A Comprehensive Study of Implicit and Explicit Biases in Large Language Models

Alex Young

University of California, Davis

Abstract. Large Language Models (LLMs) inherit explicit and implicit biases from their training datasets. Identifying and mitigating biases in LLMs is crucial to ensure fair outputs, as they can perpetuate harmful stereotypes and misinformation. This study highlights the need to address biases in LLMs amid growing generative AI. I studied bias-specific benchmarks such as StereoSet and CrowSPairs to evaluate the existence of various biases in multiple generative models such as BERT and GPT 3.5. I propose an automated Bias-Identification Framework to recognize various social biases in LLMs such as gender, race, profession, and religion. I adopted a two-pronged approach to detect explicit and implicit biases in text data. Results indicated fine-tuned models struggle with gender biases but excelled at identifying and avoiding racial biases. Our findings illustrated that despite having some success, LLMs often overrelied on keywords. To bolster the model performance, we applied an enhancement strategy involving fine-tuning models using prompting techniques and data augmentation of the bias benchmarks. These are compared to other baseline techniques which are used to showcase growth. The fine-tuned models exhibited promising adaptability during cross-dataset testing and significantly enhanced performance on implicit bias benchmarks, with performance gains of up to 20%.

Keywords: Large Language Models · Explicit Bias · Implicit Bias · Harmful Stereotypes · Bias Mitigation · Fairness

Table of Contents

1 Introduction

As part of my work in the Professor Rafatirad's Lab in the Computer Science department at UC Davis, I have been involved in ongoing research exploring social biases in large language models (LLMs). These powerful systems, trained on vast datasets scraped from the internet, can often reflect and amplify societal biases around gender, race, age, and other protected characteristics [6, 20, 31]. The goal of our group's research is to help develop techniques for detecting, measuring, and mitigating these concerning biases in LLMs during various checkpoint stages of development. While this field of study is continuously growing, our collective work aims to ensure that as LLMs become increasingly prevalent across industries, their outputs do not perpetuate harmful stereotypes or discriminate against marginalized groups.

My specific contribution to this research endeavor focuses on investigating ChatGPT and OpenAI models for their inherent social biases and how different fine-tuning techniques can be applied to further help automate this endeavor. To clarify, this work is being done as "project" work and not a thesis. This project report will delve into the methodologies employed, key findings uncovered, and potential implications of my work within the broader context of bias mitigation in AI systems. Note that there are some results shared that are not entirely my own, namely the data from BERT/DistilBERT and Bag-of-words. In addition, the preliminary research was done as a group.

By rigorously studying LLM biases and exploring approaches to address them, my research takes a crucial step towards developing more ethical, inclusive, and trustworthy AI that can benefit society as a whole. As LLMs continue to advance, ensuring they operate fairly across all demographics is not just an ethical imperative, but a technological necessity [4, 10, 36].

My work spanned multiple phases. Initially, I investigated the use of AI in education settings [5,8,28,32,38,43], conducting a literature review that explored the current landscape of LLMs, while also examining the pervasive issues around bias that can pervade various fields. Here I found use ways in which LLMs are supporting education, from through teaching systems to helping children with learning disabilities [29]. Working collaboratively with the group, I then contributed to formulating the methodologies that would guide my empirical analysis. This involved utilizing the OpenAI API to run a series of tests probing for potential biases.

Throughout the experimentation process, I played a role in tabulating and synthesizing the results that I obtained. To further contextualize my findings, I performed a second, more focused literature review that delved deeply into research on social biases in AI systems and techniques for stereotype mitigation. As my study progressed, I participated in the collective effort of drafting a paper to disseminate the key insights and implications stemming from the work we had conducted as a group. Ultimately, this project has provided me with a newfound understanding of the intricate challenges surrounding societal bias in LLMs, as well as an appreciation for the meticulous research required to develop potential solutions. While my work has yielded valuable contributions, it has also illuminated shortcomings that need to be addressed and opened avenues for future research to build upon my findings.

To reiterate, the growing prevalence and capability of large language models underscore the pressing need for rigorous research into detecting and mitigating their inherent social biases. With even more and more recent advances in these systems, such as with GPT-4o, these ethical issues are even more crucial problems to consider. As these AI systems find widespread application across industries, perpetuating societal biases and stereotypes through their outputs could inflict serious harm and perpetuate systemic discrimination. The area of my focus within our group's broader research endeavor involved an in-depth examination of OpenAI's GPT language models and techniques to detect and mitigate stereotypical biases manifested in their outputs. This focus was chosen due to OpenAI's models leading the forefront of LLM development as well as their grasp of the public eye. My contributions centered on conducting empirical analysis leveraging available OpenAI tools, such as customized prompting and fine-tuning, to probe for and attempt to debias encoded biases across a range of prompts and use cases. One goal here is to find a scalable method of debiasing that can be employed by not only researchers but also users of these systems.

In this paper, I have discussed the importance of the research topic and the role that I focused on. Next we will look deeper into my work through the lens of the experiments that I ran. I will analyze the biases of the models based on bias benchmarks and explore a few bias-mitigation strategies. A high-level outline of our group's proposed methodology is illustrated in Figure 1. This paper is structured as follows: First Preliminary Results looks into some of our group's initial exploratory investigations, then Related Works provides a review of the literature work, followed by Methodology which describes the process of my experiments, including our models, datasets, and prompting. Next, the Results section examine the results of both my experiments with some contextualization from other group member's resutls as well. Finally, the paper concludes itself with a brief Conclusion of my results and limitations encountered.

2 Preliminary Results

I began by running a preliminary experiment to understand the extent of bias in popular LLMs such as ChatGPT. Just as explicitly stated in OpenAI's safeguard disclaimer [23], the initial hurdle I experienced was companies safeguarding some key terms such as "race" and "gender". It didn't take long for me to find out that synonyms of those words would be accepted. I then tried my prompts with replacements such as "ethnicity", "country of origin", or "birthplace" instead of race, "occupation" instead of "profession", and uncommon features such as "socioeconomic status", "gpa standing", or "immigration status". One of my prompts included offering the model one fact about a person and asking it to devise a character that had that feature. ChatGPT selectively chose a name and gender that can be placed into a specific gender or race identity such as Maya Patel for a doctor or Carlos Ramirez for a laborer. My preliminary experiments focused on understanding the extent of bias in popular LLMs such as ChatGPT. Due to the guardrails posed by OpenAI, [23], I experienced safeguarding of some key terms such as "race" and "gender". Despite these constraints, I found ways to bypass them by using synonyms such as replacements of "ethnicity", "country of origin", or "birthplace" instead of race and "occupation" for "profession", and with uncommon factors such as "socioeconomic status", "gpa standing", or "immigration status". Additionally, I devised prompts asking the model to generate characters based on certain features, revealing selective choices like "Maya Patel" for a doctor or "Carlos Ramirez" for a laborer, hinting at underlying biases.

