EDITBIAS: Debiasing Stereotyped Language Models via Model Editing

Anonymous ACL submission

Abstract

Previous studies have established that pre-trained language models inherently manifest various biases. Although several debiasing strategies, such as fine-tuning a model with counterfactual data, prompt tuning, and representation projection, have been introduced, they often fall short of efficiently unlearning bias or directly altering the models' biased essence. To address these issues, we propose EDITBIAS, an efficient model editing method to remove stereotyped bias from language models with small editor networks. We design a debiasing loss to guide editor networks to conduct local edits on partial parameters for debiasing, and a remaining loss to preserve the original language modeling abilities of models during editing. Experiments demonstrate the high effectiveness and robustness of EDITBIAS on eliminating bias compared to classical debiasing baselines. Additionally, we explore the effects of bias and debiasing on language models, finding that it is challenging to debias larger and causal language models, and necessary to balance the trade-off between debiasing efforts and language modeling abilities when designing debiasing strategies. [1]

1 Introduction

In recent years, many studies have underscored the propensity of pre-trained language models (PLMs) to have social or stereotypical biases (Liang et al., 2021; Smith et al., 2022; Cheng et al., 2023a; Liu et al., 2023), such as gender bias (Sun et al., 2019; Zhao et al., 2020), race bias (Halevy et al., 2021), among others. To ensure fairness and accuracy in language models' applications, it is crucial to eliminate biases from models.

Numerous studies present various methods to mitigate bias. Some methods (Zmigrod et al., 2019; Barikeri et al., 2021) fine-tune the entire models with counterfactual data obtained by swapping out bias attribute words[2], which is slightly effective and resource-intensive, especially for large language models. Others implement debiasing with representation projection (Dev et al., 2021; Limisiewicz and Marecek, 2022; Iskander et al., 2023) or prompting (Sheng et al., 2020; Abid et al., 2021; Mattern et al., 2022; Venkit et al., 2023). For instance, SentenceDebias (Liang et al., 2020) debias sentence representations by subtracting their projection onto an estimated demographic bias subspace. Ravfogel et al. (2020) introduces Iterative Null-space Projection (INLP), a method that reduces bias in word embeddings by iteratively projecting them onto the null space of bias terms using a linear classifier. Self-Debias (Schick et al., 2021) prompts a model to scale down the probabilities of toxic tokens. However, without internal parameter modification, a model remains biased essentially and is not off-the-self for application.

An ideal debiasing approach is expected to remove bias from PLMs. Model editing (Yin et al., 2023; Zhang et al., 2024) can change specific information in PLMs by modifying partial parameters, which infers that model editing can efficiently eliminate bias. There are three kinds of editing methods:

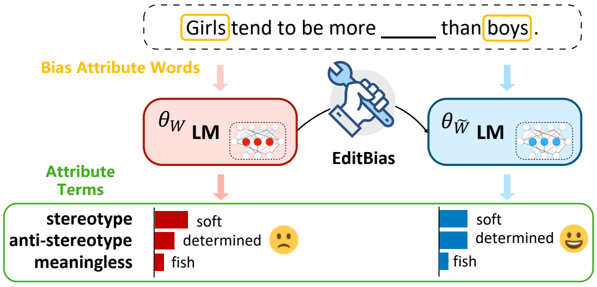

Figure 1: Debiasing a language model with EDITBIAS

[1]Code and data will be released.

[2]The bias attribute word refers to specific features or characteristics that introduce or reflect bias. For example, bias attribute words for gender bias are she, he, mother, father, and the alike. Bias attribute words for religion are Christianity, Judaism, Islam, and so on.

i) fine-tuning a model with new data (Zhu et al., 2020; Ni et al., 2023), *ii*) locating before editing (Meng et al., 2022, 2023; Dai et al., 2022; Wu et al., 2023b) *iii*) utilizing editor hyper-networks to modify PLMs' parameters (Cao et al., 2021; Mitchell et al., 2022a; Cheng et al., 2023b; Tan et al., 2023). On one hand, fine-tuning consumes computational resources and data a lot and is not suitable for large language models. According to our pre-experiments in Appendix A and Chang et al. (2023); Hase et al. (2023a), information, like knowledge and bias can not be simply interpreted as located neurons. On the other hand, small editor hyper-networks can be flexibly applied to any language model and adaptively designed to conduct any specific editing task. Thus, we introduce debiasing PLMs via model editing with editor hyper-networks in this paper.

To overcome the aforementioned shortcomings in previous debiasing methods, **EDITBIAS**, a lightweight model editing method to debias stereotyped language models, is proposed as shown in Figure 1. EDITBIAS uses editor networks to modify a small portion of the parameters, allowing the edited model to be directly deployable for applications. A symmetric debiasing loss is designed to teach the editors how to modify LMs for treating stereotypical and anti-stereotypical contexts. EDITBIAS also contains a retaining loss to avoid affecting unrelated associations during editing for preserving PLMs' modeling abilities. To demonstrate the effectiveness and robustness of EDITBIAS, we conduct experiments on StereoSet (Nadeem et al., 2021) with both masked language models and causal language models compared to four different classical debiasing baselines. The results show that EDITBIAS achieves the best performance on debiasing than all baseline methods and is robust to gender reverse and semantic generality. Furthermore, we thoroughly explore the effects of bias and the process of debiasing on language models. We find that debiasing large and causal language models poses significant challenges and highlight the necessity to balance the trade-off between the effectiveness of debiasing and maintaining language modeling performance, shedding light on future debiasing works.

