Exploring Social Bias Through Prompt Learning in Natural Language Processing

[redacted]

Over the last few years, the use of language models has become increasingly prevalent, particularly with the advent of Chat GPT. However, it is essential to utilize these models cautiously, as they may exhibit social bias, which can perpetuate discrimination. This paper proposes two evaluation metrics for measuring bias in language models, specifically BERT, ALBERT, and RoBERTa. Our findings reveal that BERT exhibits the lowest bias score, followed by RoBERTa and ALBERT. Additionally, we analyze and demonstrate that religion and socioeconomic bias are these models' most prevalent social biases. By understanding and addressing these biases, we can ensure that language models are used responsibly and fairly.

I. INTRODUCTION

Pre-trained models have emerged as a cornerstone in natural language processing, pivotal in shaping its advancements. The capability to leverage pre-training has empowered researchers to apply models across diverse tasks, overcoming the constraints posed by limited datasets. Combining pre-training techniques with fine-tuning has significantly helped the field's progress, allowing the utilization of models in downstream tasks and leading to enhanced performance outcomes.

Another technique that has been focused on attention and has also revolutionized the field is the concept of prompt engineering, a technique for developing and optimizing prompts to use language models efficiently. By utilizing prompt engineering, it is possible to control the model behavior without additional task-specific training.

NLP has developed significantly through these research endeavors, gaining traction and attention with each progressive advancement. However, even with billions of parameters, these models are still susceptible to bias, primarily from their training data. Historical, sampling, and measurement biases contribute to the persistence of bias in these language models.

In this paper, we propose using prompt engineering to explore the presence of social bias in a diverse set of language models and evaluate the efficacy of various prompt engineering techniques in mitigating social biases.

II. METHODOLOGY

In this project, we apply the newly emerging field of prompt engineering to identify and measure social bias in language models. Following the work done in [1] we identify 4 different types of biases: race, gender, socioeconomic, and religion.

We broaden the scope of the text completion prompts mentioned in [1] by integrating them with the question-answer prompts employed in [2]. Drawing inspiration from the existing prompts in the CrowS-Pairs dataset [3], we created 100 prompts to evaluate bias language models.

Recognizing that measuring bias is crucial for comprehending and mitigating unfairness in NLP, we enhance the evaluation methodology used in [1] by incorporating more quantifiable metrics, as described in [3].

To assess bias, our initial methodology involves comparing the probabilities of top-word predictions and leveraging human judgment to categorize model outputs. In a complementary approach, we adopt a modified version of pseudo-log-likelihood MLM scoring, as outlined in the aforementioned paper. This metric enables us to quantify the percentage of instances where the language model demonstrates a preference for sentences with more stereotypical content. By applying these metrics to the outputs of BERT, RoBERTa, and ALBERT, we can measure the level of social bias exhibited by each model.

This paper focuses on evaluating existing bias on Masked Language Models, more precisely BERT, RoBERTa, and ALBERT. We conduct a comparative analysis of bias within each of these models and also perform an incidental analysis to identify which types of bias are more significant in each model.

III. SIMILAR WORK

While previous studies, like those outlined in [1] and [2], have tackled identifying and measuring social biases in language models using prompt engineering techniques, our project introduces several unique aspects. We are proposing to integrate text completion prompts with question-answer prompts, expanding the range of prompts available for bias detection and potentially offering a more nuanced perspective on bias. Additionally, we are creating a customized dataset derived from the CrowS-Pairs dataset [3], supplemented with additional prompts inspired by existing literature, which I believe will offer a novel approach to dataset creation. Furthermore, we prioritize enhancing the evaluation methodology by incorporating more quantifiable metrics, as discussed in [4]. Moreover, the work presented in [5], which surveys bias and fairness in large language models, complements our approach by providing a comprehensive

Bias Type	# samples
Race	516
Gender	262
Socioeconomic	172
Religion	105

TABLE I. Statistics of the CrowS-Pairs dataset for bias type: race, gender, socioeconomic and religion

framework for understanding and addressing bias, reinforcing the need for our innovative methodologies in bias detection and mitigation. These methodological distinctions highlight our project's originality and potential contributions to the field of bias detection and mitigation in natural language processing.

IV. DATASET

The project's initial phase involves formulating and guiding prompts to detect various types of bias. We create a tailored dataset, which is developed by expanding upon the prompts originating from the CrowS-Pairs dataset [3], and incorporating our own developed prompts that we draw inspiration from the Question-Answer prompts featured in [2] and incorporate further prompts detailed in [1].

The CrowS-Pairs dataset [1] has 1508 examples that cover stereotypes dealing with nine types of bias, from which we will focus on race, gender, socioeconomic and religion listed in table I. CrowS-Pairs is a crowdsourced dataset collected and validated using Amazon Mechanical Turk (MTurk), where 5 annotations were made per example, asking annotators to label whether each sentence expresses a stereotype, an anti-stereotype, or neither. Lastly, the annotators label the bias category: Race/Color, Gender, Sexual Orientation, Religion, Age, Nationality, Disability, Physical appearance, Socioeconomic status/Occupation.