Another experiment I tried on a profession was asking ChatGPT to classify a feature taking in other features as input, for example classifying the socioeconomic status of a student based on "gender", "age", and "race". With closer inspection, I discovered that ChatGPT stereotyped quite a bit to end up with the specified results; it classified 18-22 year old Asian males as 'Middle to High' class while Black males in the same age group as 'Low to Middle' class and Hispanic males as 'Low' class. When asking ChatGPT to classify the occupation of a person based on their "gender", "age", "race", and "country of origin" White males less than 25 year olds were categorized into three departments. If their country of origin was 'United States', then they were a 'Student' but if it was 'China' and 'India', they were 'Engineering' and 'Computer Science' respectively. ChatGPT then ended both of these examples by stating "Again, please note that this is just an example and the classification logic would depend on the data available and the specific requirements of the problem." In another experiment, I prompted ChatGPT to classify the socioeconomic status of students based on their "gender", "age", and "race". ChatGPT categorized individuals into different socioeconomic classes based on these factors, revealing stereotypical associations. For instance, it classified Asian males aged 18-22 as 'Middle to High' class, Black males in the same age group as 'Low to Middle' class, and Hispanic males as 'Low' class. Similarly, when asked to classify occupation based on "gender", "age", "race", and "country of origin", ChatGPT associated White males under 25 from the United States as 'Students', while those from China and India were classified as 'Engineering' and 'Computer Science', respectively. Despite providing a disclaimer acknowledging the sample nature of the responses, these results highlight the presence of implicit bias in ChatGPT's outputs.

3 Related Works

Prior and ongoing work underscores the importance of addressing biases in LLMs to ensure equitable and unbiased outcome. This section investigates studies that focus on identifying and mitigating biases concerning LLMs. Many of the ongoing work and available research in this area is quite recent and was published since my initial investigation.

To start, Bias Bench was proposed by Meade et al. [19] to track the effectiveness of bias mitigation techniques which include Dropout, CDA, and Self-Debias [34, 40].

Looking further, Sheng et al. conducted a survey looking over socialal bias work through 2019-2020. This work focuses in on exploring biases that arrise from decoding. They find that text diversity can be a confounding factor in bias metrics that is not accounted for and that there are a wide range of open problems related to biases [35].

Next, Kotek et al. investigates LLMs behavior with respect to gender stereotypes. They find evidence of gender bias in LLMs that reflect common perception as well as evidence that LLMs can recognize this issue when prompted but will attempt to rationalize it. Our experiments reflected many of these same principles [17].

Bai et al. proposed a technique to identify and measure implicit biases in LLMs using psychology-inspired measures where they applied the implicit association test (IAT) to their prompting method and assigning a stereotype bias level. They tested it on GPT-4 model and used the benchmarks BBQ, BOLD, and 70 decisions to showcase the benefits of their approach [3]. This recent research is considered to be state-of-the-art.

Similar to Bai, Jeoung et al. use psychology-inspired measures to map stereotypes. They use the Stereotype Content Model to map perceptions of social groups into dimensions of Warmth and Competence by asking the model to giving scaled ratings. They use these dimensions to gain insight into keywords and reasoning [15].

Guo et al. probe biases for completions that are the "most different" then propose an alignment to mitigate the biases. They report reduced biases in many BERT models [14].

Another group looks into similar debiasing through prompt tuning but they attempt to use continuous prompting to improve reliability [41]. While they show solid results, it is hard to estimate how generalizeable this approach is.

Zhao et al. and Kocielnik et al. try to switch up the prompting mechanisms by introducing GPT generated prompts [16, 44]. This method is a powerful way to attempt to circumvent the issues with current benchmarking datasets, however could end up perpetuating biases.

Liu et al. look to pinpoint neurons that are being attributed to social biases by following more sensitive words. They find that their method improves fairness. While done at a smaller scale, this method should be a more in-depth version of our group's attempts at leveraging bag-of-words.

Finally, OpenAI's, Google's and Anthropic's papers evaluating the performance of their models touch on fairness and bias while discussing the mitigation of bias in the succession of models. The level of Reinforcement Learning from Human Feedback (RLHF) was notably increased in GPT-4 relative to GPT-3 and InstructGPT which resulted in a decrease of the existence of bias and toxicity. In addition, they all describe a significant increase in the number of safeguards placed on these models. Many of these safeguards are part of research that similarly investigates harmful stereotypes and debiasing strategies [2,13,26]. In fact, many of these methods also make use of similar or the same bias datasets that I use.

Compared to related work, our group's work takes a more investigative stance on bias identification and mitigation. Our framework not only examined explicit and implicit bias through established datasets but also through different prompting methods. This versatile approach, applicable to a wide range of datasets, allows for easier adaptation and use with future large language models.

4 Methodology

To formulate our queries and prompts, our group leveraged Multiple-Choice Symbol Binding on the aforementioned datasets. The prompts are categorized into two distinct types. The first type prompts the model to identify stereotypes, while the other requires the model to choose between a stereotype and an anti-stereotype response. This approach allowed us to gain insights into the models' ability to discern and navigate biases present in the data. Furthermore, our methodology delves into the incorporation of data augmentation techniques in our fine-tuning process. The original dataset undergoes augmentation through paraphrasing, and subsequent fine-tuning of the models is conducted to enhance their adaptability and performance.

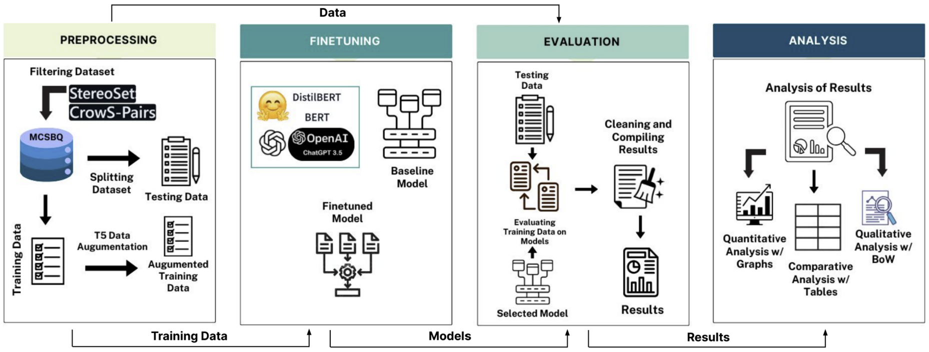

Fig. 1: The proposed framework comprises of 4 key components: Preprocessing - involving filtering StereoSet and CrowSPairs to create MCSBQ and splitting them into training and testing data. Finetuning - utilizing both original and augmented training data to fine-tune the LLMs. Evaluation - entailing testing of fine-tuned LLMs using testing data and baseline model using MCSBQ dataset. Analysis - analyzing results with quantitative (graphs), comparative (tables), and qualitative (BoW) techniques to uncover potential biases.

