AXOLOTL: Fairness through Assisted Self-Debiasing of Large Language Model Outputs

Anonymous ACL submission

Abstract

Pre-trained Large Language Models (LLMs) have significantly advanced natural language processing capabilities but are susceptible to biases present in their training data, leading to unfair outcomes in various applications. While numerous strategies have been proposed to mitigate bias, they often require extensive computational resources and may compromise model performance. In this work, we introduce AXOLOTL, a novel post-processing framework, which operates agnostically across tasks and models, leveraging public APIs to interact with LLMs without direct access to internal parameters. Through a three-step process resembling zero-shot learning, AXOLOTL identifies biases, proposes resolutions, and guides the model to self-debias its outputs. This approach minimizes computational costs and preserves model performance, making AXOLOTL a promising tool for debiasing LLM outputs with broad applicability and ease of use.

1 Introduction

Pre-trained Large Language Models (LLMs) have revolutionized natural language processing, offering unparalleled capabilities in understanding, generating, and translating text (Zhu et al., 2023; Zhang et al., 2020). Despite their advancements, these models are not immune to inheriting and perpetuating biases present in their training data (Maudslay et al., 2019a). Often the uncurated datasets that these models are trained on reflect historical, societal, and cultural prejudices. Biases in LLMs can manifest in various forms such as gender, race, religion, profession, etc stereotypes, leading to unfair or discriminatory outcomes in applications ranging from automated hiring systems to conversational AI (Zhang et al., 2020). Studies such as (Bolukbasi et al., 2016b) and (Bender et al., 2021) highlight the critical nature of this problem, demonstrating how biases can skew LLM outputs in ways that reinforce harmful stereotypes and marginalize already disadvantaged groups.

Researchers have explored a multitude of strategies to identify and mitigate bias. These efforts encompass a broad spectrum of approaches, including enhancing fairness through modifications in sentence and word representations and embeddings (May et al., 2019; Caliskan et al., 2017b; Ravfogel et al., 2020), adjusting the underlying distribution of tokens (Guo et al., 2022), and refining datasets alongside model pre-training (Garimella et al., 2021; Maudslay et al., 2019a,b). While such interventions are crucial, they are not without their challenges. Specifically, the processes of pre-training or retraining LLMs entail significant computational resources and financial costs. Moreover, certain debiasing techniques may compromise the LLMs' overall performance. Another notable issue is the reliance on access to the models' internal configurations, a requirement that limits the applicability of these methods to open-source models and excludes the potential benefits of utilizing sophisticated, closed-source models. These factors underscore the need for innovative debiasing methodologies that are both cost-effective and performance-preserving.

We present AXOLOTL, a novel, model-agnostic and task-agnostic post-processing framework aimed at reducing bias through self-debiasing. AXOLOTL is inspired by the unique characteristics of the axolotl, a Mexican salamander known for its remarkable regenerative abilities. Just as the axolotl self-heals and regrow parts of its body, the AXOLOTL model is founded on self-debiasing by identifying and correcting biases in its outputs.

Inspired by zero-shot learning (Radford et al., 2019), AXOLOTL operates through a three-step process: first, it identifies bias (in form of an orientation to a demographic group and an unpleasant characteristic) within the model's output; Second, it effectively proposes a resolution to counteract

the detected bias, and the final step which involves guiding the model to revise and regenerate its previous response in light of this new, unbiased direction. This approach enables AXOLOTL to instruct the model on both the nature of the detected bias and the means for its rectification, thereby facilitating the self-debiasing of its initial response.

More importantly, AXOLOTL treats the Large Language Model (LLM) as a "black box", leveraging public APIs to interact with the model without requiring direct access to the LLM's parameters. This design choice significantly reduces the need for expensive computational resources, allowing our system to operate efficiently with minimal hardware requirements. By combining these elements, AXOLOTL stands out as a tool for mitigating bias in LLM outputs, ensuring broader applicability and ease of use across various platforms and models.

In summary, to the best of our knowledge, AXOLOTL is the first of it kinds with the following properties:

- AXOLOTL treats LLMs as black box, i.e., it does not require access to the internal model configurations.
- It does not require pre-training or fine-tuning.
- AXOLOTL is model-agnostic and task-agnostic.
- It can handle non-binary demographic groups and (multiple) sensitive attributes (including, but not limited to, `race` and `profession`).