2 Related Work

Bias and Debiasing Many works focus on measuring bias in language models, such as societal bias (Nangia et al., 2020; Nadeem et al., 2021; Cao et al., 2022; Wan et al., 2023), cultural bias (Zheng et al., 2022; Naous et al., 2023), and multilingual bias (Zhao et al., 2020; Vashishtha et al., 2023), which provide bias measurement metrics (Hovy and Prabhumoye, 2021; Goldfarb-Tarrant et al., 2023). To mitigate bias, researchers propose various debiasing methods (Meade et al., 2022; Gallegos et al., 2023). The basic method is to fine-tune language models on counterfactual data (Lu et al., 2020; Zmigrod et al., 2019), which is costly. Except for fine-tuning, prompting (Schick et al., 2021; Guo et al., 2022) guides models to calibrate their bias. Representation projection (Liang et al., 2020; Ravfogel et al., 2020) is employed to remove bias representation out of models, which, however, cannot change the PLMs' internal bias in essence without modifying parameters. Therefore, we adopt efficiently editing partial parameters for debiasing.

Model Editing As the real world develops, some facts become obsolete and different over time. It is necessary to change, add, or erase facts stored in existing PLMs (Petroni et al., 2019; Shin et al., 2020; Li et al., 2022; Hase et al., 2023b). Model editing (Sinitsin et al., 2020) is come up with to modify information in PLMs. Editing should follow some properties (Yao et al., 2023): reliability (predicting updated facts), locality (keeping accurate on irrelevant facts), generality (editing neighboring facts without specific training), and efficiency (Mitchell et al., 2022a) (efficient in runtime and memory). The direct but inefficient editing is to finetune the whole model on new facts (Zhu et al., 2020). For locality, Dai et al. (2022); Meng et al. (2022, 2023); Ma et al. (2023a) seek the model parameters strongly related to the facts and then edit these localized hidden states. With high efficiency, edited models can be produced without changing their parameters by leveraging extra memories (Mitchell et al., 2022b) and in-context learning (Zheng et al., 2023). Also, Cao et al. (2021); Mitchell et al. (2022a) achieve fast editing by training specific editor networks. Recently, model editing methods have been applied to unlearn information from language models (Chen and Yang, 2023; Patil et al., 2023; Ishibashi and Shimodaira, 2023; Yu et al., 2023). Inspired by them, we propose an efficient model editing method EDITBIAS to unlearn bias in language models while preserving the language modeling capability and generalizing semantically related inputs.

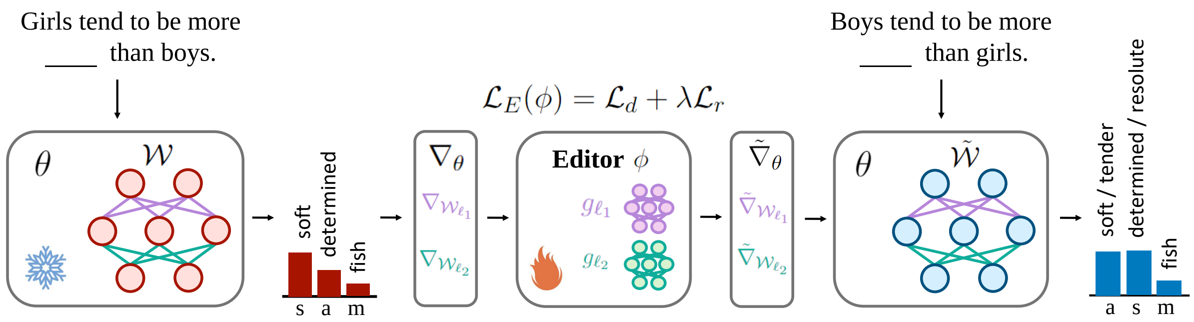

Figure 2: Debiasing a LM with EDITBIAS. Editor networks ϕ are trained to produce edits on partial parameters $\mathcal{W}$ of a LM. After editing, an unbiased PLM is obtained with the robustness of gender reverse and semantic generality. $\mathcal{L}_d$ and $\mathcal{L}_r$ refer to Equation 2 and 3 respectively. s: stereotyped. a: anti-stereotyped. m: meanless.

3 EDITBIAS

3.1 Task and Dataset

A stereotyped model is defined as a language model that exhibits stereotypical bias, such as stereotypes of generic opinions towards different demographic groups in society (Devine, 1989; Nangia et al., 2020; Bauer et al., 2023). In this paper, we study to eliminate stereotypical bias in pre-trained language models while retaining their language modeling abilities during debiasing. An ideal unbiased language model will model stereotypical contexts and anti-stereotypical contexts with the same probability. Therefore, given a biased pre-trained language model with parameters θ, the debiasing task aims to minimize its probability difference between the stereotypical context and the anti-stereotypical context. Furthermore, it is necessary to make sure that general language modeling abilities are not hurt during debiasing (Nadeem et al., 2021; Meade et al., 2022; Ma et al., 2023b; Chintam et al., 2023).

We use the intrasentence set[3] in this paper. For each instance $s \in \mathcal{S}$, there is a context sentence x with a blank (e.g., "Girls tend to be more ___ than boys.") as shown in Figure 1. When three attribute terms corresponding to stereotypical, anti-stereotypical, and meaningless associations (e.g., "soft", "determined", and "fish") fill in the blank in x, three target sentences $x_{\text{stereo}}, x_{\text{anti}}, x_{\text{mless}}$ are formed respectively as

x_{stereo}: Girls tend to be more <u>soft</u> than boys.

x_{anti}: Girls tend to be more <u>determined</u> than boys.

x_{mless}: Girls tend to be more <u>fish</u> than boys.

The optimization target of the debiasing task can be denoted as

$$l_d(x_{\text{stereo}}, x_{\text{anti}}, \theta) = \text{KL}(P_\theta(\cdot|x_{\text{stereo}}) \| P_\theta(\cdot|x_{\text{anti}})) + \text{KL}(P_\theta(\cdot|x_{\text{anti}}) \| P_\theta(\cdot|x_{\text{stereo}})) \quad (1)$$

For masked language models, P_θ is the average per-token log probability of the attribute term that fills the blank in x. For causal language models, P_θ is the average log probability of all tokens in target sentence $x_{\text{stereo/anti-stereo/mless}}$ following Nadeem et al. (2021). Meanwhile, to maintain language modeling capabilities, we hope $P_\theta(\cdot|x_{\text{mless}})$ is unchanged during debiasing.