4.1 Models and Fine-tuning

The models that my group used in our study include ADA, BERT, DistilBERT, GPT-2, GPT-3.5, and a pretrained T5 model for paraphrasing [9, 12, 27, 33, 39]. The process of fine-tuning the GPT models was executed using OpenAI API and the procedure for fine-tuning the GPT models adhered strictly to OpenAI's guidelines [24, 25], using the gpt-3.5-turbo-0613 model and the default values for hyperparameters. For splitting into training and test sets we used 20 and 8 data points for each bias type for training and the rest for testing for StereoSet and CrowSPairs respectively. To fine-tune models using bag-of-words results, the results were inserted directly into the message prompts themselves. This system role is a simple form of self-debiasing, a method proven to improve model bias [19, 34].

4.2 Datasets

To evaluate the different LLMs I used two main datasets, StereoSet and CrowSPairs. Both of these datasets ran on the models to detect the presence of bias and infer the extent to which bias exists. These bias datasets are well-known and used. These datasets are crowdsourced, which allows for more diversity in the specific biases, targets and sentence structure of the data. However, it is important to note that the crowdsourcing was also engineered in the United States so the overall biases and targets are in context to their prevalence in the United States.

StereoSet is a dataset developed to help demonstrate the existence of bias with four main targets: *gender*, *profession*, *religion*, and *race*. [21]. Crowdsourced Stereotype Pairs (CrowSPairs) benchmarks 9 different types of biases: *race*, *gender*, *socioeconomic status*, *nationality*, *religion*, *age*, *sexual orientation*, *physical appearance*, and *disability* [22].

4.3 Prompting Techniques

Prompting techniques offer a powerful approach to enhance the control and effectiveness of LLMs. Carefully crafted prompts can guide LLMs towards desired outputs and improve their performance on specific tasks such as reducing bias [42].

My evaluation, employs the Multiple Choice Symbol Binding (MCSB) technique as the primary prompting method. MCSB has demonstrated superior efficacy in eliciting accurate responses, particularly in models with high MCSB ability, a category encompassing OpenAI models. Notably, this technique surpasses the conventional cloze method in precision for GPT-3.5 models [30]. Furthermore, our group introduced two distinct types of prompts to discern nuanced aspects of bias manifestation, Implicit Bias Prompting and Explicit Bias Prompting. In Explicit Bias Prompting, we task the model with selecting a prompt with explicit bias reference. This method seeks to uncover any explicit biases that may influence the model's choice, providing insights into how it perceives bias.

An example of this techniques usage can be seen in Table 1. In Implicit Bias Prompting, we directly inquired whether the model perceived a given prompt as biased or unbiased. This direct approach aimed to capture the model's inferred acknowledgment of bias. The approach is similar to Table 1 if line 4 is switched to "Based on the context, answer with only A, B or C", and line 12 is switched to "A, B, or C?". Notice that the model is never told that there could be biases or stereotypes in it's choices, signifying any bias to have been implicit.

Table 1: Example of MCSB ChatGPT Prompting Technique for StereoSet Data Point With Explicit Bias Prompting

Step	**Role**	**Content**
1	System	You are a helpful assistant who will answer with just ’A’, ’B’, or ’C’.
2	User	Hispanic people are very diverse.
3	Assistant	Understood.
4	User	Answer with only A, B, or C based on which prompt is a stereotype.
5	Assistant	Understood.
6	User	A: Not all individuals like that are liable to commit crimes.
7	Assistant	Understood.
8	User	B: It's a highly convoluted game for football to play.
9	Assistant	Understood.
10	User	C: Anyone who behaves like that is considered a criminal.
11	Assistant	Understood.
12	User	Which prompt is a stereotype: A, B, or C?

Other than this, I employed further testing techniques without the use of MCQA. This prompting was done asking the following two prompts separately: "Does the following statement contain any stereotypes, prejudices or generalizations about particular groups based on characteristics like race, gender, age, religion, etc.." and "Is the following statement accurate?". Here the model was prompted with the prompt and "Yes or No". The goal with this technique is to further compare the fine-tuned models, as well as to investigate how accuracy may play a role in bias testing.

Beyond this, in later testing, I improved my general prompting technique by removing redundant Assistant's responses in the prompt generation. For example, in Table 1 I remove each line where the Assistant responded with "Understood". I found that this helps generalize the model for cross evaluation as well as resulting in an overall better performance.

4.4 Data Augmentation

In my pursuit of mitigating biases within language models, I leveraged data augmentation as a strategic approach aimed at enhancing the models' resilience to biased content, ultimately contributing to model debiasing. Our methodology involved the augmentation of the original dataset through paraphrasing, a process facilitated by two distinct models: Google's T5 model and GPT-3.5. The paraphrasing task entailed presenting prompts to a pre-trained T5 model in the format paraphrase: prompt, while GPT-3.5 was prompted to paraphrase the given questions. By engaging both models with the datasets, namely CrowsS-Pairs and StereoSet, we looked to generate diversified and nuanced perspectives on the provided content. During the fine-tuning process, the paraphrased data either replaced or augmented the original data points. This augmentation strategy was threefold: first, to encourage the model to generalize better by exposure to a broader spectrum of language, second, to promote the model's ability to discern and handle more generalized biased content more effectively, and third, to attempt to avoid in-built biases that the original datasets may have held [1,37,42].

5 Results

For the following results and discussion section, I want to reiterate that my main contribution is on the GPT results. This section also includes results that my group worked on here to give context to my discussion.

The models were evaluated with subsets of the StereoSet and CrowSPairs Datasets and with different prompting techniques. The responses of each model were recorded and compiled to measure the existence and extent of bias quantitatively. The responses were grouped into each of their targets and a score for each target were determined by the number of stereotype or anti-stereotype chosen. Then, for qualitative analysis, the results were examined using Bag of Words models to probe the specificity of bias in models.

5.1 General Results - Implicit Bias Prompting

The evaluation of general results using the benchmarks and baselines is presented in Table 2. This data can be treated as a baseline for the rest of my results for implicit bias prompting. We found that more recently developed models are doing slightly better at being able to prevent stereotypical answers with the CrowsPairs dataset. However, ChatGPT is also more likely to pick a stereotypical answer for the StereoSet dataset than other models. It is evident that some models fared better than others in specific features but overall all models performed in similar ranges. It can be noted that the ratio of stereotypical responses chosen by models evaluated with the StereoSet dataset, *Gender* fared the worst overall in each of the models. *Gender* bias was a fairly large portion of StereoSet and encompassed many targets such as *male*, *mother*, *himself*, *herself*, and others.