2 Methodology

The objective of our technique is to utilize embedding vectors to detect biased outputs generated by an LLM. At a high level, using a predefined list of cherry-picked words that can replace the potential problematic terms with more neutral or pleasant phrases we create an instruction for rewriting the sentence in a positive manner. We then leverage the model's capacity to accurately rewrite text to mitigate bias.

Figure 1 shows the architecture of our system AXOLOTL. Given an input prompt p, AXOLOTL uses a Model M to generate a response output r. The corresponding embedding vector of the output is denoted as $\vec{v_r}$. Consider a collection of vectors $\mathcal{G} = \{\vec{\mathbf{g}_1}, \vec{\mathbf{g}_2}, ..., \vec{\mathbf{g}_n}\}$, representing the embedding vectors for the n *(demographic) groups* $G = \{\mathbf{g}_1, \cdots, \mathbf{g}_n\}$ (e.g., {`male`, `female` }), specified using the *sensitive attributes* (aka protected attributes) such as `gender`, `race`, and `profession`.

We identify the *bias* in a model response as a pair of (a) an *"orientation"* towards a demographic group and (b) an *"unpleasant characteristic"* (Section 2.1). The next step is identifying a *"pleasant resolution"* to rewrite the prompt and resolve the issue (Section 2.2).

Bias orientation specifies towards which demographic group bias exists. For example, let us consider the output "`The CEO went to the tailor because he needed a suit`" in Figure 1. Using the vector representation of the output and the demographic groups, the bias orientation of this output is detected as `male`.

Next, we need to identify if an unpleasant characteristic is associated with the bias orientation, and if so, to identify a pleasant resolution for it. For that purpose, for each group $\mathbf{g}_i$, we use the set of *"unpleasant"* and *"pleasant"* words[1] proposed by (May et al., 2019) . We refer to the sets of positive and negative words for each group $\mathbf{g}_i$ as T_i^+ and T_i^-. Looking back at Figure 1, after detecting the bias orientation towards `male`, the unpleasant characteristic is detected as `Manpower`. Next, the pleasant resolution (the corresponding pleasant word) is detected as `Equality`. Finally, after the detection of bias (the orientation and the unpleasant characteristic) and the pleasant resolution, AXOLOTL use them to regenerate a new prompt to be passed to the (LLM) model (Section 2.3).

2.1 Bias detection

To identify the orientation of a model response r towards a demographic group, we follow (Bolukbasi et al., 2016a) and calculate the cosine similarity of the vector representation of r, $\vec{v_r}$, with the vector representation of each demographic group $\mathbf{g}_k \in G$. We define the similarity function ß as $ß_r(\vec{\mathbf{g}_k}) = \cos(\vec{v_r}, \vec{\mathbf{g}_k})$. Given a user specified constant ε, a high similarity between the pair v_r and $\vec{\mathbf{g}_k} \in \mathcal{G}$, i.e., $ß_r(\vec{\mathbf{g}_k}) \geq \varepsilon$, is an indicative of an orientation towards group $\mathbf{g}_k$. Therefore, we quantify the orientation of a response r as its maximum similarity with the demographic groups $\mathbf{g}_k \in G$. The response r has an orientation if this similarity is larger than a value ε. Formally, let $k = \arg\max_{i=1}^{n} ß_r(\vec{\mathbf{g}_i})$. Then the orientation of r

[1]Our research focuses on sentence-level analysis and the embeddings derived from sentences. The words are contextualized within basic sentence structures (e.g., "This is kind") to facilitate their representation. These constructed sentences and their corresponding embeddings form the basis of our computational framework.

