Prompt-Based Debiasing for Large Language Models

Richard Huang, Jatin Chadha

Abstract

Large Language Models (LLMs) are subject to various biases as a consequence of their training data, which can be further reflected in their responses to user-input prompts.This can serve as a potential issue for commercialized LLM use, especially in situations where protected classes should not be used as a factor in inference. In this project, we propose a prompting schematic to test and help limit potential gender biases in state of the art LLMs using the WinoBias Gender Coreference Resolution dataset.

1 Introduction

LLMs perpetuate biases based on training data, including racial and gender biases.[2] As LLMs are adopted for crucial decision-making, it becomes imperative to mitigate these biases. Li et al.'s paper "Steering LLMs Toward Unbiased Responses: A Causality-Guided Debiasing Framework" discusses reducing social bias through guided prompt design.[1] The authors employ strategies that encourage bias-free reasoning on the WinoBias dataset, evaluating how LLMs assign stereotypical gender pronouns to occupations.

In this project, we apply a new prompting strategy inspired by the reference paper on the WinoBias dataset to further test selected LLMs' coreference resolution capabilities. The strategy we propose involves "Entity-Neutral Prompting" in which we prompt the LLM to replace all occupational entities with neutrals like Person 1 and Person 2, before performing coreference resolution.

2 Data and Models

2.1 Data

The Winobias[3] dataset contains 3,168 sentences split evenly into Type 1 and 2 sentences. Type 1 sentences contain semantic cues and do not contain syntactic cues while Type 2 sentences contain both semantic and syntactic cues. In our transformed dataset after cleaning, raw sentences contain brackets around the profession and pronoun while clean sentences simply contain the sentence without brackets. The profession column simply lists the profession name such as "accountant" or "janitor" extracted from the sentence and the pronoun column lists the pronoun used in the sentence such as "his" or "her." The stereotype column lists whether the sentence follows gender-occupation stereotypes with values of "pro" or "anti."

We tested our prompting strategies on only Type 2 sentences from the WinoBias dataset as they better represent real-world language use. Rate-limited and cost-prohibitive API access also influenced our choice to focus on Type 2 sentences.

2.2 Models

To test our prompting strategies, we utilize three proprietary state of the art models often used commercially: GPT 3.5-turbo, LLAMA-2 (70 billion parameters), and Gemini-1.0-pro, the latter being the most advanced of the three. We access all models through their respective APIs and apply our prompting strategies with a Python script.

3 Approaches

Li et al. discuss three strategies for prompt-based LLM debiasing. Strategy 1 focuses on selecting mechanisms for demographic information and agnostic facts, while Strategy 2 aims to counteract existing biases by managing dependencies between demographic representations and entities, ensuring no new biases are introduced. Strategy 3 involves selecting mechanisms for demographic information and aware text.

Our project builds off the debiasing prompting strategies introduced in the reference paper. We introduce our strategy "Entity-Neutral Prompting", a prompt-engineering implementation of "Fairness through Unawareness" that removes potential biases from specific occupations being associated with certain demographics by replacing entities with neutral identifiers "Person 1" and "Person 2." Thus, rather than just focusing on mitigating demographic biases, our approach entirely eliminated those demographic factors whether through using neutral identifiers or entirely removing gender-specific information from the sentences. An example of a prompt and intended transformed sentence is as follows: "Replace all occupational entities in the original sentence with neutrals 'Person 1,' 'Person 2,' etc. After generating the transformed sentence, identify who 'he' refers to by replacing it with the corresponding 'Person' label." "Person 1 reviewed the report that Person 2 had prepared, and he was pleased with the thoroughness."

4 Results

	Llama 2 (%)	GPT 3.5 (%)	Gemini 1.0 Pro (%)
Overall Original Accuracy	83.52	83.84	88.38
Overall New Accuracy	90.15	88.64	94.57
Pro-Stereotype Original Accuracy	90.15	93.18	95.45
Pro-Stereotype New Accuracy	97.60	94.19	99.37
Anti-Stereotype Original Accuracy	71.59	74.49	81.31
Anti-Stereotype New Accuracy	82.70	83.08	89.77
Male Original Accuracy	87.03	84.51	90.43
Male New Accuracy	92.95	90.05	96.35
Female Original Accuracy	80.00	83.16	86.33
Female New Accuracy	87.34	87.22	92.78

Figure 1: Accuracy table by model, separated by stereotype/sex

Each sentence contains the primary profession as the target profession to be inferred, and accuracy was calculated as the number of professions each LLM inferred correctly from the prompt divided by the total number of sentences. We can see that the new prompting scheme showed great improvements over the original query, with each model performing approximately 4%-11% better for each of the categories (except GPT 3.5 in the pro-stereotype category). Of the three LLMs we picked, Gemini 1.0 Pro had the highest overall original accuracy and highest overall new accuracy while Llama 2 saw the largest improvement between overall original accuracy and overall new accuracy. All three LLMs performed worse in the anti-stereotype categories and female categories than their opposite categories, indicating potential occupational, demographic, or gender bias within these models.

5 Conclusion

Overall, the application of entity-neutral prompting showed promising results in mitigating gender bias, specifically with occupational entities and referring pronouns. The new accuracy values are better than the original accuracy values across all models and categories, indicating that our approach was effective.

The lower accuracy results of anti-stereotype and female compared to pro-stereotype and male category sentences reaffirm our belief of potential bias within LLMs and their training corpora.

Our work was limited to user-sided prompt engineering, which does not address implicit biases in LLMs as these are internal mechanisms of their training data. While users of LLMs can be more aware of potential biases in responses and take steps to mitigate them, a more effective solution is for LLM developers to employ techniques to reduce these biases within the model itself and select appropriate data to train the model on. When these biases are mitigated at the source, a higher baseline accuracy would be established and user prompt engineering would not be a necessity to mitigate social bias but rather a safeguard against it.

References

[1] Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Steering llms towards unbiased responses: A causality-guided debiasing framework, 2024.

[2] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation, 2019.

[3] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods, 2018.