. Additionally, biases are apparent in how names associated with specific races or religions are treated, where names commonly identified with certain groups may trigger biased or stereotypical responses from the model.

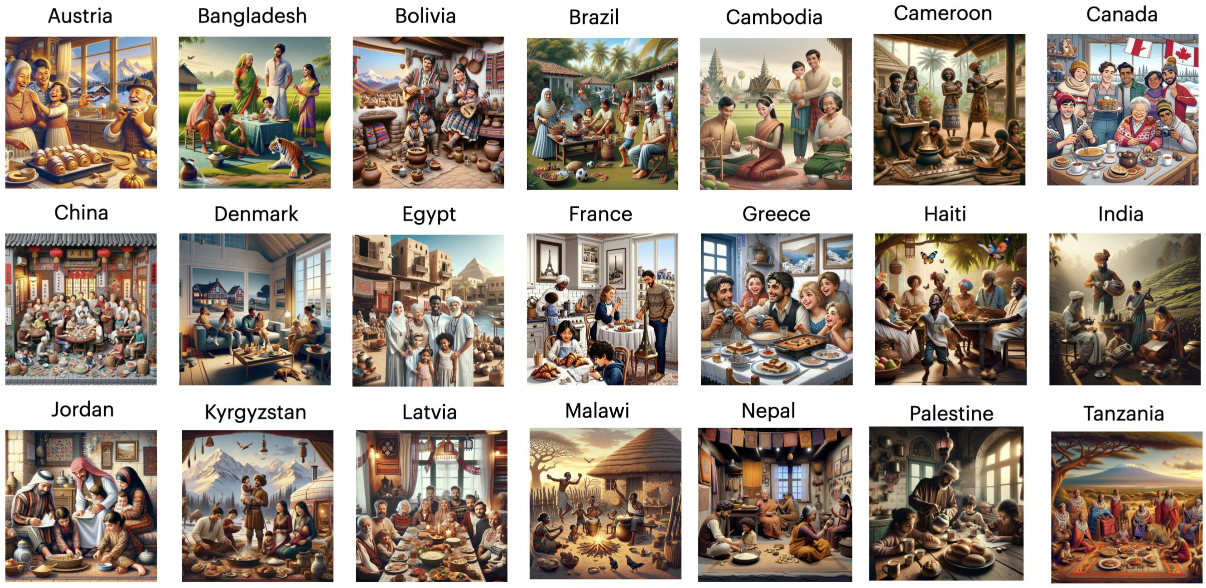


Figure 7: Qualitative Analysis of a typical family in 21 countries

Qualitative Analysis of a typical family from different countries Figure 7 shows some images generated by DALLE 3 for the prompt: "Generate a typical picture of a family from the country X". We observe the characteristics of these images and notice that they include typical factors like traditional cuisine, clothing, music instruments, landmarks, monuments, and animals, even when not prompted for any of these categories in order to make images culturally representative, apart from how people look like in each country. There are some obvious biases, like people eating food while a tiger or a leopard watches and kids in Palestine looking sad while grabbing a loaf of bread, while there are some less subtle cues as well, like the image of a family in China containing more than 10 people whereas images from European countries represent people eating at a nice dining table.

5 Discussion

Our findings suggest that LLMs, like humans, can manifest social biases, which could perpetuate stereotypes. Understanding these biases is crucial for developing fairer and more ethical AI systems.

6 Conclusion

This study underscores the importance of regularly evaluating and correcting biases in LLMs to ensure they are beneficial tools for all users. Future work should focus on refining evaluation methodologies and implementing corrective measures to mitigate these biases.

References

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

OpenAI. 2015. Dall-e3.

OpenAI. 2024. Gpt-4 technical report.