3.2 Debising via Model Editing

According to Section 1, to conduct effective and efficient debiasing, we propose **EDITBIAS**, a model editing method to debiasing stereotyped LMs as shown in Figure 2.

EDITBIAS adopts lightweight model hyper editor networks ϕ to conduct debiasing edits on PLMs' partial weights $\mathcal{W}$, following Cao et al. (2021); Mitchell et al. (2022a); Tan et al. (2023). A pre-trained language model represents inputs $\mathcal{X}$ as $P_\Theta(\mathcal{X})$. A model editor for debiasing is a function: $(\mathcal{X}_{\text{stereo}}, \mathcal{X}_{\text{anti}}) \times \mathcal{L} \times \Theta \times \Phi \to \Theta$, which maps an stereotypical input x_{stereo} and its corresponding anti-stereotypical input x_{anti}, loss function $l_d : (\mathcal{X}_{\text{stereo}}, \mathcal{X}_{\text{anti}}) \times \Theta \to \mathbb{R}$, biased pre-trained language model parameters $\theta_{\mathcal{W}}$, and editor paramters ϕ to new unbiased model parameters $\theta_{\tilde{\mathcal{W}}}$. The input to an editor network g_ℓ is the fine-tuning gradient $\nabla_{\mathcal{W}_\ell} l_d(x_{\text{stereo}}, x_{\text{anti}}, \theta)$ at the layer $\ell, \ell \in \{1, L\}$. The editor network will output the layer's parameter edit $\tilde{\nabla}_{\mathcal{W}_\ell}$, which is helpful to eliminate bias, to update $\mathcal{W}_\ell$. To be specific, EDITBIAS uses a debiasing training set $\mathcal{S}_{\text{edit}}^{\text{train}}$ and a development set $\mathcal{S}_{\text{edit}}^{\text{dev}}$ to learn parameters ϕ_ℓ for each of the editor network g_ℓ. They are initialized as ϕ_0 at the time

[3]Following Meade et al. (2022); Yu et al. (2023), we utilize only the *intrasentence* portion in StereoSet, which generally adapts to the debiasing task and various language models.

step 0. The partial weights $\mathcal{W}$ (e.g., the weights of the last three layers) we would like to edit are selected before training. At the time step $t-1$, an edit is conducted by ϕ and produces parameter updates $\tilde{\mathcal{W}} \leftarrow EDIT(\theta_{\mathcal{W}}, \mathcal{W}, \phi_{t-1}, x_{\text{stereo}}, x_{\text{anti}})$ with the rank-1 gradient decomposing from Mitchell et al. (2022a). Then editable weights are modified by $\tilde{\mathcal{W}}_\ell = \mathcal{W}_\ell - \alpha_\ell \tilde{\nabla}_{\mathcal{W}_\ell}$ for the layer ℓ, which is back-propagated into g_ℓ. We design two training losses for EDITBIAS using the edited weights $\tilde{\mathcal{W}}$ to teach editor networks how to conduct edits on $\mathcal{W}$. One is a **debiasing loss**:

$$\begin{aligned}\mathcal{L}_d = \text{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{stereo}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{anti}})) \\ + \text{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{anti}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{stereo}}))\end{aligned} \tag{2}$$

Debiasing aims to make a language model equally treat the stereotypical contexts and anti-stereotypical contexts for fairness according to Section 3.1, which is different from knowledge editing. Thus, we design $\mathcal{L}_d$ as symmetric KL divergence losses to guide editor networks to modify $\mathcal{W}$ for debiasing. Moreover, to avoid negative effects on the language modeling abilities, another loss is a **retaining loss** designed to keep the probability of meaningless terms unchangeable during editing:

$$\mathcal{L}_r = \text{KL}(P_{\theta_{\mathcal{W}}}(\cdot|x_{\text{mless}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{mless}})) \tag{3}$$

The total **training loss** of EDITBIAS is $\mathcal{L}_E(\phi_{t-1}) = \mathcal{L}_d + \lambda \mathcal{L}_r$. At the training step t, ϕ is updated by an Adam optimizer (Kingma and Ba, 2015) , which is denoted as $\phi_t \leftarrow \text{Adam}(\phi_{t-1}, \nabla_\phi \mathcal{L}_E(\phi_{t-1}))$. For evaluation, model editors produce debiasing edits on a held-out set $\mathcal{S}^{te}_{edit}$. Because the effectiveness of instance-editing, using one instance in each editing operation, is limited (Cao et al., 2021; Meng et al., 2022, 2023; Ma et al., 2023a; Gu et al., 2024), EDITBIAS adopts batch-editing, using one batch samples in one edit for the debiasing scenario. During training and testing, the same batch size is used for optimal debiasing performance.

4 Experiments

This section elaborates on experiments and results of **EDITBIAS**, along with a more in-depth analysis and discussion about bias and debiasing effects in pre-trained language models.

4.1 Setups

Dataset We utilize StereoSet (Nadeem et al., 2021) to conduct all experiments. There are three reasons. Firstly, it is widely used (Liang et al., 2021; Meade et al., 2022; Smith et al., 2022; Joniak and Aizawa, 2022; Limisiewicz et al., 2023; Omrani et al., 2023; Ma et al., 2023b; Xie and Lukasiewicz, 2023; Yu et al., 2023; Yang et al., 2023) to evaluate different types of bias in pre-trained language models, including gender, race, and religion bias. Secondly, the meaningless attribute terms in StereoSet can be applied for modeling ability maintenance. Other datasets have no meaningless association data. Thirdly, the data size of StereoSet is large enough for training compared with other bias datasets. Since current bias datasets are created for measurement, their sizes are usually small. For example, Crows-Pairs (Nangia et al., 2020) only has 1508 samples without train/test splits. Comparatively, more than 8000 samples in StereoSet are suitable for our work. Gender, race, and religion bias data from StereoSet are considered in this work. We stochastically split all samples related to gender, race, and religion bias in the test set (6,392 samples) of the *intrasentence* StereoSet by 8:1 as $\mathcal{S}^{\text{train}}_{\text{edit}}$ and $\mathcal{S}^{\text{dev}}_{\text{edit}}$ respectively and use the development set (2,106 samples) as $\mathcal{S}^{\text{test}}_{\text{edit}}$.