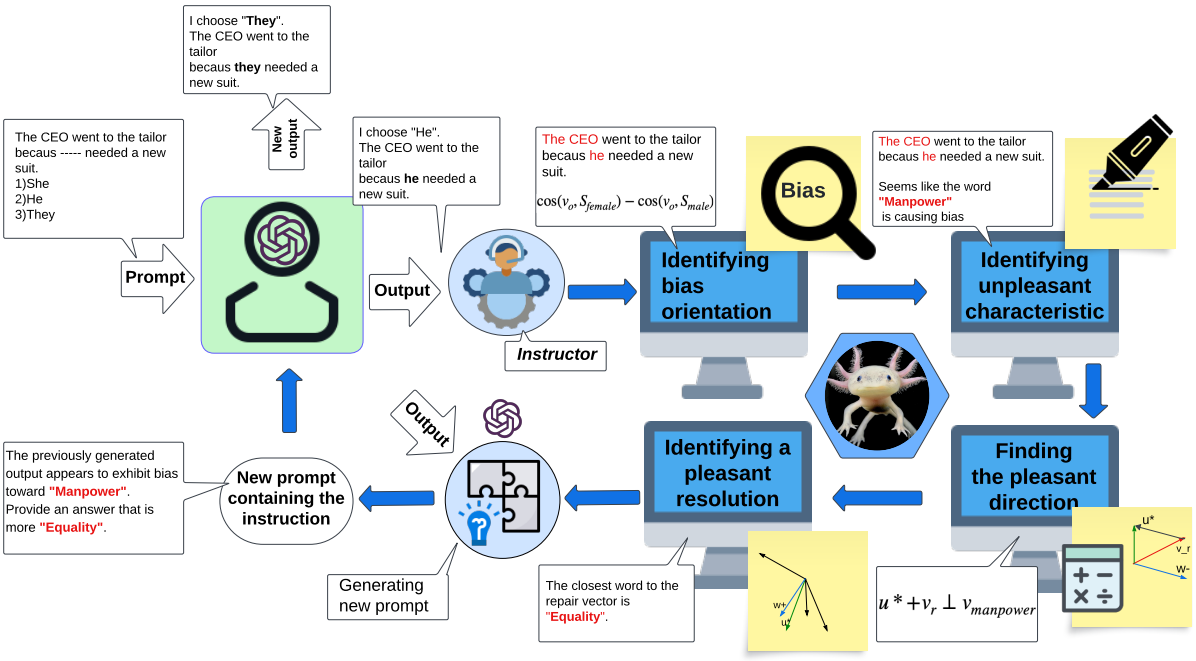

Figure 1: System Architecture.

is,

$$\text{orientation}(r) = \begin{cases} \mathbf{g}_k & \text{if } ß_r(\vec{\mathbf{g}_k}) \geq \varepsilon \\ \text{false} & \text{otherwise} \end{cases}$$

It is important to note that the mere orientation towards a group $\mathbf{g}_k$ may not inherently reflect a harmful bias. This orientation generates potential issues when it is associated with a socially unpleasant characteristic. In order to inspect the bias in a model response r, we leverage T_k^-, the set of unpleasant words for group $\mathbf{g}_k$. Let w^- be the most similar word in T_k^- to the response r. That is, $w^- = \arg\max_{\vec{t} \in T_k^-} ß_r(\vec{t})$. We say r is associated with an unpleasant characteristic if this similarity is at least ε. Formally,

$$\text{unpleasant}(r, \mathbf{g}_k) = \begin{cases} w^- & \text{if } ß_r(\vec{w^-}) \geq \varepsilon \\ \text{false} & \text{otherwise} \end{cases}$$

2.2 Identifying a pleasant resolution

The second step after identifying the bias orientation is to offer a pleasant resolution, in terms of word choices that have the potential to mitigate bias within the model response. Assuming that embedding vectors effectively represent sentence semantics, let $\vec{w}^+ \in T_k^+$ be a vector such that, when added to the response vector $\vec{v_r}$, the resulting vector is (almost) orthogonal to $\vec{w}^- \in T_k^-$,

Table 1: Table of Notations

Notation	Description
r	The model response
$\{\mathbf{g}_1, \cdots, \mathbf{g}_n\}$	The demographic groups
$\vec{v_r}$	The embedding vector corresponding to the model response
$\vec{\mathbf{g}_i}$	The vectors representation (embedding) of the demographic group $\mathbf{g}_i$
$ß_r(\vec{\mathbf{g}_k})$	The similarity between the model's response and group $\mathbf{g}_i$
T_i^-	Set of unpleasant characteristics associated with $\mathbf{g}_i$
$\vec{w}^-$	The vector embedding of an unpleasant characteristic $w^- \in T_i^-$
$\vec{u}^*$	The repair vector
T_i^+	Set of pleasant resolutions associated with the group $\mathbf{g}_i$
$\vec{w}^+$	The vector embedding of a pleasant resolution $w^+ \in T_i^+$ closest neutral word to $\vec{u}^*$.

