Bias and Mitigation in Large Language Models

Chirag Bellara
Purdue University Fort Wayne
Department of Computer Science
bellc03@pfw.edu

Pakshal Bhandari
Purdue University Fort Wayne
Department of Computer Science
bhanp02@pfw.edu

Abstract

Pretrained language models, particularly masked language models (MLMs), have shown remarkable performance across various natural language processing (NLP) tasks. However, there is evidence to suggest that they may perpetuate cultural biases present in the training corpora, which can lead to harmful representations of marginalized groups. To address this issue, we introduce the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs), a dataset designed to measure social biases in language models against protected demographic groups in the US.

As the field of NLP continues to evolve, this dataset can serve as a valuable benchmark to assess progress towards building less biased models. By leveraging CrowS-Pairs, researchers and developers can work towards creating more inclusive and respectful language models that do not perpetuate harmful stereotypes and biases.

1 Introduction

Large Language Models (LLMs) have revolutionized Natural Language Generation (NLG) tasks, exhibiting human-like text generation capabilities, including dialogue generation. However, their training on vast unsupervised data often overlooks ideological balance, raising concerns about potential biases towards specific extremes. This oversight prompts questions about guiding LLMs toward unbiased generation.

Identifying and quantifying the learned biases enables us to measure progress as we build less biased, models that propagate less harm in their myriad downstream applications. We do reliable quantitative benchmark that measures these models' acquisition of major categories of social biases.The dataset CrowS-Pairs consists of 1508 examples covering nine types of biases, including race, religion, and age. In each pair, a model is presented with two sentences: one that is more stereotypical and another that is less stereotypical.

2 Motivation

Detecting and mitigating biases in LLMs are essential for combating discrimination, protecting individual rights, and ensuring equitable treatment for all members of society.

It fuels stereotypes and dehumanization, particularly against marginalized groups(Dev et al., 2022). LLMs are used to analyze online interactions, detect abuse, and distress, and predict social cues based on demographics. As LLMs shape public discourse and decision-making, addressing biases is crucial to promote fairness and inclusivity in digital spaces (Blackwell et al., 2017), (Guda et al., 2021). Here we evaluate multiple masked language models to evaluate which models are more susceptible to bias.

3 Scope, Results & Analysis

3.1 Past Work

Since the crackdown on the organizations that own the rights to the most used LLMs, several steps have been taken by these LLM companies to ensure as little bias as possible. However, there have been several instances in the past that reported these biases. Because it is impossible to fully eliminate bad behavior from the final fine-tuned model, companies add additional safeguards around how the model is used. These safeguards usually include checking if the user's input is appropriate, and/or checking if the model's output is appropriate. Implementation in software may involve rule-based systems, keyword checking (e.g. looking for swear words or slurs), and/or machine learning models to identify and classify the input and outputs as appropriate.
The LLM companies do not reveal the exact working of these safeguarding mechanisms that are put

in place. “The lack of technical disclosures or reports on jailbreak prevention mechanisms leaves a void in understanding how various providers fortify their LLM chatbot services. [...] The exact methodologies employed by service providers remain a well-guarded secret. We do not know whether they are effective enough." (Deng et al., 2024) The key finding of Deng et al.'s paper, *DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models* (Wang et al., 2024) is that LLMs can be easily misled to produce toxic, biased outputs.
The paper *Gender bias and stereotypes in Large Language Models* explores the issue of gender bias in LLM's. In this paper, the authors use prompts with professional anecdotes to try and prove that the models will assume certain roles to be associated with certain genders all the time. Some of the key findings of this paper include:

- LLMs are 3-6 times more likely to choose an occupation that stereotypically aligns with a person's gender.
- LLMs ignore crucial ambiguities in sentence structure 95% of the time in our study items, but when explicitly prompted, they recognize the ambiguity.
- LLMs provide explanations for their choices that are factually inaccurate and likely obscure the true reason behind their predictions. That is, they provide rationalizations of their biased behavior.

3.2 Dataset Used & Flow of Current Work

We utilize the CrowS-Pairs dataset[1] (Nangia et al., 2020), a comprehensive benchmark designed to gauge the extent of stereotypical biases embedded within masked language models (MLMs). This dataset comprises 1,508 instances spanning nine categories of biases: race/color, gender/gender identity, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status. Each instance consists of a pair of sentences. The initial sentence consistently pertains to a historically disadvantaged group in the United States, while the subsequent sentence portrays a contrasting advantaged group. Notably, the first sentence may either exemplify or contravene a prevalent stereotype. Moreover, the second sentence represents a minimal alteration of the first, with modifications limited to the identifiers of the respective groups.