Metrics We use the Stereotype Score and Language Modeling Score from StereoSet (Nadeem et al., 2021) to measure debiasing performance and language modeling performance respectively. The Stereotype Score (*SS*) is the percentage of samples in which a model prefers stereotypical contexts to anti-stereotypical contexts.

$$\text{SS}(\theta) = \mathbb{E}_{s \in \mathcal{S}^{test}_{\text{edit}}} \mathbb{1}\left[P_\theta(\cdot|x_{\text{stereo}}) > P_\theta(\cdot|x_{\text{anti}})\right]$$

The Language Modeling Score (*LMS*) is the percentage of examples in which a model ranks the meaningful associations over meaningless associations to measure a model's language modeling abilities for each attribute term.

$$\begin{aligned}\text{LMS}(\theta) = \frac{1}{2}\mathbb{E}_{s \in \mathcal{S}^{test}_{\text{edit}}} \mathbb{1}\left[P_\theta(\cdot|x_{\text{stereo}}) > P_\theta(\cdot|x_{\text{mless}})\right] \\ + \frac{1}{2}\mathbb{E}_{s \in \mathcal{S}^{test}_{\text{edit}}} \mathbb{1}\left[P_\theta(\cdot|x_{\text{anti}}) > P_\theta(\cdot|x_{\text{mless}})\right]\end{aligned}$$

An ideal unbiased model has a *SS* of 50% and an ideal debiasing will not change the *LMS* before and after debiasing.

Methods and Models Compared with EDITBIAS, four distinguishing baseline debiasing methods from Meade et al. (2022) are implemented[4]:

[4] https://github.com/McGill-NLP/bias-bench

Method	RoBERTa-base						RoBERTa-large					
	SS (%) $\rightarrow$ 50%			ΔLMS (%) $\rightarrow$ 0			SS (%) $\rightarrow$ 50%			ΔLMS (%) $\rightarrow$ 0		
	gender	race	religion	gender	race	religion	gender	race	religion	gender	race	religion
Pre-edit	65.78	62.34	59.54	89.53	89.85	86.46	69.35	62.80	50.76	90.14	90.71	87.98
CDA	62.81	62.14	57.55	-0.65	-1.07	+1.79	64.62	60.08	57.67	-1.31	-1.47	+1.39
SentenceDebias	64.17	60.00	55.85	**-0.59**	**-0.18**	-3.34	68.52	62.77	46.30	**+0.22**	**-0.06**	-1.68
Self-Debias	67.25	60.57	57.00	-0.84	-0.26	**-1.02**	66.03	59.95	51.69	-0.81	-0.21	**-0.96**
INLP	61.93	59.44	56.40	-1.49	+0.34	-1.90	68.66	60.60	53.25	-0.39	-1.30	-3.65
EDITBIAS	**49.67**	**48.48**	**51.04**	-34.74	-44.00	-52.69	**51.10**	**45.80**	**50.97**	-64.06	-57.52	-41.34

Method	GPT2-base						GPT2-medium					
	SS (%) $\rightarrow$ 50%			ΔLMS (%) $\rightarrow$ 0			SS (%) $\rightarrow$ 50%			ΔLMS (%) $\rightarrow$ 0		
	gender	race	religion	gender	race	religion	gender	race	religion	gender	race	religion
Pre-edit	62.67	60.57	58.02	93.28	89.76	88.46	65.58	61.63	62.57	93.39	92.30	90.46
CDA	60.33	58.70	59.97	**-0.81**	-1.94	-0.17	63.29	61.36	61.79	**-0.21**	-3.02	**0.00**
SentenceDebias	56.57	55.39	**50.65**	-10.55	+1.76	**+0.10**	67.99	58.97	56.64	+0.29	+1.52	+0.34
Self-Debias	62.32	58.95	57.00	-3.43	**+0.09**	-2.20	60.28	57.29	57.61	-3.47	-4.12	-1.35
INLP	59.87	55.51	55.73	-14.04	-1.34	-1.29	63.17	60.00	58.57	-5.15	**-1.49**	-2.48
EDITBIAS	**46.98**	**53.03**	53.53	-8.80	-15.53	-25.54	**48.20**	**53.29**	**55.84**	-8.97	-26.36	-44.81

Table 1: Performance of EDITBIAS compared with baselines. **Pre-edit** represents the exact *SS* and *LMS* of pre-trained language models before debiasing. **ΔLMS** (%) refers to the absolute change in *LMS* (%) during debiasing.

counterfactual data augmentation (CDA) (Zmigrod et al., 2019), SentenceDebias (Liang et al., 2020), Self-Debias (Schick et al., 2021), and iterative nullspace projection (INLP) (Ravfogel et al., 2020). Different from all baselines, our editor networks can be trained and validated with a mixture of all three types of bias, instead of dealing with only one particular bias at a time. As for testing, EDITBIAS is evaluated on gender, race, and religion bias samples from $\mathcal{S}_{\text{edit}}^{\text{test}}$ separately. The λ is determined by grid searching in each training ranging from {0.5, 1.0, 1.5, 2.0, 2.5, 3.0}. We implement parameter-efficient model editing utilizing low-rank gradient decomposition (Mitchell et al., 2022a). MLPs in different Transformer blocks in pre-trained language models are selected to be edited in this paper according to preliminary experiments described in Section 4.4. EDITBIAS is a **model-agnostic** debiasing method and can be applied to any open-source language model, such as LLaMA2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), QWen (Bai et al., 2023) and GLM (Zeng et al., 2023). Due to computational constraints, we conduct experiments on relatively small language models in this paper, including both masked language models, RoBERTa-base and RoBERTa-large (Liu et al., 2019), and causal language models, GPT2-base and GPT2-large (Radford et al., 2019) with HuggingFace (Wolf et al., 2019). We report the best debiasing performance among different edited positions in Table 1 (the last layer for RoBERTA-base, the penultimate layer for RoBERTa-large, and the first two layers for GPT2-base and GPT2-medium).