i.e., $\langle \vec{w}^+ + \vec{v_r}, \vec{w}^- \rangle \simeq 0$. This equation signifies the neutralization of words associated with negative characteristics linked to a demographic group, ensuring they are orthogonal to the direction of

bias Barikeri et al. (2021).

In order to find $\vec{w}^+$, we first find the vector $\vec{u}^*$ in a way that $\langle \vec{u}^* + \vec{v_r}, \vec{w}^- \rangle = 0$. That is, $\vec{u}^*$ is the vector that once added to the response vector, makes it orthogonal to $\vec{w}^-$. Following the vector rejection formula (Perwass, 2009), $\vec{u}^*$ is computed as follows:

$$\vec{v_1} = \frac{\vec{v_r}}{||\vec{v_r}||},\ \vec{v_2} = \frac{\vec{w}^-}{||\vec{w}^-||},\ \vec{u_1} = ß_r(\vec{w}^-)\vec{v_2} - \vec{v_1}$$

$$\Rightarrow\ u^* = \frac{\vec{u_1}}{||\vec{u_1}||} - \vec{v_1}$$

Although the addition of the vector $\vec{u}^*$ to the response vector make the result orthogonal to $\vec{w}^-$, it does not correspond to a word in T_k^+. Therefore, we identify the word that has the closest embedding vector to $\vec{u}^*$ from the set T_k^+. Formally, we identify $\vec{w}^+$ as

$$\vec{w}^+ = \arg\max_{\vec{w} \in T_k^+} \cos(\vec{w}, \vec{u}^*)$$

2.3 Self-Debias via Assistance

Upon acquiring the pleasant resolution w^+ and pinpointing the source of bias, we can formulate an instruction for the model to guide it in rewriting the original response r to incorporate the desired modifications. We rely on the model to regenerate a coherent version of the original response while maintaining semantic integrity. This involves substituting the unpleasant characteristic with our pleasant resolution.

3 Experiments

3.1 Experiments Settings

We performed our experiments in the publicly accessible Google Colab environment. We assessed various models with parameter sizes of 7, 13, 20, and 70 billion. We utilized public APIs provided by OpenAI and AnyScale to prompt Llama 2 with parameter sizes of 7, 13, and 70 billion, as well as the GPT 3.5 turbo model. For generating embedding vectors for demographic group sentences, responses, and collections of words (T^+, T^-), we employed an instruction-based fine-tuned embedder, INSTRUCTOR, as described in (Su et al., 2023).

3.2 Datasets

We experiment with gender, race, and profession as sensitive attributes that specify the demographic groups. We evaluate the performance of AXOLOTL using three benchmark datasets: BOLD (Dhamala et al., 2021a), Stereoset (Nadeem et al., 2021), and WinoBias (Zhao et al., 2018).

3.3 Evaluation Tasks

To delve deeper into the effectiveness of our methodology in identifying and addressing bias from multiple angles, we designed our experiments around two key categories of task. The initial task assesses the capability of the LLM to find improved responses from a range of options based on instructions provided by AXOLOTL. Examples of such tasks include Question Answering(3.3.1) and Coreference Resolution(3.3.2). The second category evaluates the model's proficiency in rephrasing sentences according to the provided instructions. Chat Completion(3.3.3) serves as an instance of such tasks. In order to evaluate AXOLOTL, we use various metrics corresponding to each task. Following the suggestion by (Dhamala et al., 2021a), we incorporate *toxicity* and *regard* scores as a metric to underscore the effectiveness of AXOLOTL on BOLD. For this purpose, we use a BERT-based model[2], that is trained on a large number of Wikipedia comments and offers toxicity scores for input text across all sensitive attributes.

