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

LLMs, trained on vast amounts of data, can inadvertently learn and perpetuate unfair biases related to gender, race, profession, religion, and political ideologies. The research aims to identify these biases through carefully designed prompts and benchmark datasets, quantify their extent using metrics like toxicity and stereotype scores, and propose techniques to reduce bias while preserving model capabilities. The findings demonstrate that state-of-the-art LLMs exhibit varying degrees of bias, with larger models potentially amplifying biases more effectively. Counterfactual data augmentation and the AI Fairness 360 toolkit are presented as promising bias mitigation approaches, enabling the development of more ethical and responsible language models.

1 Introduction

Large language models (LLMs) have revolutionized the field of natural language processing, enabling remarkable capabilities in text generation, question answering, and language understanding. However, these powerful models, trained on vast amounts of data, can unintentionally learn and perpetuate societal biases present in their training corpus. These biases can manifest as unfair or discriminatory patterns in the model's outputs, potentially leading to harmful outcomes and undermining trust in AI systems.

As LLMs become increasingly prevalent in various applications, addressing the issue of bias has become a crucial ethical and technical challenge. This project aims to shed light on the extent and nature of biases present in state-of-the-art LLMs, quantify their impact, and explore strategies to mitigate these biases while preserving the models' impressive capabilities. By evaluating multiple LLMs and comparing their performances, we aim to gain insights into the factors that contribute to the amplification or reduction of biases, such as model architecture, parameter count, and training data composition. Furthermore, this research explores two promising bias mitigation techniques: counterfactual data augmentation (CDA) and the AI Fairness 360 toolkit.

The pursuit of ethical and responsible AI development stands as a paramount motivation driving this research endeavor. As language models pervade numerous domains, from content generation to decision-making systems, it becomes imperative to identify and address inherent biases that could perpetuate harmful societal stereotypes, discrimination, or unfair outcomes. By proactively tackling bias, we not only uphold ethical principles but also foster trust and confidence among users, paving the way for increased adoption and positive real-world impact of these powerful technologies.

Moreover, bias mitigation in large language models holds the potential for significant competitive advantage and innovation. Organizations that prioritize the development of debiased, inclusive, and trustworthy AI systems can gain a crucial edge in an increasingly AI-driven landscape. Such efforts can lead to more reliable products and services, improved decision-making processes, and a positive reputation for embracing ethical AI practices. Ultimately, this project's motivation stems from a commitment to preventing negative outcomes, promoting user trust and adoption, enhancing AI credibility and reliability, and unlocking the full potential of language models to benefit society as a whole.

2 Related Works

Bias 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 dig-

ital spaces (Blackwell et al., 2017), (Guda et al., 2021).

Previous research has made significant strides in detecting and mitigating biases within language models. A novel approach was proposed in "Mitigating Unwanted Biases with Adversarial Learning in Language Models," (Zhang et al., 2018) which employs adversarial learning techniques to identify and neutralize biased language patterns. Similarly, "Measuring and Reducing Stereotypes in Word Embeddings." (Bolukbasi et al., 2016) address the detection and mitigation of biases in word embeddings.

In addition to bias detection methods, researchers have also explored various techniques for mitigating biases in language models. The concept of counterfactual fairness was introduced in, "Counterfactual Fairness,"(Kusner et al., 2018) which aims to ensure fair treatment of individuals by considering counterfactual outcomes in predictive models. By assessing fairness through hypothetical scenarios, their approach offers insights into mitigating biases in machine learning algorithms and promoting equitable outcomes.

3 Methodology

In this study, we will be exploring three major fronts of bias and mitigation in large language models. For each of these fronts, we will be following certain methodologies and experiments that would give us further insight into the data, the behavior of the models, and the patterns being used to generate this biased text.

I. **Scrutinize:** Finding out if there is any bias.
To detect bias, the methodology involves crafting careful prompts to trigger biased responses from language models, such as using gender stereotypes or sociopolitical stances. These prompts are then evaluated across multiple models, with generated outputs manually inspected for signs of unfair bias.
Here we delve into the detection of bias within large language models, such as ChatGPT, LLaMa, Gemini, and Gemma. By employing tailored prompts engineered to provoke biased responses related to gender, ethnicity, and socio-political stances, this experiment aims to shed light on the inherent biases present in these models. For checking the presence of bias in models we jailbroke multiple large language models like ChatGPT, LLaMa, Gemini, and Gemma. For this, we used specific prompts designed to elicit biased responses related to gender, ethnicity, and political/social stances. We employ "jailbreaking" techniques to bypass the safety constraints and filters of these language models. This allows the models to generate more unconstrained and potentially biased outputs.
The generated outputs from the different models were manually inspected and compared. The goal here was to get the LLMs to give responses that were either biased or radical. We notice that GPT is trained well to avoid responding to potentially biased text. Llama and Gemini do present some radical views for the prompts given. Proceeding with this we know that the goal has to be to minimize the bias in the training data itself.