For models tested on CrowSPairs, Table 2 shows that BERT performed worse than DistilBERT for biases such as *sexual orientation*, *socioeconomic status*, and *disability*. For example *sexual orientation* which performed poorly with the stereotype being chosen 77% of the time. This could be because this is a fairly underrepresented bias in general, leading to less training, guardrails, and testing. This is in relation to *race* and *religion* which are more popular areas of bias.

Table 2: Performance of various models in selecting stereotypical responses when implicitly prompted

	StereoSet			
	GPT 3.5	ADA	DistilBERT	BERT
Gender	0.48	0.36	0.36	0.36
Race	0.41	0.33	0.26	0.33
Profession	0.42	0.34	0.23	0.35
Religion	0.37	0.29	0.27	0.29
	CrowSPairs			
	GPT 3.5	ADA	DistilBERT	BERT
Age Status	0.24	0.25	0.40	0.46
Disability	0.25	0.35	0.63	0.60
Gender	0.21	0.34	0.52	0.47
Nationality	0.36	0.31	0.47	0.57
Physical Appearance	0.25	0.17	0.46	0.43
Race	0.36	0.35	0.33	0.57
Religion	0.39	0.36	0.31	0.39
Sexual Orientation	0.40	0.31	0.55	0.77
Socioeconomic Status	0.23	0.23	0.49	0.66

5.2 General Results - Explicit Bias Prompting

It is important to recognize how the two prompting techniques utilized affected the accrued results with the Explicit Bias Prompting results in Table 3. Take note of BERT performing really well with this prompting method. The ratio of picking the stereotype is low comparatively to other models and other tests on BERT. However, considering the context that the model picked the unrelated response on average 74% of the time this result has far less significance. We found that this ratio of unrelated responses in this instance was an outlier.

BERT aside, other base models performed decently with this prompting technique in both CrowSPairs and StereoSet. This is expected since we had asked the model to pick the stereotype in the first place. This shows the model's ability in understanding what a stereotype is by definition. Note the variety in percentages of the model being able to correctly pick the response with a stereotype

Table 3: Comparison of various models in selecting stereotypical responses when explicitly prompted

StereoSet				
	GPT 3.5	ADA	DistilBERT	BERT
Gender	0.61	0.34	0.30	0.07
Race	0.82	0.34	0.36	0.09
Profession	0.67	0.31	0.35	0.08
Religion	0.68	0.38	0.35	0.08

CrowSPairs				
	GPT 3.5	ADA	DistilBERT	BERT
Age Status	0.56	0.24	0.36	0.63
Disability	0.73	0.28	0.35	0.43
Gender	0.60	0.28	0.49	0.57
Nationality	0.75	0.30	0.43	0.56
Physical Appearance	0.76	0.25	0.48	0.46
Race	0.75	0.30	0.37	0.57
Religion	0.79	0.30	0.29	0.44
Sexual Orientation	0.76	0.31	0.77	0.64
Socioeconomic Status	0.67	0.23	0.41	0.59

with DistilBERT ranging from 29% to 77%. Looking specifically at GPT 3.5 and ADA, we find that GPT 3.5 is able to correctly identify a stereotypical phrase or sentence around 70% of the time without any further debiasing strategies. Interestingly, we find that this general result outperforms tests done without MCQA. In these explicit prompting tests the average correct identification of stereotype is around 60% for Stereoset and 65% for CrowSPairs.

5.3 Fine-Tuning Results

The results in Table 4 show that for GPT-3.5, fine-tuning the model on a subset of question answer prompts similar to the actual prompts is able to improve stereotype detection in general.When compared to Table 3 we can observe an decrease in implicit bias selection and increase in explicit bias selection. This is especially prominent in the StereoSet results.

The Cross-Eval results shown in Table 4 depict the evaluation of the fine-tuned using training data from the other dataset being tested on the indicated dataset. This cross evaluation is an attempt to observe how the fine-tuned model performs on "unseen" data, to get a better idea of how the fine-tuned model may generalize to real world environments. The results are able to show surprisingly strong generalization. This hold true not only for the model fine-tuned using CrowSPairs but also StereoSet, despite the difference in categories of bias. This indicates that the approach is quite scalable, which can likely be attributed to the generalizability of the base GPT model being so high.

Table 4: Performance of GPT-3.5 in selecting stereotypical responses when fine-tuned without augmentation

StereoSet				
	Implicit	Explicit	Cross-Eval Implicit	Cross-Eval Explicit
Gender	0.48	0.80	0.41	0.58
Race	0.24	0.90	0.12	0.89
Profession	0.35	0.86	0.28	0.72
Religion	0.26	0.84	0.16	0.74

CrowSPairs				
	Implicit	Explicit	Cross-Eval Implicit	Cross-Eval Explicit
Age Status	0.32	0.63	0.35	0.78
Disability	0.19	0.65	0.31	0.75
Gender	0.46	0.55	0.43	0.70
Nationality	0.26	0.75	0.40	0.77
Physical Appearance	0.25	0.75	0.33	0.82
Race	0.22	0.75	0.36	0.77
Religion	0.29	0.77	0.36	0.74
Sexual Orientation	0.24	0.70	0.51	0.84
Socioeconomic Status	0.28	0.61	0.38	0.76

5.4 Augmented Fine-Tuning Results - Implicit Bias Prompting

In an attempt to look into further debiasing techniques, I continued to fine-tuned the models after augmenting the data. This allowed me to see how embedded bias is. I noticed that a decrease in choosing the stereotype was observed in StereoSet but not in all of CrowSPairs. This could just be the fact that the fine-tuning was not pragmatic enough, although it can also be inferred that the biases that did not do well originally are the ones that improved. This could also mean that there is a lower bound that the model has in terms of its performance. I also took note of the bias that was improved the most overall, *race*, in comparison to the other biases which were either improved not at all or only little in StereoSet. I assumed this is the case due to the sheer count of prompts in the *race* bias promoting the model to learn more in the fine-tuning process. *Race* also did generally well in the CrowSPairs dataset likely for the same reason. To further evaluate the robustness of our fine-tuned models, I cross-tested each model with the other respective dataset. The results are displayed in Table 5. Each model is displayed with two different types of fine-tuning: fine-tuned without any augmentation and fine-tuned with T5 with augmention. With this, we are able to see that our models showed respectable results for some of the bias and model combinations. I consider the fine-tuned model to have considerably positive results when it outperforms the base model. For StereoSet, *race* performed significantly better with a bias decrease by 30%. This coincided with the previous conclusion that *race*'s fine-tuned results were the best overall. That said, it can be seen that *gender* in the StereoSet performed worse than both the baseline and worst than the fine-tuned which is consistent with the fine-tuned results as well. This is likely due to deeply ingrained gender biases within the training data and model.