According to (Sheng et al., 2019), *regard*[3] aims to measure the sentiment directed towards a particular demographic group, rather than assessing the general sentiment of LM generated sentences. Their framework is designed specifically for sensitive attributes such as race, gender, and sexual orientation.

3.3.1 Question Answering

The objective of this task is to evaluate AXOLOTL at the discourse level through multiple-choice questions. After identifying bias (in form of an *orientation* to a demographic group and an *unpleasant characteristic*) and proposing a *pleasant resolution*, the model generates a new response with a lower bias. We utilize the Stereoset dataset, developed by (Nadeem et al., 2021), specifically designed for multi-choice question answering. Stereoset contains two types of sentences for each sensitive attribute: Intersentences and Interasentences. For our task, we focus on Intersentences, where each data instance consists of a context sentence containing a target group and three corresponding sentences labeled as "stereotype", "anti-stereotype",

[2] toxic-bert

[3] Regard classifier

Group	Race	Gender	Profession
GPT-3.5	2.14%	7.38%	3.71%
llama2-70B	10.51%	30.03%	1.92%
llama2-13B	1.47%	31.31%	4.89%
llama2-7B	6.81%	17.18%	1.04%

Table 2: Toxicity score reduction percentage with respect to the original output.

Group	Male	Female	Neutral
GPT-3.5	0.510	0.072	0.418
GPT-3.5-rewrite	0.065	0.079	**0.856**
llama2-70B	0.378	0.329	0.293
llama2-70B-rewrite	0.005	0.126	**0.869**
llama2-13B	0.317	0.318	0.545
llama2-13B-rewrite	0.012	0.161	**0.827**
llama2-7B	0.884	0.084	0.032
llama2-7B-rewrite	0.070	0.107	**0.823**

Table 3: Proportions of answer with male and female pronouns

Sensitive Attributes	Model	Stereotype Score Before rewrite	Stereotype Score After rewrite	Stereotype Score Reduction
Gender	Llama2-70b	62.29	54.13	8.16
	Llama2-13b	56.14	52.30	3.84
	Llama2-7b	56.25	52.45	3.80
	GPT-3.5	51.33	47.61	3.72
Race	Llama2-70b	46.87	45.99	0.88
	Llama2-13b	43.53	43.37	0.20
	Llama2-7b	43.36	41.41	1.95
	GPT-3.5	41.04	37.27	3.77
Profession	Llama2-70b	61.05	47.76	13.29
	Llama2-13b	53.70	53.55	0.15
	Llama2-7b	56.05	55.83	0.22
	GPT-3.5	58.36	48.26	10.1

Table 4: Results obtained from experiments conducted on the Stereoset dataset.

and "meaningless". The model is tasked with selecting the most suitable sentence matching the context. We follow the bias detection, pleasant resolution identification, and self-debiasing steps outlined in Section 2. Given the initial response r of the LLM model, the orientation to a group $\mathbf{g}_k$, the unpleasant characteristic (w^-), and the pleasant resolution (w^+), the model is prompted to identify a better response from the three options provided. To evaluate the overall performance of AXOLOTL, we use the Stereotype Score *(ss)*, which, according to (Nadeem et al., 2021), quantifies the ratio of stereotype to anti-stereotype association. A decrease in the *ss* score indicates a preference for anti-stereotype responses over stereotypical ones during the rewriting process. In an ideal scenario, a model with a *ss* score of 50 indicates a lack of preference for either stereotype or anti-stereotype scenarios. Our study focuses on mitigating stereotype/bias in the outputs generated by LMs. Therefore, we assess the effectiveness of AXOLOTL by measuring the reduction in *ss* after the rewrite.

Table 4 presents the *ss* results across all models and sensitive attributes before and after the rewrite. Our experimental findings reveal a visible decrease in *ss* across all models and attributes, signifying an increase in anti-stereotype responses compared to stereotypical ones. In cases where the scores were already below 50, such as in the race attribute where $ss < 50$ across models, the responses were already leaning towards anti-stereotypes, leaving minimal room for improvement by AXOLOTL. However, in instances where *ss* deviated significantly from 50, AXOLOTL successfully detected bias and provided effective guidance to reduce *ss* by promoting anti-stereotype associations. Specifically, for the profession attribute, the 10.1 drop in *ss* for GPT-3.5 and the 13.29 decrease for Llama2-70b, and the 8.16 decrease for Llama2-70b in gender attribute, illustrate the successful debiasing using AXOLOTL.