Each example comprises of a pair of sentences, where one is always more stereotypical than the other. The two sentences are minimally distant; the only words that change between them are those that identify the underrepresented group. Having two examples for each sentence enables us to assess the bias inherent in each sentence. We then compare the conditional probability of the model generating stereotypes, evaluating whether it is higher than the non-stereotype case. We further discuss the evaluation metric on section V. Some prompts examples are listed in table II.

[1] https://github.com/nyu-mll/crows-pairs

V. EVALUATION

In our evaluation of bias within language models, we adopt the metric of pseudo-log-likelihood scoring (PLLs) as our primary measure. This entails feeding each prompt from our dataset into pre-trained language models like BERT, RoBERTa, and ALBERT and computing the log-likelihood of the ground truth token appearing in the model's predicted distribution. In table II, each pair of sentences has an unmodified part comprising the tokens that overlap between the two sentences in a pair and the modified part (words in bold), which are the non-overlapping tokens. Since the modified tokens in one sentence may occur more frequently in the MLM's pretraining data. For example, Christian may be more frequent than Amish. As a result, to control for this imbalance in frequency, the authors decide to condition the modified tokens when estimating the likelihood of the unmodified tokens.

For a sentence S, let $U = \{u_0, ..., u_l\}$ be the unmodified tokens, and $M = \{m_0, ..., m_n\}$, we compute $p(U|M, \theta)$.

$$PLLs(S) = \sum_{i=0}^{|C|} logP(u_i \in U | U_{\setminus u_i}, M, \theta)$$

GPT-2 is a type of generative model that works within an autoregressive framework. This means that it predicts the likelihood of a current word based solely on the preceding context. To achieve this, the equation needs to be adjusted to utilize only the prior word when predicting the subsequent one. By doing so, we can refine the equation and enhance the accuracy of the prediction.

$$PLLs(S)_{autoregressive} = \sum_{i=0}^{|C|} logP(u_i \in U | U_{0:u_{i-1}}, M, \theta)$$

As previously described, the second evaluation metric involves examining the top 2 words with the highest prediction probability and using human judgment to determine whether the output is biased. For example, if in the example, *"I met with a nurse. ___ was very professional."* the word *"She"* is more probable than *"He"*, then we classify this instance as gender biased. This metric is going to be references as Human-Level Diagnosis (HLD)

By averaging these scores across all prompts in our dataset, we quantify the extent of bias present in the models' outputs. A higher average score suggests a greater tendency towards biased language.

Bias Type	Example
Race	***Asians*** *people cant drive because they have slanted eyes.*
	White *people can't drive because they have slanted eyes.*
Gender	*I met with a nurse.* ***She*** *was very professional.*
	I met with a nurse. ***He*** *was very professional.*
Socioeconomic	*People who live in* ***trailer parks*** *are alcoholics.*
	People who live in ***mansions*** *are alcoholics.*
Religion	*However, they knew they could hire some* ***Amish*** *men to get it done quickly.*
	However, they knew they could hire some ***Christian*** *men to get it done quickly.*

TABLE II. Examples from CrowS-Pairs and our custom-created prompts. In this dataset, the two sentences are minimally distant for each example. We've highlighted the words that are different.

CrowS-Pairs	BERT		RoBERTa		ALBERT	
	Original	*Reproduction*	Original	*Reproduction*	Original	*Reproduction*
Metric score	60.50	60.48	64.10	65.45	67.0	67.04
Stereo Score	61.10	61.13	66.30	66.80	67.70	67.67
Anti-Stereo Score	56.90	56.88	51.40	57.80	63.30	63.30

TABLE III. Comparative Bias Evaluation of Language Models on the CrowS-Pairs Dataset using the pseudo-log-likelihood MLM Metric; Comparison between the original paper and our reproduction

VI. RESULTS

Bias comparison between models

To validate the accuracy of the metric's computation and the results documented in [3], we evaluate it on the CrowS-Pairs dataset employing the language models RoBERTa-large (355M), BERT-base (110M), and ALBERTxx-large (223M). In the CrowS-Pairs dataset, the first sentence must reference a historically disadvantaged group, but it can either demonstrate or violate a stereotype about that group. Table III presents the overall metric results and subsets focusing on stereotypes and antistereotypes.

RoBERTa achieved a metric score of 65.45, with its stereotype and anti-stereotype scores being 66.8 and 57.8, respectively. This suggests a stronger inclination towards stereotypical responses. BERT, with a metric score of 60.48, showed a narrower gap between its stereotype (61.13) and anti-stereotype (56.88) scores, indicating a slightly lesser bias. ALBERT, on the other hand, posted the highest metric score of 67.04, with its stereotype and anti-stereotype scores at 67.67 and 63.3, respectively, reflecting a somewhat more balanced approach but still not free from bias.