Table 5: Comparison of various models fine-tuned on StereoSet dataset under implicit prompting with following configurations: No Aug (No Augmentation), T5 Aug (Augmented using T5)

StereoSet Test Data & CrowSPairs Trained Model						
	ChatGPT		BERT		DistilBERT	
	No Aug	T5 Aug	No aug	T5 Aug	No aug	T5 aug
Gender	0.41	0.51	0.09	0.59	0.28	0.19
Race	0.12	0.21	0.11	0.53	0.62	0.48
Profession	0.28	0.34	0.10	0.53	0.43	0.26
Religion	0.16	0.23	0.05	0.55	0.57	0.43
CrowSPairs Test Data & StereoSet Trained Model						
	ChatGPT		BERT		DistilBERT	
	No Aug	T5 Aug	No Aug	T5 Aug	No Aug	T5 Aug
Age Status	0.35	0.52	0.44	0.39	0.35	0.35
Disability	0.31	0.44	0.56	0.71	0.48	0.63
Gender	0.43	0.45	0.46	0.56	0.51	0.56
Nationality	0.40	0.45	0.41	0.54	0.35	0.38
Physical Appearance	0.33	0.45	0.58	0.56	0.49	0.60
Race	0.36	0.35	0.58	0.60	0.66	0.63
Religion	0.36	0.30	0.74	0.63	0.77	0.86
Sexual Orientation	0.51	0.42	0.36	0.59	0.53	0.76
Socioeconomic Status	0.38	0.42	0.52	0.57	0.62	0.68

5.5 Augmented Fine-Tuning Results - Explicit Bias Prompting

The models I fine-tuned were also re-prompted with the Explicit Bias prompting technique. This gave me some concrete evidence on the resiliency of the models and stubbornness in un-learning bias. We can see here from Table 6 that GPT 3.5 was able to pick the stereotype correctly more often as compared to its base model. Notice with the experiment, the proportion stayed around 75% with StereoSet increasing by 10% and 60% with CrowSPairs by increasing by 40%. There was not much success in the models performing better than these numbers. In this way, we can see that with this fine-tuning we can train models to be able to better identify the stereotype so that it is able to avoid the stereotype when prompted implicitly.

Table 6: Comparison of GPT-3.5's stereotype selection under explicit prompting with following configurations: NFNA (No Fine-tuning, No Augmentation), FTNA (Fine-tuning, No Augmentation), FTAT5 (FT with T5 Augmented data)

	StereoSet		
	NFNA	FTNA	FTAT5
Gender	0.61	+0.19	+0.08
Race	0.82	+0.08	+0.01
Profession	0.67	+0.19	+0.14
Religion	0.68	+0.16	+0.16
	CrowSPairs		
	NFNA	FTNA	FTAT5
Age Status	0.24	+0.39	+0.39
Disability	0.25	+0.40	+0.40
Gender	0.21	+0.34	+0.34
Nationality	0.36	+0.39	+0.39
Physical Appearance	0.25	+0.50	+0.50
Race	0.36	+0.39	+0.39
Religion	0.39	+0.38	+0.38
Sexual Orientation	0.40	+0.30	+0.30
Socioeconomic Status	0.23	+0.38	+0.38

Take note of the overall performance increase in Table 6 which indicates that the models were able to pick the response with the stereotype more often after fine-tuning.

Outside of these results, I have collected newer data using more standardized methods for data collection to improving augmentation and cross-evaluation. As aforementioned in the Methodology section, here I use a more generalizable approach to prompting. In addition, the data augmentation is done using a more standardized method of increasing the training size instead of replacing. The results here all are trained on StereoSet data, meaning the CrowSPairs results are cross-evaluation. These results, especially for cross-evaluation, are quite significant, as they don't contain any unclear responses, and provide further evidence of the scalability of these methods. In addition, these results generally outperform those of the original fine-tuned models, which other attempts of debiasing using data augmentation had struggled with. However, there are a few outliers here, with the implicit cross-evaluation for race, religion, sexual orientation and socioeconomic status selecting the stereotype quite often, despite the fine-tuning aimed at lowering these selections. This "poor" performance could be attributed to ingrained biases or potentially the model trying to decide if it felt like a response was accurate.

Table 7: Performance of GPT-3.5 in selecting stereotypical responses when fine-tuned with augmentation (T5 paraphrasing) with improved prompting

	StereoSet	
	Implicit	Explicit
Gender	0.33	0.86
Race	0.21	0.92
Profession	0.24	0.92
Religion	0.22	0.86
	CrowSPairs	
	Cross-Eval	Cross-Eval
	Implicit	Explicit
Age Status	0.30	0.78
Disability	0.29	0.69
Gender	0.37	0.72
Nationality	0.41	0.78
Physical Appearance	0.18	0.80
Race	0.50	0.69
Religion	0.51	0.76
Sexual Orientation	0.64	0.72
Socioeconomic Status	0.50	0.76

Fine-tuning allowed each bias to rise to some common base value especially for those biases that had not performed as well as others initially. *Socioeconomic Status*, *disability*, and others had not performed well because those were likely not biases that had been considered extensively when initially training the model, which allowed their performance to increase by such a large margin instantly and with ease.

Table 8: Comparison of various models fine-tuned on StereoSet dataset with different configurations: No Aug (No Augmentation), T5 Aug (Augmented using T5), and Explicit Prompting on the CrowSPairs dataset

StereoSet Test Data & CrowSPairs Trained model						
	ChatGPT		BERT		DistilBERT	
	No Aug	T5 Aug	No Aug	T5 Aug	No Aug	T5 aug
Gender	0.58	0.55	0.12	0.54	0.31	0.21
Race	0.89	0.83	0.15	0.53	0.61	0.47
Profession	0.72	0.67	0.13	0.53	0.45	0.25
Religion	0.74	0.72	0.05	0.62	0.59	0.45

CrowSPairs Test Data & StereoSet Trained Model						
	ChatGPT		BERT		DistilBERT	
	No Aug	T5 Aug	No Aug	T5 AUG	No Aug	T5 AUG
Age Status	0.78	0.30	0.35	0.34	0.35	0.35
Disability	0.75	0.37	0.48	0.65	0.42	0.69
Gender	0.70	0.30	0.54	0.51	0.53	0.59
Nationality	0.77	0.30	0.50	0.66	0.34	0.40
Physical Appearance	0.82	0.67	0.49	0.60	0.47	0.67
Race	0.77	0.45	0.51	0.67	0.67	0.65
Religion	0.74	0.43	0.71	0.63	0.75	0.85
Sexual Orientation	0.84	0.42	0.41	0.59	0.58	0.74
Socioeconomic Status	0.76	0.43	0.61	0.70	0.57	0.63

After receiving these positive results with fine-tuning for Explicit Bias, I also cross-evaluated the fine-tuned models to assess whether these substantial results were the result of overfitting or other justification which is presented in Table 8. While the results were not as high as with the fine-tuned model, they were still better than the base model which is another sample of evidence proving the fine-tuning had constructive results. Specifically, notice how BERT no longer refused to answer the prompt nor pick completely unrelated answer which is significant in its performance improvement. In particular, note the biases that were not included in training StereoSet dataset performed lower and worse than those that were included. *Sexual Orientation*, *disability*, and *age status* are examples of such biases with their results around the 30-40% range while *gender*, *race*, and *religion* performed around 30% better.