- *sent_more*: The sentence deemed more stereotypical.
- *sent_less*: The sentence deemed less stereotypical.
- *stereo_antistereo*: Indicates the direction of stereo typicality within the pair. A "stereo" designation signifies that "sent_more" embodies a stereotype associated with a historically disadvantaged group, whereas an "antistereo" classification indicates that "sent_less" challenges a stereotype associated with such a group. In either scenario, the counterpart sentence describes a contrasting advantaged group.
- *bias_type*: Specifies the type of bias evident in the instance.
- *annotations*: Specifies the type of bias evident in the instance.
- *anon_writer*: The anonymized id of the writer.
- *anon_annotators*: The anonymized ids of the annotators.

In Update 2, the primary focus of our work was to quantify how much bias is present in widely used LLMs. To do this, we use the CrowS dataset as mentioned above. The goal is to measure how well can LLMs classify stereotyped and non-stereotypes texts. We use pre-trained models of BERT, RoBERTa, AlBERT, DistilBERT, GPT2, and XLNet to train them on the CrowS dataset and use the prompts to measure the extent of stereotype classification. The score that is used is a metric that calculates the percentage of examples for which the LM prefers the more stereotyping sentence (or, equivalently, the less anti-stereotyping sentence). For a sentence S, let U = {u0, . . . , ul} be the unmodified tokens, and M = {m0, . . . , mn} be the modified tokens such that

$$S = U \cup M$$

The score is the estimate of the probability of the unmodified tokens conditioned on the modified tokens(Nangia et al., 2020),

$$Score = p(U \mid M, \theta)$$

[1]This dataset is public and is available to download on https://github.com/nyu-mll/crows-pairs/tree/master/data

Model Name	Metric Score	Stereotype Score	Anti-Stereotype Score
BERT	60.48	61.09	56.88
RoBERTA	65.45	66.80	57.80
ALBERT	**67.04**	**67.67**	**63.30**
DistilBERT	56.90	57.41	54.13
XLNet	53.05	53.61	50.00

Table 1: Model performances CrowS-Pairs dataset. Higher numbers indicate higher model bias.

3.3 Current Analysis

The results (Table 1) demonstrate that all five language models evaluated exhibit substantial bias to varying degrees. There is generally consistency in the rankings across the different metrics, with XLNet performing the best (least biased) and ALBERT performing the worst (most biased) across all three metrics.

One potential factor contributing to the difference in bias scores could be the varying number of parameters across these models. ALBERT, being a relatively large model with a substantially higher number of parameters compared to the other models, may have acquired and amplified more biases during the pre-training process. The large number of parameters in ALBERT could have enabled it to capture and overfit biased patterns present in the pre-training data more effectively.

However, it's important to note that a higher number of parameters does not necessarily equate to increased bias. Larger models often demonstrate better performance on various downstream tasks due to their increased capacity to learn and generalize from data. Therefore, it is highly possible that despite exhibiting a higher bias score in this particular evaluation, ALBERT may still outperform the other models on specific downstream tasks that require more complex reasoning and understanding.

The trade-off between model size, performance, and bias is a complex issue that requires careful consideration. While smaller models like XLNet may exhibit lower biases, they may also have limitations in terms of their overall performance and ability to handle more challenging tasks. Conversely, larger models like ALBERT, while potentially more biased, could still be valuable in certain applications where their increased capacity for learning and generalization outweighs the need for minimizing biases.

It's crucial to understand that bias is a multifaceted problem, and these results only provide a snapshot of the biases present in these models under the specific evaluation setup and metrics used. Additional analysis, using different datasets, tasks, and bias measurement techniques, may be necessary to gain a more comprehensive understanding of the biases inherent in these language models.

Ultimately, the decision of which model to use in a particular application should consider not only the bias scores but also the specific requirements, constraints, and trade-offs involved. Efforts should be made to mitigate biases through techniques such as fine-tuning, data augmentation, and architectural modifications, while also considering the potential impact of biases on the intended use case and target user groups.

3.4 Upcoming Results & Analysis

For the final update, we plan to add some new ways to better measure the biases in the language models we're working with. This will involve leveraging more techniques and metrics to quantify and analyze different types of biases, such as gender, racial, and ideological biases.

One thing we'll do is test the models using special datasets with examples designed to reveal specific biases, like gender or racial biases. By seeing how the models respond, we can get a clearer picture of what biases exist and how strong they are. We'll also look at methods used in machine learning to uncover hidden biases, like slightly changing the inputs and seeing how that affects the model's outputs.

After thoroughly measuring the biases, we want to find ways to reduce them. Some ideas include:

1. Changing how the models are trained to avoid learning biases in the first place.

2. Using special techniques to "de-bias" the training data before feeding it to the models.

3. Adjusting the models' outputs after they're generated to remove biased content.

4. Modifying the internal design of the models to make them less prone to developing biases.

5. Getting human feedback to help identify and correct biases.

At the end of this project, the goal is to develop language models that are as unbiased and fair as possible, so they don't discriminate or promote harmful stereotypes when used in the real world.

References

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2024. Masterkey: Automated jailbreaking of large language model chatbots. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *CoRR*, abs/2102.00272.

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

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.