II. **Quantify:** Measure the bias present.
To quantify the degree of bias present, the methodology involves using benchmark datasets like CrowS-Pairs and BOLD that are carefully curated to test for specific biases like gender, profession, and stereotype associations. Pre-trained language models are evaluated on these datasets by computing metrics such as toxicity scores that measure unsafe/offensive content and bias scores that quantify stereotypical vs. anti-stereotypical associations using techniques like likelihood ratios or log probabilities.

- **Datasets**
  Utilizing diverse bias evaluation datasets across different tasks is important for comprehensively measuring bias in large language models. For our study, we will focus on using two specific datasets – CrowS-Pairs: Crowdsourced Stereotype Pairs (Nangia et al., 2020) for evaluating stereotype associations, and BOLD: Bias in Open-Ended Language Generation Dataset (Dhamala et al., 2021) for analyzing biases in open-ended text generation from language models.
  - **CrowS-Pairs: Crowdsourced Stereotype Pairs**
    The CrowS-Pairs dataset[1] contains 1,508 examples in total. Each example consists of a pair of sentences,

[1] CrowS-Pairs dataset.

Model	Toxicity Score	Professional Bias	Gender Bias
GPT2	0.033	0.222	0.138
DistilGPT2	0.044	0.204	0.158
XLNet	0.393	0.257	0.182
BERT	-	0.057	0.033
DistilBERT	-	0.016	0.042
RoBERTa	-	0.071	0.052

Table 1: Toxicity, professional bias, and gender bias scored of models.

along with several columns/fields providing metadata and annotations about those sentences. The key columns/fields are:

(a) **sent_more** – contains one of the two sentences in the pair.

(b) **sent_less** – contains the other sentence, which is a minimally edited version of the sent_more sentence.

(c) **stereo_antistereo** – indicates whether the sent_more sentence demonstrates a stereotype (stereo) or violates/contrasts a stereotype (antistereo).

(d) **bias_type** – specifies the type of bias present, chosen from race/color, gender/gender identity, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status.

(e) **annotations** – lists the bias type annotations provided by crowd-workers for that example.

(f) **anon_writer** – anonymized ID for the writer who generated the original sentences.

(g) **anon_annotators** – Anonymized IDs for the crowd-workers who annotated the bias types for that example.

– **BOLD: Bias in Open-Ended Language Generation Dataset**
BOLD[2] contains a large set of carefully constructed prompts spanning multiple bias domains, with columns indicating the prompt text, domain, and potential sub-groups. These prompts can be used to analyze biases in how language models complete the open-ended generations.
BOLD contains a total of 23,679 text generation prompts across 5 domains - profession, gender, race, religious ideologies, and political ideologies. Each prompt is a partial sentence or text snippet designed to probe biases when language models continue generating text to complete the prompt. The prompts were collected and constructed using data from Wikipedia articles.
The dataset contains columns/fields for:

(a) The prompt text itself

(b) The domain the prompt belongs to (e.g. gender, race, etc.)

(c) Potentially sub-groups within each domain (e.g. specific professions, religions, etc.)

The number of prompts is not evenly distributed across domains. For example, "Profession" has 10,195 prompts distributed across 18 classes whereas "Religious ideologies" has only 639 prompts across 7 classes.

- **Experiments**
Now that we have established the presence of bias, the next step is to quantify this bias [3]. For this, we use curated bias benchmark datasets to systematically quantify different types of biases present in several widely used large language models. The calculated bias metric scores provided a numerical comparison of the varying degrees of bias across models. For comparison, we use the following pre-trained models, GPT, XLNet,

[2] BOLD Dataset

[3] Bias Detection and Mitigation in LLMs code files.

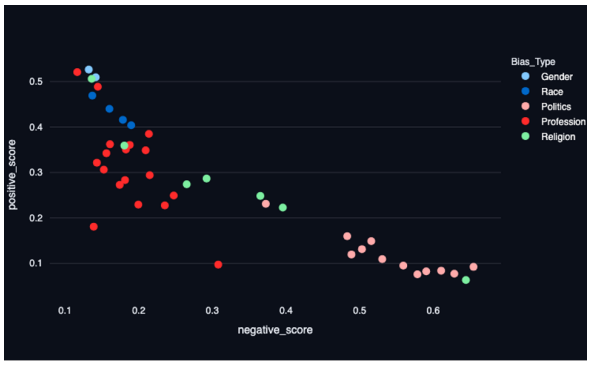

Figure 1: Positive and negative toxicity scores for GPT2 model.

BERT, and variations of the GPT and BERT models. For evaluating these models, we use the following scores,

(a) *Toxicity Score*: Measured using Hugging Face's evaluate library to detect offensive/unsafe content.

(b) *Profession Bias Score*: Specific metric to measure bias towards certain professions.

(c) *Gender Bias Score*: Metric to quantify gender bias

(d) *Stereotype Score*: Calculated using log probability ratios to quantify stereotypical vs anti-stereotypical associations.