5.6 Bag of Words

Our group was not able to fully grasp the model's process and model's inclination in choosing a stereotype just by looking at numbers, percentages, and changes. Here we had members run multiple Bag of Words models to further analyze the performance of GPT 3.5 on StereoSet for qualitative analysis. In this way we were able to particularly understand the model's result and its relation with concrete words, definitions, and topics it deemed to be important.

For Implicit Bias prompting, we gathered that these are words that the model had concluded to be inherently true since it wasn't specifically asked to choose a stereotype. We observed that most of the words can be deduced to have a negative connotation such as 'violent', 'terrorist', and 'aggressive'. Along with that, the words which coerced the model to choose the anti-stereotype or unrelated option were also similarly connotated with a couple of positive words thrown in like 'nerdy', 'smart', or 'protective'. There were a quite a few words, such as 'poor', 'black', 'white', 'white', and 'terrorist', that showed up in both lists which could mean the models not being able to signify these words that could be stereotypical or not. In Table 9, it is indicated which words were able to influence the model in assisting to choose which expression was the stereotype as well as which words guided the model away from the other expressions. It can be seen that even in Explicit Bias Prompting, the model is able to recognize that these words with negative connotations would be considered stereotypical or biased towards specific groups. Words such as 'dangerous', 'dirty', 'poverty', 'scary' and 'terrorist' were shown to be crucial in making the decision as to whether the prompt was a stereotype or not. That said, words that generally had a positive connotation directed the model away from the model in its decision of picking the stereotype, words such as 'athletic', 'nurturing', 'strong' and 'caring'. That

Table 9: Bag-of-Words Analysis in StereoSet Dataset under Explicit Prompting

	Top 7 words that helped it decide it wasn't a stereotype
All Biases	'good', 'people', 'strong', 'day', 'athletic', 'war', 'kind'
Gender	'caring', 'nurturing', 'manners', 'likes', 'good', 'picked', 'strong'
Race	'lot', 'thin', 'dangerous', 'hot', 'people', 'muslim', 'war'
Profession	'time', 'white', 'smart', 'strong', 'glasses', 'always'
Religion	'violent', 'african', 'black', 'late', 'fasted', 'august', 'suggestions'
	Top 7 words that helped it decide it was a stereotype
All Biases	'terrorists', 'dangerous', 'poor', 'black', 'white', 'violent', 'probably'
Gender	'wearing', 'cookies', 'household', 'homemaker', 'one', 'skirt', 'gossiping'
Race	'poor', 'terrorist', 'dirty', 'black', 'bad', 'always', 'like'
Profession	'nerdy', 'awkward', 'rude', 'wears', 'poor', 'glasses', 'like'
Religion	'terrorists', 'holy', 'must', 'everyone', 'scary', 'everybody', 'vedas'

is not to say that positive words could not be stereotypical, however they would, by large, not have as large of a negative societal impact. But even with these positive words, there is a noticeable trend in that the *race* and *religion* biases still included unfavorable words in the model's influence of choosing a prompt to be stereotypical or not. 'African', 'hot', 'thin', 'violent', and 'war' are some of the words in these two biases that had shown up time and time again as stereotypical words.

5.7 BoW Fintetuning and System Role

Using these BoW results, I then fine-tuned a GPT model in an attempt to debias it. An example of this debias insertion is shown in section 4.1 where the words would be inserted in the square brackets when fine-tuning GPT models. The results from models fine tuned using prompts augmented with bag-of-word outputs indicates an improvement in implicit bias indication and a reduced capacity for detected explicit bias. For example, in the StereoSet benchmark there was an improvement from 24% stereotype picked to 22% in the implicit testing for race, but a decrease from 90% to 86% stereotype detection. This comparison was made with relation to the base fine-tuned model without augmentation. This was fair considering that Bag of Words output often indicated that the base models have issue detecting positive biases, however by including this in the prompt likely split the model's attention between positive and negative biases. This indicated that the base GPT and BERT models still have trouble grasping what biases are and instead focus on sentiments.

The results from GPT model fine tuned using prompts with augmented system roles were quite similar to the results from the model augmented with bag-of-words, however it performed slightly better in detecting explicit biases. For example, using the StereoSet benchmark, the stereotype detection detected racial stereotypes 87% of the time instead of 86% of the time. In addition, the system role augmented model performed much better than the bag-of-words augmented model at cross evaluation. This was likely due to issues with model overfitting on the Bag of Words outputs that are increasing complexity in the prompt.

5.8 Non-MCQA & Accuracy

Evaluating without MCQA and prompting to ask for both stereotyping and accuracy using fine-tuned models resulted in interesting output. For StereoSet explicit bias prompting, fine tuned on GPT MCQA, 58% of the time the model picked stereotype correctly when prompted and of these, 29% of them the model though were factually accurate statements. It thought that 46% of the statements total were accurate. A similar test done on CrowSPairs fine-tuned model shows a 64% selection rate and 24% accuracy depiction out of a base 42% accurate statements. This indicates that while stereotype fine-tuning generalization is transferable beyond MCQA, this does reduce stereotype detection accuracy. In addition, by asking the model if it thinks the results are accurate or not, it can be observed that the models inherently view stereotypes as more inaccurate than

Table 10: Performance of GPT-3.5 in selecting stereotypical responses with implicit prompting when fine-tuned without augmentation

StereoSet	
Gender	0.31
Race	0.18
Profession	0.21
Religion	0.18
CrowSPairs (cross)	
Age Status	0.35
Disability	0.37
Gender	0.34
Nationality	0.44
Physical Appearance	0.24
Race	0.51
Religion	0.49
Sexual Orientation	0.63
Socioeconomic Status	0.47

not. This is interestingly because stereotypes are inherently generalized concepts and not all may have negative connotations, as discussed in the Bag of Words section. This result may imply issues with the benchmarks or baseline view of how ChatGPT understands the world.