4.2 Main Results

EDITBIAS achieves the best debiasing performance on all types of bias compared to all debiasing baselines. According to the Stereotype Scores, EDITBIAS can reduce *SS* to less than 56% and more than 46% while most *SS* of debiased models with previous debiasing baselines are above 60%, which demonstrates EDITBIAS leads to significant improvement for debiasing performance. For instance, as for the *SS* of RoBERTa-base, EDITBIAS yields an improvement of ↑11.60, ↑7.92, and ↑4.81 on the absolute difference from 50% for gender, race, and religion bias respectively, compared with the best *SS* among all baselines. The main reason is that the parameters that may be associated with bias are explicitly edited, which is illustrated in Section 4.4 and Appendix A. Additionally, EDITBIAS obtains much better debiasing performance by training small editor networks in a few training steps (e.g., 14 steps for RoBERTa-base and 226 steps for GPT2-base) than fine-tuning an entire model in 2000 steps with CDA, which indicates the high efficiency of our EDITBIAS. Compared to prompting and representation projections baselines that can only calibrate models' output distributions instead of language models themselves, EDITBIAS produces off-the-shelf LMs that can be directly used for application and substantially outperforms them because modifying parameters effectively changes the internal representations and distributions of language models. Moreover, EDITBIAS presents excellent performance on every bias type though editor networks are trained to produce edits on a mixture of different types of bias at a time. It is illustrated that our method can generalize debiasing success to various bias, compared to debiasing baselines that can only deal with one particular bias at a time, such as creating a bias subspace of a certain bias in SentenceBias.

Editing debiasing parameters harms the original language modeling abilities. Unfortunately, EDITBIAS damages LMs' language modeling capabilities, though $\mathcal{L}_r$ is considered. *LMS* drops more than 10 (%), especially for editing top layers of RoBERTa. It is consistent with Gu et al. (2024); Gupta et al. (2024) that editing exhibits notable shortcomings in maintaining the inherent modeling capabilities of language models. Because rich semantic information and text patterns are captured by parameters of language models during pre-training (Geva et al., 2021), directly modifying some parameters will hurt the intrinsic encoding mechanisms. As a result, the whole language modeling abilities are destroyed, showing that the model's semantic recognition between meaningful and meaningless associations is ambiguous.

Debiasing larger models is more difficult. Comparing the results of models with different sizes, we observe that the difficulty of debiasing and the modeling effects from editing increase with the model size. Specifically, the sum of absolute difference *SS* from 50% for three types of bias is 1.89 of RoBERTa-base and 9.58 of GPT2-base while it is 6.27 of RoBERTA-large and 10.93 of GPT2-medium. And the *LMS* drops of RoBERTa-large and GPT2-medium during debiasing are larger than those of RoBERTa-base and GPT2-base respectively, indicating that larger models are more sensitive to bias (Vig et al., 2020b). According to the *SS* of pre-edit models, larger models are more biased likely because they capture more bias from the huger pre-training corpus. Meanwhile, with stronger language modeling abilities, it is harder for larger models to unlearn bias, and debiasing via model editing will definitely hurt the modeling capabilities to a large degree if we expect to implement successful debiasing. Although debiasing relatively large models is hard, empirical results demonstrate that EDITBIAS has great potential to debias large language models, with the advantage of efficiently modifying small portions of parameters compared to fine-tuning the whole model.

4.3 Ablation Study on Retaining Loss $\mathcal{L}_r$

Method	SS (%)			LMS (%)		
RoBERTa-base	gender	race	religion	gender	race	religion
w/o $\mathcal{L}_r$	47.37	46.06	51.92	-44.77	-52.47	-64.89
w $\mathcal{L}_r$	49.67	48.48	51.04	-34.74	-44.00	-52.69
GPT2-base	gender	race	religion	gender	race	religion
w/o $\mathcal{L}_r$	53.70	51.96	55.81	-43.27	-43.17	-53.33
w $\mathcal{L}_r$	46.98	53.03	53.53	-8.80	-15.53	-25.54

Table 2: Ablation study on the retaining loss $\mathcal{L}_r$.

We perform an ablation study to show the effectiveness of the retaining loss for maintaining language modeling abilities during debiasing. We disable the remaining loss and train editor networks with the same hyperparameters as the training process using the remaining loss. Results are shown in Table 2. There are large drops on *LMS* if the retaining loss is not deployed during editing. Specifically, the *LMS* drops of GPT2-base increase absolutely by 34.47, 27.64, and 27.79 for gender, race, and religion bias respectively during debiasing without $\mathcal{L}_r$, which illustrates that the remaining loss plays an important role in reducing harm to the language modeling abilities during editing.

4.4 Further Discussion on Editing Positions and Models for Debiasing

In EDITBIAS, MLPs in some Transformer blocks are selected to be edited for unlearning bias. To pursue optimal performance, it is necessary to carefully consider which hidden states to be edited. Before embarking on our main experimental investigation, therefore, preliminary experiments are conducted to explore bias effects in PLMs. Following causal tracing from Meng et al. (2022), we propose bias tracing to track bias effects in PLMs in Appendix A. It is observed that MLPs in several early and last Transformer blocks exert a substantial influence on bias captured in language models. Based on our findings and some existing works that demonstrate editing MLPs can modify knowledge associations in PLMs (Geva et al., 2021; Mitchell et al., 2022a; Meng et al., 2022, 2023; Gupta et al., 2023; Wu et al., 2023a), EDITBIAS edits MLPs in the first three and last three blocks for the debiasing task. To comprehensively explore the effects of the debiasing language models via model editing, we edit MLPs in different encoder & decoder