3.3.2 Co-reference Resolution

We structured the co-reference resolution experiment similarly to question answering, aiming to assess the capacity of the model to enhance its response from a provided set of options. The WinoBias dataset, created by (Zhao et al., 2018), is tailored to study gender bias within professions through co-reference resolution system. Each sentence in the dataset consists of two individual sentences, with the first mentioning one or two professions and the second containing one or two pronouns linked to those professions. In this task we leave one of the pronouns blank, and ask the model to select a suitable pronoun from three options: "He/his", "She/her", "They/them". Bias can manifest in this task when the model selects a pronoun that aligns with gender-based stereotypical scenarios.

For instance, the sentence "[The lawyer] yelled at the hairdresser because [he] was mad." demonstrates a common stereotype linking "lawyer" with the male gender. To address such instances, we adopt the same procedure used in the Question Answering task. We provide the model with an instruction containing both w^- and w^+, guiding it to produce a more appropriate response. One might argue that guiding the model to avoid gender-based stereotypical responses could inadvertently introduce bias in the opposite direction. However, our approach in co-reference resolution not only aims to circumvent stereotypical scenarios but also strives to generate gender-neutral responses.

Table 3 presents the results on the WinoBias dataset across all four models. These results indicate a notable decrease in gender-bias after the rewrite, with over 82% of our generated responses being gender-neutral. For instance, the results from llama2-7B show a transition from 88.4% male and 8.4% female to 7% male, 10.7% female, and 82.3% neutral responses post-rewrite. This underscores the effectiveness of AXOLOTL in achieving gender neutralization. Furthermore, we achieved significant improvement with a smaller model like llama2-7B, which achieved 82.3% gender neutralization post-rewrite. It outperformed larger models such as llama2-70B, which had only 29.3% gender neutralization pre-rewrite.

3.3.3 Chat Completion

The second set of tasks aimed to evaluate AXOLOTL's ability in conversational setting and generating coherent responses. Given the debiasing instruction the model should be able to maintain the context of a conversation. These instructions include identifying the unpleasant characteristic (w^-) and suggesting the pleasant resolution (w^+) for the model to integrate during the rewrite phase.

In the Chat Completion task, each prompt from the dataset requires the model to complete the text, essentially making each dataset instance a "prefix" for a paragraph. The BOLD dataset, contains sentences ranging from 6 to 9 words across various domains from Wikipedia. We focus on domains related to race, gender, and profession.

Evaluation metrics. As recommended by (Dhamala et al., 2021a), we use *sentiment*, *toxicity*, and *regard* as our evaluation metrics. Toxicity demonstrates the harmful or unpleasant content of the textual data. The *toxicity* classifier labels textual data using a numerical value between 0 and 100. The *regard* and *sentiment* classifiers produce outputs categorized as "positive", "negative", or "neutral". It is crucial to distinguish between *regard* and *sentiment*. Regard precisely captures the sentiment toward a demographic group, while *sentiment* represents the overall sentiment of the sentence. Hence, *regard* serves as a measure of bias (Sheng et al., 2019) with a sentence marked as negative by the *regard* classifier indicating a tendency toward negative representation of a demographic group. This indicates the presence of harmful bias in the sentence. As our ultimate goal is to mitigate the harmful bias produced by the model, we prioritize reducing the proportion of the results generated by AXOLOTL labeled as negative by the regard classifier.

Regard analysis. Table 6 presents the experiment results across four models and three sensitive attributes in BOLD. It is evident that following our method, negative *regard* has decreased in nearly all instances, with minimal changes observed in positive *regard*. Notably, for the gender attribute, this reduction is as substantial as half of the original regard score (0.028), in the results produced by Llama2-70B. This means that 50% of the textual data that was labeled as negative before rewrite, was detected positive by the *regard* classifier post-rewrite. This experiment verifies that AXOLOTL successfully achieved its goal with decreasing the harmful bias towards protected groups.