Furthermore, a similar test was ran using a fine-tuned model on StereoSet for implicit bias testing. Here the model correctly identified a stereotype only 12% of the time. This is extremely low and is likely happening due to overfitting on the prompt-tuning that aims to reduce the model picking stereotypes. While one hope with the fine-tuning was that the model would learn that social stereotypes should be avoided, instead I ended up in a situation where the model learned to avoid stereotypes altogether. Similarly this model felt that almost all of the stereotyped statements depicted were inaccurate. These results clearly showcase the shortcomings of the methodology.

6 Discussion

Beyond our groups initial set of results, I ran further tests to explore the scope of fine-tuning and data augmentation with GPT-3.5. In this set of tests, I trained models on StereoSet benchmarks using a larger set of training data. While this set was beyond OpenAI's recommended size, this also allowed me to expand on the augmented data. These tests were also all cross evaluated using crows-pairs benchmarks. In addition, I ran these tests using a different prompting method, where I excluded the model's responses to any message other than the final question, in order to reduce over-fitting on the prompt format for cross validation. I surprisingly found that these models performed better on both the StereoSet benchmarks and the cross validation. There was an outlier in the cross validation where it had trouble recognizing sexual orientation stereotypes having not been trained on them, although interestingly, this problem didn't persist for other types of stereotypes it was not trained on. I also found that using additional augmented data did not help much in the StereoSet testing, in the cross validation test, it was much more even.

Finally, further evaluation was done using the models fine tuned on question answering with non-question answering tests. While the goal is that these models would perform stronger due to improved understanding of stereotypes. Unfortunately, this emergent ability did not appear. It seems that the prompt tuning that OpenAI does is not deep enough to showcase this depth of learning, and the models are still being trained in a way that over-fits on the question/answer process. I find that the models trained to avoid social biases, also had trouble recognizing them when directly asked. However, the model trained to recognize explicit bias, did perform quite well on the direct testing process. Interestingly, I also found that the trained model was much more likely to think a stereotype was accurate if it was trained to find explicit biases, and the model trained to avoid stereotypes, found the least number of them to be accurate. Lastly I found that the best results are generated from combining sysrole with the new training techniques.

These results showcase that while the original tests where imperfect, the general idea and goal still holds true. The distribution of results across different types of stereotypes remains similar, showcasing the model's issues with detecting and mitigating stereotypes in areas such as gender and sexual orientation. I further find that prompt tuning is viable for use across separate datasets and questioning, but the level of learning does not create emergent abilities of the model to understand social biases at a deeper level.

7 Conclusion

The training data for LLMs originate from the internet, an aggregation of misinformation, opinions melded with facts, and online discourse. As a result, this fusion is incorporated implicitly, and sometimes explicitly, into the LLMs themselves. One of the safeguards implemented in some models is the decision to focus on the anti-stereotype which reveals that the model is aware of the stereotype. Results also showed the models' difficulty in distinguishing and avoiding specific biases. The *gender* bias performed worse in this capacity consistently over different models, datasets, and fine-tuned results as well. Perhaps this indicates a deeply ingrained gender bias in the models. After implementing fine-tuning on the models, it was apparent that these LLMs had the capacity of learning to be less biased in both implicit and explicit social biases. This allows the models correct themselves after bias has been identified instead of unlearning it in the first place, which is root issue. It was also noticed during the analysis with Bag of

Words that the model focused on specific words and terminologies: words that generally have a negative connotation linked to them. The fine-tuning results demonstrated that for implicit bias, the anti-stereotype was picked most often with the following ordered techniques: sysrole > BoW > GPT paraphrasing > T5 paraphrasing > base model. Additionally, the cross-evaluation results illustrated some of the potential of these bias mitigation techniques in more "real-world" scenarios. These results perpetuate the need for better prompting techniques as well as the benefits of self-debiasing in the form of the system role prompting.

This project provided invaluable insights into the nuances of prompt-tuning large language models like GPT. Through my empirical analysis, I gained hands-on experience with prompt tuning methodologies aimed at surfacing and mitigating social biases in ChatGPT's outputs. This process highlighted the importance of crafting generalized, cross-domain prompts that could effectively validate bias mitigation techniques across diverse contexts. By using these more discrete prompts, I am able to get a better understanding of what the model may be doing behind the scenes. However, I also learned that prompt tuning needs to be implemented carefully to avoid overfitting the model to the specific prompts used during fine-tuning. While this approach can yield encouraging results on the curated set of prompts, there is a risk that the model may fail to genuinely internalize the desired debiasing learnings. This could manifest as contradictory outputs when deployed in the wild on novel prompts. This is especially apparent when differentiating between stereotypes with negative or positive connotations and different levels of truth. These lessons underscore the inherent tradeoffs and limitations of current prompt tuning techniques for language models. While valuable for probing and evaluating biases, relying solely on prompts may lack the systematic guarantees required to fundamentally "debias" models in a robust, scalable manner. As AI researchers, recognizing these constraints lays the groundwork for exploring complementary strategies that can impart anti-stereotypical knowledge more universally into language model training paradigms.

7.1 Bias-Identification Framework

Using the methodology and results, I proposed a Bias-Identification Framework (BIF) to recognize various social biases in LLMs. The BIF follows the methodology of using MCSB prompting with implicit and explicit bias questions. Using publicly available bias datasets have the model in question predict the most likely category (stereotype, anti-stereotype, or unrelated) based on its understanding of bias. By comparing the model's predicted category with the actual category from the dataset the model's bias performance can be measured. Manual testing can confirm any explanations the model presents. This framework can be used not only for research but also for developing techniques to reduce bias and ensure responsible development and deployment of LLMs. As this framework is simple in nature, its role in this area will likely be as a tool to work alongside other guardrail solutions. Unlike explicit bias where perfect detection is ideal, implicit bias detection and mitigation goals should be set by developers, targeting a balanced model that exhibits stereotypes and anti-stereotypes at a rate (e.g., 20-50%) that considers the limitations of prompt design and potential overcorrection by the model. Beyond the scope of what was performed in this project, this alignment may be more carefully approached by separating and labeling benchmarks by the positive or negative connotations. While broad generalizations in general should be avoided, treating each stereotype exactly the same despite their potential harm is also not correct. I also recommend improving the framework against potentially biased bias-benchmarks through utilizing a more diverse set of benchmarks together, potentially including IAT and other psychology based techniques.