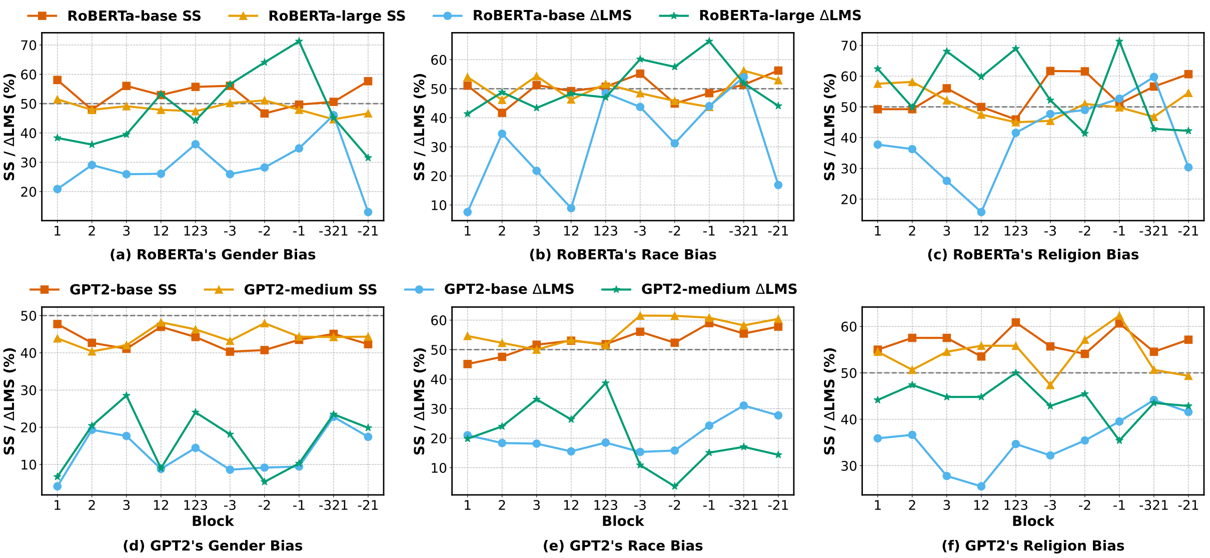

Figure 3: *SS* (%) and ΔLMS (%) drops of debiased language models after editing MLPs in different encoder & decoder blocks. 1/2/3: the first/second/third block. 12: the first 2 blocks. 123: the first 3 blocks. -1/-2/-3, the last/penultimate/antepenultimate block, -321: the last 3 blocks. -21: the last 2 blocks.

Model	Blocks	Gender	Race	Religion	SUM
RoBERTa-base	Early	24.84	**12.14**	**11.67**	48.65
	Last	**18.03**	19.40	41.53	**78.96**
RoBERTa-large	Early	**9.18**	**17.52**	25.27	**51.97**
	Last	12.08	21.16	**13.47**	46.71
GPT2-base	Early	**27.28**	**13.88**	34.45	75.61
	Last	38.07	30.63	32.13	**100.83**
GPT2-medium	Early	29.22	**11.74**	**21.42**	62.38
	Last	25.93	52.47	23.40	**101.80**

Table 3: The sum of the absolute differences between *SS* and 50%. Early (Last) blocks: 1, 2, 3, 12, and 123 (-3, -2, -1, -321, and -21) blocks.

blocks with EDITBIAS, and measure the resulting debiasing performance and modeling capabilities in this section. The *SS* and *LMS* drops of debiased language models are shown in Figure 3.

Debiasing causal language models is harder than mask language models. According to Figure 3, the Stereotype Scores of debiased RoBERTa are generally better and stabler than that of GPT2 and the *LMS* drops of RoBERTa are mostly larger and more unstable than that of GPT2, which indicates that it is more difficult to debias GPT2 than RoBERTa utilizing model editing. The reason is likely the different architectures of RoBERTa and GPT2. The bidirectional Transformer in RoBERTa might make the model more sensitive to changes in weights during debiasing than GPT2 with a unidirectional decoder-only structure because it integrates context from both directions when modeling bias. Based on the successful debiasing and relatively small *LMS* drops of GPT2, we can theoretically surmise that for most causal language models, debiasing them with editing methods is reliable and leads to a relatively little impact on modeling abilities, especially for current decoder-only large language models, like GPT-Neo (Black et al., 2021) and LLaMA2-70b (Touvron et al., 2023).

Editing MLPs in early blocks can achieve better debiasing performance than editing MLPs in upper blocks. According to Figure 3 and Table 3, in most cases, *SS* of debiased language models are closer to 50% after editing MLPs in bottom layers than in upper layers. Early layers capture basic linguistic features like syntax and common word associations while upper layers delve into deeper semantic relationships, contextual understanding, and high-level language features (Geva et al., 2021). Since biases often manifest in fundamental linguistic patterns, like the co-occurrence of bias attribute words and attribute terms, modifying early layers allows for correction at the source of these representations. Biases encoded in the early layers are propagated and potentially amplified through the network as information passes through subsequent layers. Since upper layers build on the representations formed by lower layers, biases present at the beginning can become deeply embedded and more complex to disentangle at later stages. By targeting debiasing efforts at the early stages, it's possible to prevent the propagation of biases, making the over-all debiasing process more effective. In contrast, the upper layers specialize in context-specific and complex language tasks. Editing biases in these layers might only address specific manifestations of bias and not the underlying bias itself.

The trade-off, mitigating biases in language models without significantly compromising the language modeling performance, is worth studying further. From Figure 3, we can see that achieving good debiasing performance comes at the cost of sacrificing language modeling capabilities. Editing for debiasing often involves altering the model's parameters to optimize the *SS*. However, these parameters were also optimized to perform well on language tasks, contributing to the *LMS*. When adjustments are made to reduce bias, they can interfere with the model's learned patterns, leading to a decrease in language modeling performance. Therefore, tackling biases arising from complex and deeply ingrained patterns within the training data without affecting the intricate structure of learned representations is challenging, which inspires us to seek methods to balance debiasing and modeling performance in the future.

4.5 Reversing Gender Attribute Words

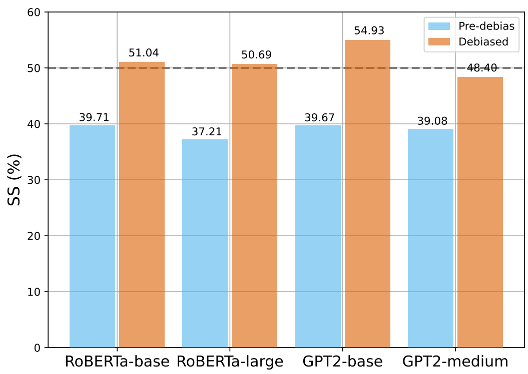

Figure 4: Gender Reverse Robustness. *Pre-debias* refers to *SS* of pre-trained language models on the gender reverse test set before debiasing. *Debiased* refers to *SS* of debiased models by EDITBIAS.

A robust gender debiasing method can calibrate a model's treatment to the two genders, male and female, equally. For instance, given the two sentences "Girls tend to be more ___ than boys." and "Boys tend to be more ___ than girls.", a debiased model will equivalently model the stereotypical term "soft" and the anti-stereotypical term "determined" in both two sentences though only the first sentence is used for training. To evaluate this robustness, a gender counterfactual test set $\mathcal{S}^{\text{test}}_{\text{gender*}}$ is created (Appendix C). We reverse all gender attribute words in the gender bias samples from $\mathcal{S}^{\text{test}}_{\text{edit}}$ to construct the set. For example, "boys", "father", and "Female" are changed into "girls", "mother", and "Male" respectively. Then the test set is used to examine the robustness of EDITBIAS, the implementation of which is the same as Table 1. The results in Figure 4 show that EDITBIAS is robust enough to unlearn gender counterfactual bias.

4.6 Semantic Generality

Model / SS (%)	Pre-debias			EditBias		
	gender	race	religion	gender	race	religion
RoBERTa-base	52.97	55.25	61.83	51.10	51.92	52.33
RoBERTa-large	50.39	54.20	60.50	51.37	48.53	47.53
GPT2-base	52.21	55.62	57.65	48.23	55.95	49.95
GPT2-medium	53.11	56.18	62.62	50.29	48.95	48.05

Table 4: *SS* (%) on the synonym-augmented test set.

Similar to the generality principle of knowledge editing, a robust debiasing method should ensure the debiased language model demonstrates unbiased behavior on a group of semantically similar attribute terms with attribute terms used in training, showcasing its adaptability to the nuanced and dynamic nature of language. To evaluate this robustness of EDITBIAS, we curate a synonym-augmented test set that substitutes attribute terms in $\mathcal{S}^{\text{test}}_{\text{edit}}$ with their synonyms generated by WordNet (Miller, 1995) using NLTK (Bird and Loper, 2004). Results in Table 4 show that our debiasing method can generally remove bias in the language models' neighboring semantic modeling space.

5 Conclusion

We propose **EDITBIAS**, an efficient model editing method to debias language models by modifying a small portion of PLMs' parameters with $\mathcal{L}_d$ and $\mathcal{L}_r$. Experiments illustrate that EDITBIAS presents much better debiasing performance than classical debiasing methods, and is robust in gender reverse and semantic generality though it hurts models' original language modeling abilities. Meanwhile, we comprehensively investigate debiasing and bias effects on language models, concluding that debiasing larger and causal language models is difficult, and it is important to consider the trade-off between debiasing and language modeling performance when designing debiasing methods. We hope our findings can give insights into future debiasing works and the NLP community.

Limitations and Future Works

More experiments to extend the debiasing method. In this work, we only study one benchmark dataset with its corresponding metrics. To extend the generality of our work, more bias datasets and metrics with various formats, from different domains and perspectives will be utilized in experiments, such as Stanceosaurus (Zheng et al., 2022) and HOLISTICBIAS (Smith et al., 2022). Due to the limited GPU resources, some larger language models have not been explored, such as LLaMA2 (Touvron et al., 2023), GLM (Zeng et al., 2023), and GPT-Neo (Black et al., 2021). We will conduct experiments with with more datasets and models in the future.

New bias editing methods with less modeling harm and without training costs. Though EDITBIAS obtains great performance on debiasing, alleviating its harm to the language modeling ability is significant and challenging. For instance, to reduce the modeling damage, we will try to edit neurons within a tiny disturbance, such as altering a small term in Taylor expansions of these activations. When compared to locate-then-edit approaches, like ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), as a meta-learning method, EDITBIAS necessitates additional training stages for hyper-networks, potentially leading to increased time and memory costs. In the future, we will try different editing methods without explicit training using large corpora.

References

H. Abdi and L. J. Williams. 2010. Principal component analysis. *WIREs Computational Statistics*, 2:433–459.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 298–306. ACM.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

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.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3727–3742. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in english language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1276–1295. Association for Computational Linguistics.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. 2023. Do localization methods actually localize memorized data in llms? *CoRR*, abs/2311.09060.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12041–12052. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1504–1532. Association for Computational Linguistics.

Siyuan Cheng, Ningyu Zhang, Bozhong Tian, Zelin Dai, Feiyu Xiong, Wei Guo, and Huajun Chen. 2023b. Editing language model-based knowledge graph embeddings. *CoRR*, abs/2301.10405.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2021. Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5034–5050. Association for Computational Linguistics.

Patricia G Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *CoRR*, abs/2309.00770.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring \textlessmask\textgreater: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2209–2225. Association for Computational Linguistics.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *CoRR*, abs/2401.04700.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1012–1023. Association for Computational Linguistics.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. *CoRR*, abs/2401.07453.

Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing common sense in transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8214–8232. Association for Computational Linguistics.

Matan Halevy, Camille Harris, Amy S. Bruckman, Diyi Yang, and Ayanna M. Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *EAAMO 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtual Event, USA, October 5 - 9, 2021*, pages 7:1–7:11. ACM.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023a. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *CoRR*, abs/2301.04213.

Peter Hase, Mona T. Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023b. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2706–2723. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Lang. Linguistics Compass*, 15(8).