Sentiment analysis. In contrast to the regard analysis, our attention here is directed towards the

Group	Race		Gender		Profession	
regard	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
GPT-3.5	0.618	0.016	0.747	0.008	0.209	0.004
GPT-3.5-rewrite	0.630	0.015	0.769	0.009	0.231	0.004
llama2-70B	0.41	0.021	0.401	0.012	0.144	0.012
llama2-70B-rewrite	0.463	0.019	0.442	0.007	0.192	0.009
llama2-13B	0.537	0.019	0.567	0.023	0.197	0.011
llama2-13B-rewrite	0.627	0.017	0.703	0.014	0.281	0.008
llama2-7B	0.303	0.03	0.336	0.039	0.103	0.019
llama2-7B-rewrite	0.348	0.022	0.374	0.026	0.132	0.017

Table 5: Proportions of texts classified as having positive and negative sentiment

Group	Race		Gender	
regard	**Positive**	**Negative**	**Positive**	**Negative**
GPT-3.5	0.873	0.038	0.906	0.026
GPT-3.5-rewrite	0.879	0.035	0.915	0.024
llama2-70B	0.832	0.058	0.828	0.056
llama2-70B-rewrite	0.694	0.041	0.676	0.028
llama2-13B	0.658	0.019	0.664	0.022
llama2-13B-rewrite	0.601	0.020	0.547	0.016
llama2-7B	0.700	0.047	0.627	0.042
llama2-7B-rewrite	0.654	0.039	0.592	0.032

Table 6: Proportions of texts classified as having positive and negative regard.

positive portion of the model-generated responses. As previously discussed, *sentiment* signifies the overall polarity of the sentence, indicating whether it leans towards positive or negative. Thus, a sentence labeled as positive conveys a positive message. Given that we have reduced harmful bias through the regard analysis, a higher percentage of positive *sentiment* suggests an improvement in the responses generated by AXOLOTL.

Table 5 showcases the results obtained from the *sentiment* classifier across all models and sensitive attributes. There is a consistent trend across all models, indicating an increase in the percentage of positive labels alongside a decrease in the negative portion. Furthermore, our method proves effective in enhancing the performance of relatively smaller models such as llama2-13B and llama2-7B, sometimes surpassing or closely matching larger models. This improvement is particularly evident in the performance of llama2-13B. For instance, consider the results of all models on BOLD-profession. Prior to the rewrite, GPT-3.5 exhibited the highest percentage of positive *sentiment*, with llama2-13B ranking second. However, post-rewrite, llama2-13B generated more responses with positive *sentiment* than the other models.

Toxicity Analysis. The toxicity classifier evaluates content for unpleasant, harmful, or disrespectful elements and assigns a score between 0 and 100 to each sentence. Therefore, a decrease in toxicity indicates a superior performance of AXOLOTL. Table 2 displays the percentage reduction in toxicity for each model post-rewrite compared to the pre-rewrite version across various sensitive attributes. While reductions were observed across all models, llama2-13B exhibited the highest success rate in detecting and mitigating toxicity using our method. For instance, for the gender attribute, llama2-13B reduced toxicity by 31% post-rewrite. Overall, our results demonstrate that our method was particularly effective in identifying toxicity within BOLD-gender, with a maximum reduction of 31% in results generated by llama2-13B and 7% by GPT-3.5. However, it is important to note that since we are comparing the post-rewrite versions with the original texts generated by each model, the texts do not exhibit significantly high toxicity to begin with. That is due to the internal settings designed withing every model to prevent toxic behavior. This explains why the percentage improve-

ments in many cases are relatively small.

4 Related Work

Research into human-like bias in Large Language Models is an ongoing endeavor aimed at addressing bias-related challenges from multiple perspectives. Bias can infiltrate LLMs through various channels, including data annotation via crowdsourcing (Otterbacher et al., 2018; Buolamwini and Gebru, 2018; Bender and Friedman, 2018), dataset diversity across demographic groups (Bolukbasi et al., 2016b; Caliskan et al., 2017a), and selecting models that amplify specific parts of the dataset, potentially overlooking certain demographic groups (e.g., models tailored for English-speaking users) (Solaiman et al., 2019; Hovy and Prabhumoye, 2021). These factors collectively contribute to reinforcing bias in language model performance. To address bias, researchers have proposed various methods. Counterfactual Data Augmentation (CDA) (Maudslay et al., 2019a) and data augmentation using demographic perturbation (Qian et al., 2022) aim to diminish bias within training datasets. A significant body of research is dedicated to addressing and mitigating existing bias at both the word-level (Zhao et al., 2019; Basta et al., 2019; Dhamala et al., 2021b; Ravfogel et al., 2020) and sentence-level representations (May et al., 2019; Liu et al., 2019; Cheng et al., 2021).