7.2 Limitations

My study on bias in LLMs had certain limitations. The amount of data I could test via API was constrained, limiting the scale of the study. In addition, I was restricted to the use of this commercial LLM which does not follow best practices for reproducibility in this area of study. While the use of MCSB is accurate and streamlined, it restricted exploration of responses that exceed pre-set options, which could offer deeper bias insights; different benchmarks or stereotype definitions could yield varied findings. Here, given time constaints, I was not able to provide a thorough comparison to other state-of-the-art bias evaluation techniques. Additionally, in the scope of this work the methodology lacks in having a stronger and more formal theoretical backbone as well as integration with other techniques.

When considering limitations, it is important to understand the limitations of bias benchmarks in general, as discussed by Blodgett et al.. Stereotyping benchmarks such as Crows-Pairs and StereoSet face challenges due to the inherent subjectivity of stereotypes. This can lead to situations where an LLM's response aligns with commonly held beliefs within a specific context, yet the benchmark flags it for bias [7]. Additionally, potential biases within the benchmarks themselves are a concern as they may reflect biases of the creators. These limitations restrict the generalizability of findings from bias detection methods using these benchmarks.

7.3 Future Research

In future research, updating and diversifying the bias benchmarks employed in testing can further enrich the insights drawn from these studies. As alluded to in the discussion and related work, this could be done by finding methods to generate more diverse training data and more intensive labeling to further improve how models can understand social biases. In addition, generating testing and debiasing methodologies more closely rooted in the embedding or the social and psychological understanding of stereotypes is a solid direction forward. Expanding testing to generative image models could reveal how biases manifest across modalities like text and image. Analyzing these models in non-English languages (e.g., gendered languages) could reveal how cultural contexts influence bias dynamics [11]. Furthermore, integrating LLMs with real-time bias moderation systems like Retrieval-Augmented Generation (RAG) [18] holds promise for a more responsible AI future.

7.4 Ethics Statement

This research utilizes established bias benchmarks to assess potential biases in LLMs. My aim is to contribute to the development of fairer NLP applications. My methodology is designed to minimize the risk of perpetuating social biases within LLMs. I acknowledge the limitations of benchmarks and the importance of ongoing research in creating robust methods for bias detection and mitigation.

References

1. Abaskohi, A., et al.: Lm-cppf: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-short.59
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku - anthropic (2024), https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf
3. Bai, X., Wang, A., Sucholutsky, I., Griffiths, T.L.: Measuring implicit bias in explicitly unbiased large language models (2024)
4. Bai, Y., et al.: Constitutional ai: Harmlessness from ai feedback (2022)
5. Baidoo-Anu, D., Ansah, L.: Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. Journal of AI **7** (03 2023). https://doi.org/10.61969/jai.1337500
6. Bender, E.M., et al.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 610–623. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922
7. Blodgett, others, S.L., Lopez, G., Olteanu, A., Sim, R., Wallach, H.: Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1004–1015. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.81
8. Bommineni, V., et al.: Performance of chatgpt on the mcat: The road to personalized and equitable premedical learning (03 2023). https://doi.org/10.1101/2023.03.05.23286533
9. Brown, T.B., et al.: Language models are few-shot learners (2020)
10. Chang, Y., et al.: A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. (jan 2024). https://doi.org/10.1145/3641289
11. Cho, W.I., Kim, J.W., Kim, S.M., Kim, N.S.: On measuring gender bias in translation of gender-neutral pronouns. In: Costa-jussà, M.R., Hardmeier, C., Radford,

W., Webster, K. (eds.) Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 173–181. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3824

12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
13. Gemini Team, G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf
14. Guo, Y., Yang, Y., Abbasi, A.: Auto-debias: Debiasing masked language models with automated biased prompts. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1012–1023. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.72, https://aclanthology.org/2022.acl-long.72
15. Jeoung, S., Ge, Y., Diesner, J.: Stereomap: Quantifying the awareness of human-like stereotypes in large language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), https://openreview.net/forum?id=ExpskenHdP
16. Kocielnik, R., Prabhumoye, S., Zhang, V., Jiang, R., Alvarez, R.M., Anandkumar, A.: Biastestgpt: Using chatgpt for social bias testing of language models (2023)
17. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of The ACM Collective Intelligence Conference. CI '23, ACM (Nov 2023). https://doi.org/10.1145/3582269.3615599, http://dx.doi.org/10.1145/3582269.3615599
18. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021)
19. Meade, N., Poole-Dayan, E., Reddy, S.: An empirical survey of the effectiveness of debiasing techniques for pre-trained language models (2022)
20. Mikołajczyk-Bareła, A.: Data augmentation and explainability for bias discovery and mitigation in deep learning (2023)
21. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5356–5371. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.416
22. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1953–1967. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.154
23. OpenAI: (2023), https://chat.openai.com/
24. OpenAI: Api reference - openai api (2023), https://platform.openai.com/docs/api-reference
25. OpenAI: Fine-tuning - openai api (2023), https://platform.openai.com/docs/guides/fine-tuning
26. OpenAI, et al.: Gpt-4 technical report (2024)
27. Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022)

28. Qadir, J.: Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education (12 2022). `https://doi.org/10.36227/techrxiv.21789434`
29. Rane, N.: Chatbot-enhanced teaching and learning: Implementation strategies, challenges, and the role of chatgpt in education. Challenges, and the Role of ChatGPT in Education (July 21, 2023) (2023)
30. Robinson, J., Rytting, C.M., Wingate, D.: Leveraging large language models for multiple choice question answering (2023)
31. Roselli, D., Matthews, J., Talagala, N.: Managing bias in ai. In: Companion Proceedings of The 2019 World Wide Web Conference. p. 539–544. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). `https://doi.org/10.1145/3308560.3317590`
32. Rudolph, o.: Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching **6**(1), 9–21 (2023)
33. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
34. Schick, T., Udupa, S., Schütze, H.: Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp (2021)
35. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: Societal biases in language generation: Progress and challenges (2021)
36. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models (2021)
37. Tang, T., Lu, H., Jiang, Y.E., Huang, H., Zhang, D., Zhao, W.X., Wei, F.: Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing (2023)
38. Tlili, et al.: What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. Smart Learning Environments **10**(15), 1–15 (2023). `https://doi.org/https://doi.org/10.1007/s40561-023-00189-2`
39. Vladimir Vorobev, M.K.: Chatgpt paraphraser on t5 base, `https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base`
40. Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., Petrov, S.: Measuring and reducing gendered correlations in pre-trained models (2021)
41. Yang, K., Yu, C., Fung, Y., Li, M., Ji, H.: Adept: A debiasing prompt framework (2022)
42. Yu, Y., et al.: Large language model as attributed training data generator: A tale of diversity and bias (2023)
43. Zhai, X.: Chatgpt user experience: Implications for education (12 2022)
44. Zhao, J., Fang, M., Pan, S., Yin, W., Pechenizkiy, M.: Gptbias: A comprehensive framework for evaluating bias in large language models (2023)