The pre-trained models were provided examples from CrowS-Pairs and BOLD datasets. Their outputs were recorded and scored using the above evaluation metrics. As seen in Figure 1 the positive toxicity scores for GPT2 are highest for gender bias while the negative toxicity scores are the highest for political ideologies. Similarly, all the scores for the remaining models are given inTable 1.
We additionally assessed the stereotype metric score (Nangia et al., 2020), a bias metric tailored specifically for the CrowS-Pairs dataset by its creators. Looking at the results in Table 2, ALBERT has the highest metric score (67.04) as well as the highest stereotype score (67.67), indicating it exhibits substantial stereotypical biases in its outputs compared to other models. On the other hand, XLNet has the lowest metric score (53.05), stereotype score (53.61), and anti-stereotype score (50.0), suggesting it displays reduced stereotypical associations relative to other models evaluated. The BERT-based models (BERT, DistilBERT, RoBERTa) demonstrate varying degrees of bias, falling in between the extreme cases of ALBERT and XLNet for these metrics.

III. **Countermeasures:** Provide strategies to minimize, mitigate, and counter the bias present. The key strategies to mitigate biases in large language models involve data augmentation and optimization techniques. Counterfactual data augmentation (CDA) (Mouli et al., 2022) aims to reduce biases by generating counterfactual examples that challenge stereotypical assumptions during fine-tuning. Toolkits like AI Fairness 360 (Bellamy et al., 2018) provide a range of pre-processing, in-processing, and post-processing methods to modify the training data, learning procedure, or model outputs to align with fairness constraints.

- **CDA: Counterfactual Data Augmentation**
  CDA is a data augmentation technique designed to mitigate language model biases. It involves generating modified versions of the training data by making controlled changes while preserving the original meaning of the text. The key idea behind CDA is to present alternative perspectives that challenge the model's preconceived notions or bias tendencies. The process works by identifying and replacing specific demographic mentions or stereotypical associations in the training data with counterfactual examples. For instance, to counter gender bias, CDA can generate sentences describing people in non-stereotypical roles, such as "John, a skilled nurse, gently tended to the elderly patient's wounds." or "The brave firefighter Emily rushed into the burning building to rescue the trapped residents." Similarly, CDA can be used to generate balanced news article excerpts or other text that presents different political or ideological viewpoints, countering potential biases. By exposing the language model to these augmented, coun-

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

terfactual examples during fine-tuning, CDA aims to make the model more robust and less likely to produce biased or unfair outputs.

- **AI Fairness 360**
  AIF360[4] is an open-source library developed by IBM and the research community to help detect and mitigate bias in AI systems, including language models. It provides a comprehensive toolkit with over 70 fairness metrics and a wide range of bias mitigation algorithms. AIF360 offers three main approaches to mitigate bias:
  - Pre-processing: These techniques modify the training data itself to reduce bias. Examples include reweighing training examples, optimized data pre-processing, and data transformations to remove sensitive attributes.
  - In-processing: These methods modify the learning procedure or model architecture to reduce bias during training. Techniques include adding discrimination-aware regularization terms, adversarial debiasing, and constraint-based optimization.
  - Post-processing: These algorithms modify the model's outputs or predictions in a post-hoc manner to align with fairness constraints, treating the model as a black box.

  AIF360 provides a flexible and modular framework, allowing researchers and practitioners to combine different pre-processing, in-processing, and post-processing techniques tailored to their specific use case and fairness requirements. By leveraging the tools and algorithms in AIF360, language model developers can systematically detect and mitigate various types of biases, such as gender, racial, or age biases, in their models' outputs.

4 Analysis

The analysis of bias in language models, particularly in the context of the BOLD dataset prompts, reveals significant disparities among the models evaluated. With its considerable parameter count, ALBERT emerges as the most biased model across all metrics examined. This finding aligns with the hypothesis that larger models, such as ALBERT, may inadvertently amplify biases present in their pre-training data. The extensive parameter count of ALBERT potentially facilitates the capture and exaggeration of biased patterns more efficiently than smaller counterparts.

Conversely, GPT exhibits the least bias among the models assessed. This result suggests that while model size may contribute to bias amplification, other factors, such as architecture and training methodology, also play crucial roles. The superior performance of GPT in mitigating biases underscores the importance of exploring diverse approaches in language model design and training.

Furthermore, the analysis highlights the potential trade-offs between model size, bias mitigation, and task complexity. While smaller models like XLNet demonstrate reduced biases compared to larger counterparts, their performance and capacity to address intricate tasks may be limited. This observation underscores the intricate interplay between model architecture, size, and task suitability in the pursuit of bias mitigation without compromising performance.

Overall, the analysis underscores the multifaceted nature of bias in language models and the need for nuanced approaches to address it effectively.

[4] AI Fairness 360 homepage.

References

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

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).

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings.

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.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

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

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2018. Counterfactual fairness.

S Chandra Mouli, Yangze Zhou, and Bruno Ribeiro. 2022. Bias challenges in counterfactual data augmentation.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning.