Evaluating the Efficacy of Prompting Techniques for Debiasing Language Model Outputs (Student Abstract)

*Shaz Furniturewala[1], *Surgan Jandial[2], *Abhinav Java[2], *Simra Shahid[2], Pragyan Banerjee[1], Balaji Krishnamurthy[2], Sumit Bhatia[2], Kokil Jaidka[3]
[1]BITS Pilani
[2] MDSR Labs, Adobe
[3] National University of Singapore
mohammadshaz529@gmail.com

Abstract

Achieving fairness in Large Language Models (LLMs) continues to pose a persistent challenge, as these models are prone to inheriting biases from their training data, which can subsequently impact their performance in various applications. There is a need to systematically explore whether structured prompting techniques can offer opportunities for debiased text generation by LLMs. In this work, we designed an evaluative framework to test the efficacy of different prompts for debiasing text along different dimensions. We aim to devise a general structured prompting approach to achieve fairness that generalizes well to different texts and LLMs.

Introduction

Pre-trained language models (LMs) comprise data scraped from a net; thus, they form an approximation of human information, including unvetted sources (e.g internet forums) (Ferrara 2023), human stereotypes and biases and a skew in the representation of various social groups. Furthermore, when they are finetuned for downstream tasks, these biases get propagated into the finetuned models as well (Ladhak et al. 2023). One of the most effective ways to remove bias from LMs is to augment the training data using methods like Counterfactual Data Augmentation (Chen et al. 2023). However, as LMs get larger with training corpora reaching trillions of tokens, creating unbiased corpora is incredibly hard, and retraining LMs of that scale becomes increasingly computationally expensive. With the proliferation of prompt engineering, intricate strategies are developed and evaluated for a wide range of tasks; however, their adaptation and applicability to debiasing tasks lack attention beyond the few initial works (Borchers et al. 2022).

In this work, we propose to address this gap by adapting and evaluating the state-of-the-art prompting approaches on a standard debiasing benchmark. Our experiments encompass popular language models, and the findings offer encouraging insights for further exploration of a general debiasing prompting framework.

*These authors contributed equally.

Experimental Setup

We followed a zero-shot prompting approach with the StereoSet dataset and a variety of prompting methods on a variety of LLMs, as described in the following paragraph.

Baseline Models and Approaches. We experimented with varied model sizes – GPT-2 Small (224M), GPT-2 Large (774M) (Brown et al. 2020), Pythia (6.9B) Biderman et al. (2023), and LLaMA2 (13B) (Touvron et al. 2023). For each of the models, we provided the same inputs together with some prepended prompts as detailed below:

- **Instruction**: "The following text is unbiased and not discriminatory:"
- **Role/Persona**: "You are an unbiased person who does not discriminate against people. "
- **Zero-Shot CoT**: "Generate some text that treats all groups of people the same and does not discriminate between them. Let's think step by step:"
- **Self Debias**: Developed by Schick, Udupa, and Schütze (2021), this is a post-hoc debiasing approach that is currently one of the highest performing algorithms for fair text generation that does not require fine-tuning. It uses specific prefixes to generate biased text and penalizes words with a high probability in the biased distribution.
- **Implicative Introspection**: This is our novel iterative prompting method that combines the initial LLM output with a new directive: "What is the implication of the given text?". These two outputs are collectively provided to generate a new unbiased output informed by the LLMs criticism of its own generation.

Evaluation Metrics. We report our findings with StereoSet (Nadeem, Bethke, and Reddy 2020), a natural language dataset that measures stereotypical bias of four types: gender, race, religion, and profession. It includes sentences like 'The boy was [BLANK] in school,' and three options to fill the [BLANK], and words like ('rowdy,' 'calm,' and 'mirror') which constitute a stereotype, an anti-stereotype, and an unrelated word. The bias (SS) score is the fraction of sentences where the stereotype is most likely and language modeling (LM) is measured by the fraction of sentences where the unrelated sentence is not the most likely. The ICAT score is a combination of the two.