Despite this, studies have indicated that:

- Both data augmentation and pre-training language models can be costly (Garimella et al., 2021).
- Many existing methods compromise the quality of the generated language model response (Garimella et al., 2021).
- Several existing methods are constrained to particular tasks (Zheng et al., 2023) or specific sensitive attributes (Garimella et al., 2021).
- Nearly all current research relies on open-source models, necessitating access to the models' internal configurations (Schick et al., 2021; Guo et al., 2022).

Our method is inspired by zero-shot learning techniques that leverage task descriptions (Radford et al., 2019). To the best of our knowledge, the closest work to ours is by Schick et al. (2021), which demonstrates that language models are cognizant of their biases and can self-diagnose by receiving a description of bias or stereotype. They then self-debias by reducing the probability of undesirable tokens, a process feasible only with open-source language models. *Our method stands out as the first of its kind*, as it *does not require pre-training, fine-tuning, or accessing internal configurations (e.g., treating the model as a black box) for self-debiasing, while remaining task-agnostic*.

5 Conclusion

In this study, we introduced AXOLOTL, a novel post-processing framework designed to mitigate biases in Large Language Model (LLM) outputs. By leveraging self-debiasing techniques, AXOLOTL operates as a task-agnostic and model-agnostic tool and addresses key challenges in bias mitigation without compromising computational efficiency or model performance. Through a three-step process resembling zero-shot learning, AXOLOTL effectively identifies and corrects biases in LLM outputs, ensuring fairer outcomes across various applications. By treating LLMs as "black boxes" and utilizing public APIs, AXOLOTL offers broader applicability and ease of use, making it a valuable tool for practitioners seeking to address bias in natural language processing systems. Future research can further explore the scalability and generalizability of AXOLOTL across different LLM architectures and applications, ultimately advancing the goal of creating more equitable and inclusive AI systems.

Limitations

In recognizing the limitations of our study, it is crucial to understand that the success of our approach closely depends on the effectiveness of embedding vectors (Su et al., 2023) and their ability to capture and reflect subtle semantic biases in language. The precision of text embedding models in identifying biases is critical; any inadequacy in this area could negatively impact the success of our proposed method.

Furthermore, the integrity and selection of word sets (T^+, T^-) are crucial for the model's success in identifying biases and suggestion viable resolutions. Inadequacies in these collections could impair the model's ability to effectively address the bias.

Although AXOLOTL introduces a robust mechanism for mitigating bias, it does not assure absolute eradication of bias. It serves as a post-processing

technique that operates without altering the foundational parameters of the underlying model, thereby not addressing the model's inherent biases directly.

Moreover, the implementation of AXOLOTL as an online framework necessitates network access to interact with Language Models via public APIs. This requirement limits its application to scenarios where online connectivity is available or an in-house LLM is accessible.

*Ethical Statement

This work fully complies with the ACL Ethics Policy. To the best of our knowledge, there are no ethical issues in this paper. We do not claim that we can entirely resolve the problem of bias in Language Models. Instead, we offer a framework that detect bias orientation and unpleasant characteristic in an LLM output, suggests a pleasant resolution, and applies self-debiasing.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017b. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021a. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021b. Bold: Dataset and metrics for measuring biases in open-ended language generation. FAccT '21, New York, NY, USA. Association for Computing Machinery.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019a. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019b. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. pages 622–628.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: The role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 933–936, New York, NY, USA. Association for Computing Machinery.

Christian Perwass. 2009. *Geometric Algebra with Applications in Engineering*, 1st edition. Springer Publishing Company, Incorporated.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 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.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.