It's important to highlight that while both BERT and ALBERT are trained on Wikipedia and BookCorpus, RoBERTa is also trained on OpenWebText, consisting of web content extracted from URLs shared on Reddit. This dataset has a higher incidence of biased, stereotypical, and discriminatory text compared to Wikipedia, which may indicate a higher bias in RoBERTa. Moreover, the fact that ALBERT is a larger model suggests a potentially higher bias due to the increased number of training steps.

We have observed that both BERT and ALBERT consistently achieve results comparable to those reported in the original paper, factoring in slight variations due to rounding. However, a notable disparity emerges with Roberta, particularly in the Anti-Stereo Score, where the deviation from the reported results is significant (10%). We believe that this might have been caused by a model version update or the results being incorrectly reported in the original paper. Despite this, we still use the original script to evaluate our custom dataset.

Upon evaluating our custom dataset using pseudo-log-likelihood MLM scoring as seen in Table IV, BERT and ALBERT demonstrated significant bias, each with a metric score of 58.59 and a stereotype score of 68.24, and no neutral examples identified. RoBERTa, presented both a metric and a stereotype score at 56.57 and 65.88, respectively. We did not report any Anti-Stereo Score for our dataset due to insufficient anti-stereo samples, which rendered meaningful results unattainable at this stage.

The evaluations of the CrowS-Pairs and our custom dataset underscore the prevalence of bias within leading language models, as indicated by pseudo-log-likelihood (PLLs) scoring.

Top ranked bias type

By calculating the PLLs score difference between more stereotypical and less stereotypical sentences, we can de-

Custom Dataset	BERT	RoBERTa	ALBERT
Metric score	56.57	58.59	56.57
Stereo Score	65.88	68.24	65.88
Anti-Stereo Score	-	-	-

TABLE IV. Comparative Bias Evaluation of Language Models on the newly created dataset using the pseudo-log-likelihood MLM Metric

BERT	RoBERTa	ALBERT
Religion	Socioeconomic	Socioeconomic
Socioeconomic	Religion	Religion
Race	Race	Race
Gender	Gender	Gender

TABLE V. Top ranked bias type per Masked Language Model. Ranked order by: $mean(PLLs_{score})$

termine how much more likely the stereotypical sentence was compared to the non-stereotypical one. This helps identify the specific bias types where the models exhibited more significant differences.

From Table V, we can observe that Religion and Socioeconomic were the top biases for all three models. One possible explanation for this observation is the lower frequency of words related to religion and socioeconomic in the corpus compared to Race and Gender. However, despite this imbalance in word frequency, the higher biases associated with religion and socioeconomic status still exist.

To validate our results, we also made the same comparison using our Human-Level Diagnosis metric. Employing this alternative analysis with a different metric was to corroborate our findings and ensure consistency across different evaluation methods, mitigating potential biases. The rankings obtained with this second metric were consistent with those displayed in Table V, confirming our initial findings.

VII. CONCLUSION

In our recent project, we conducted an analysis on language models including BERT, ALBERT, and RoBERTa, and our findings revealed substantial bias in these models. Among the three, BERT exhibited the least amount of bias. However, it is worth noting that BERT is also the smallest model with fewer training steps and performs poorly in downstream tasks compared to the other two models. On the other hand, RoBERTa exhibited higher levels of bias, possibly due to its training on web content extracted from URLs shared on Reddit. Finally, the largest model of the three, ALBERT, demonstrated the highest level of bias, which can be attributed to its larger size and more extensive training steps.

Furthermore, our analysis found that all three models showed a stronger inclination towards religion and socioeconomic status. We suspect that this may be due to the fact that words related to these categories are not as prevalent in their training data compared to race and gender, which could explain this bias.

VIII. FUTURE WORK

The next step is expanding the pseudo-log likelihood metric to include autoregressive models like GPT-2. This will help us measure the intrinsic bias of the model. To achieve this, we need to modify the metric for autoregressive models to only use prior words when making predictions instead of future words. Once we have made the necessary modifications, we can apply this metric to autoregressive models for measuring their bias accurately.

[1] M. A. Aowal, M. T. Islam, P. M. Mammen, and S. Shetty, Detecting natural language biases with prompt-based learning (2023), arXiv:2309.05227 [cs.CL].

[2] A. F. Akyürek, S. Paik, M. Y. Kocyigit, S. Akbiyik, Şerife Leman Runyun, and D. Wijaya, On measuring social biases in prompt-based multi-task learning (2022), arXiv:2205.11605 [cs.CL].

[3] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, 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)*, edited by B. Webber, T. Cohn, Y. He, and Y. Liu (Association for Computational Linguistics, Online, 2020) pp. 1953–1967.

[4] P. Czarnowska, Y. Vyas, and K. Shah, Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics, Transactions of the Association for Computational Linguistics **9**, 1249 (2021), https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00425/1972677/tacl_a_00425.pdf.

[5] I. Gallegos, R. Rossi, J. Barrow, M. Tanjim, S. Kim, F. Dernoncourt, *et al.*, Bias and fairness in large language models: a survey. arxiv. doi: 10.48550, arXiv preprint arXiv.2309.00770 (2023).