Method	Gender	Prof.	Race	Religion	Overall
GPT-2 Small	62.65	61.31	58.90	63.26	60.42
+ SDB	60.84	59.68	57.78	60.40	58.96
+ Instruction	61.95	61.11	58.18	62.32	59.89
+ Zero-Shot CoT	60.53	61.22	57.47	63.39	59.46
+ Role/Persona	**59.83**	59.85	**55.84**	**57.95**	57.90
+ Implication	60.93	**57.64**	56.89	59.89	**57.78**
GPT-2 Large	67.64	64.43	62.35	66.35	63.93
+ SDB	63.39	60.74	58.47	62.20	60.06
+ Instruction	65.83	63.88	62.96	67.61	63.83
+ Zero-Shot CoT	68.22	65.86	61.97	68.40	64.43
+ Role/Persona	64.35	62.02	62.22	66.56	62.57
+ Implication	**59.03**	**55.95**	**54.43**	**60.35**	**55.79**
Pythia (6.9B)	69.39	65.18	63.52	66.30	64.97
+ SDB	64.60	60.41	58.81	60.50	60.18
+ Instruction	67.95	64.70	64.89	69.62	65.37
+ Zero-Shot CoT	69.59	65.26	66.95	68.87	66.72
+ Role/Persona	66.30	66.18	64.58	68.70	64.79
+ Implication	**61.78**	**59.12**	**57.77**	**57.61**	**58.76**
LLaMA 2 (13B)	70.54	65.47	64.47	69.60	65.78
+ SDB	63.65	61.11	59.48	65.24	60.82
+ Instruction	70.40	65.93	63.80	70.18	65.65
+ Zero-Shot CoT	72.19	66.62	68.48	74.01	68.45
+ Role/Persona	68.52	65.85	67.72	67.92	67.13
+ Implication	**62.30**	**60.08**	**59.19**	**61.42**	**59.99**

Table 1: StereoSet (SS) scores should be closer to 50%

Method	GPT-2 Small LM	GPT-2 Small ICAT	GPT-2 Large LM	GPT-2 Large ICAT
Base Model	91.01	72.04	91.77	66.21
+ SDB	89.07	73.11	88.52	**70.71**
+ Instruction	**92.00**	73.80	**93.15**	67.38
+ Zero-Shot CoT	90.90	73.69	92.68	65.94
+ Role/Persona	90.76	76.41	91.31	68.35
+ Implication	84.94	71.72	74.44	65.83

Method	Pythia (6.9B) LM	Pythia (6.9B) ICAT	LLaMA 2 (13B) LM	LLaMA 2 (13B) ICAT
Base Model	92.96	65.13	92.96	63.62
+ SDB	89.07	**70.93**	89.41	**70.06**
+ Instruction	**92.74**	64.22	**93.22**	64.05
+ Zero-Shot CoT	92.48	61.55	**93.49**	58.99
+ Role/Persona	**93.11**	65.56	93.42	61.41
+ Implication	84.02	69.29	83.00	66.42

Table 2: LM and ICAT scores should be closer to 100%

Results and Future Work

Tables 1 and 2 offer two main takeaways. First, by and large, the reasoning-based prompts (Zero-Shot CoT, Implication) outperform other approaches at reducing bias; yet, they appear to suffer from fluency issues with a low LM score, where a simpler Instruction suffices. Second, the debiasing is quite uniform over the different kinds of bias, with the greater overall improvements observed in the largest (13B) LLaMA 2 model with the self-debiasing approach.

For future work, our findings inspire a more elaborate research design comprising a larger repertoire of datasets, such as the BOLD and CrowSPairs datasets, and more LLMs. We will also further experiment with developing debiasing prompts that can benefit from a combination of instructive and reasoning approaches. The supplementary includes the code to replicate our experiments.

References

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv:2304.01373.

Borchers, C.; Gala, D.; Gilburt, B.; Oravkin, E.; Bounsi, W.; Asano, Y. M.; and Kirk, H. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 212–224. Seattle, Washington: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Z.; Gao, Q.; Bosselut, A.; Sabharwal, A.; and Richardson, K. 2023. DISCO: Distilling Counterfactuals with Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5514–5528.

Ferrara, E. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. arXiv:2304.03738.

Ladhak, F.; Durmus, E.; Suzgun, M.; Zhang, T.; Jurafsky, D.; McKeown, K.; and Hashimoto, T. 2023. When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3206–3219. Dubrovnik, Croatia: Association for Computational Linguistics.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.