Yoichi Ishibashi and Hidetoshi Shimodaira. 2023. Knowledge sanitization of large language models. *CoRR*, abs/2309.11852.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5961–5977. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Przemyslaw K. Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. *CoRR*, abs/2207.02463.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1720–1732. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5502–5515. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Tomasz Limisiewicz and David Marecek. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. *CoRR*, abs/2206.10744.

Tomasz Limisiewicz, David Marecek, and Tomás Musil. 2023. Debiasing algorithm through model adaptation. *CoRR*, abs/2310.18913.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *CoRR*, abs/2308.05374.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *Preprint*, abs/1907.11692.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security - Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. 2023a. Untying the reversal curse via bidirectional language model editing. *CoRR*, abs/2310.10322.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023b. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11328–11345. Association for Computational Linguistics.

Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *CoRR*, abs/2212.10678.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1878–1898. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of

Proceedings of Machine Learning Research, pages 15817–15831. PMLR.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. Association for Computational Linguistics.

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 2020, Online, November 16-20, 2020*, pages 1953–1967. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *Prepring arXiv*, abs/2305.14456.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2023. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. *CoRR*, abs/2311.08011.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4123–4139. Association for Computational Linguistics.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *CoRR*, abs/2309.17410.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Trans. Assoc. Comput. Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3239–3254. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry V. Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9180–9211. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1630–1640. Association for Computational Linguistics.

Chenmien Tan, Ge Zhang, and Jie Fu. 2023. Massive editing for large language models via meta learning. *CoRR*, abs/2311.04661.

Hugo Touvron, Louis Martin, Kevin Stone, and et.al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association* for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 307–318. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao K. Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 116–122. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020a. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020b. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3730–3748. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023a. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *CoRR*, abs/2308.09954.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023b. DEPN: detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2875–2886. Association for Computational Linguistics.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15730–15745. Association for Computational Linguistics.

Ke Yang, Charles Yu, Yi Ren Fung, Manling Li, and Heng Ji. 2023. ADEPT: A debiasing prompt framework. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10780–10788. AAAI Press.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10222–10240. Association for Computational Linguistics.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. 2023. History matters: Temporal knowledge editing in large language model. *CoRR*, abs/2312.05497.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6032–6048. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2896–2907. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *CoRR*, abs/2305.12740.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2132–2151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *CoRR*, abs/2012.00363.

Ran Zmigrod, S. J. Mielke, Hanna M. Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1651–1661. Association for Computational Linguistics.

A Bias Tracing

ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) utilize causal tracing (Vig et al., 2020a) to locate facts memorized in the parameters of a pre-trained autoregressive transformer. After they find the specific hidden state with the strongest effect on individual facts, they modify these localized parameters for changing facts. Inspired by causal tracing, we propose bias tracing to seek the exact hidden states that contribute most to bias exhibited in the language models including masked language models and causal language models, which will guide us to select positions to edit for debiasing.

A.1 Tracing Bias Associations

Following Meng et al. (2022), we analyze all internal activations of a language model $\mathcal{M}$ during three runs: a clean run eliciting the bias in language models, a corrupted run disrupting the bias context modeling, and a corrupted-with-restoration run measuring bias exhibited in a single state.

- As for the **clean** run, we obtain $P_\theta(\cdot|x_{\text{stereo}})$ and $P_\theta(\cdot|x_{\text{anti}})$ for each sample in the datasets, and collect all hidden activations $\{h_i^l|i \in [1, K], l \in [1, L]\}$ for each token i and each layer l, given the input text $x = [x_1, \ldots, x_K]$ and the $\mathcal{M}$ with L layers.

- In the **corrupted** run, noise is added to the embedding of all bias attribute words in the input. For the embedding h_i^0 in the token sequences of bias attributes words to be corrupted, we set $\hat{h}_i^0 := h_i^0 + \tau$, where $\tau \sim \mathcal{N}(0; \sigma)$.[5] Then, $\mathcal{M}$ runs based on the corrupted embeddings and we collected the following corrupted activations $\{\hat{h}_i^l|i \in [1, K], l \in [1, L]\}$. Since the existence of bias attribute words in a context is the reason why a context presents bias, corrupting the embedding of bias attribute words will remove the bias effects on the following language modeling process.

- With noisy embeddings, we restore specific hidden states of some token i (the bias attribute word, the attribute term, or the token before the attribute term) and layer l (the Transformer block, the attention layer, or the MLP layer) in the **corrupted-with-restoration** run, which lets $\mathcal{M}$ output the clean state h_i^l. The following forward-running executes without more intervention.

We calculate the absolute log probability difference between x_{stereo} and x_{anti}, $f_d(\theta, x_{\text{stereo}}, x_{\text{anti}}) = |\log P_\theta(\cdot|x_{\text{stereo}}) - \log P_\theta(\cdot|x_{\text{anti}})|$, to measure bias in a language model. The larger the difference is, the more biased $\mathcal{M}$ is. By running the network twice, bias tracing computes the bias effect of activations. The normal clean run occurs first to obtain all clean activations. Secondly, embeddings of bias attribute words are corrupted and the lowest difference is obtained. Then the corrupted activations $\hat{h}_i^l$ of a certain token i and layer l are restored to their original values h_i^l from the same token i and the same layer l. If an activation restoration of a token i^* and layer l^* causes a larger difference than a restoration from other tokens and layers, we can know that the activations of the token i^* and layer l^* give more impetus to bias.

A.2 Bias Tracing Results

We conduct gender bias tracing on the *intrasentence* part of StereoSet at every layer and every token. The average bias effects of 500 samples with GPT2-XL after a corrupted run and a corrupted-with-restoration run are shown in Figure 5 (a) and (b), respectively.

Bias best corresponds to the states of MLPs at lower layers. Figure 5 (a) illustrates that at layer 0-13, transformer block states and MLPs play a much more significant role in bias than attention layers, with peaking at layer 8. This reveals that language models intensively present bias in the foundational representations learned by lower layers, and these early presentations can influence the

[5] σ is three times the standard deviation of 1000 subject embeddings from `https://rome.baulab.info/data/dsets/known_1000.json`

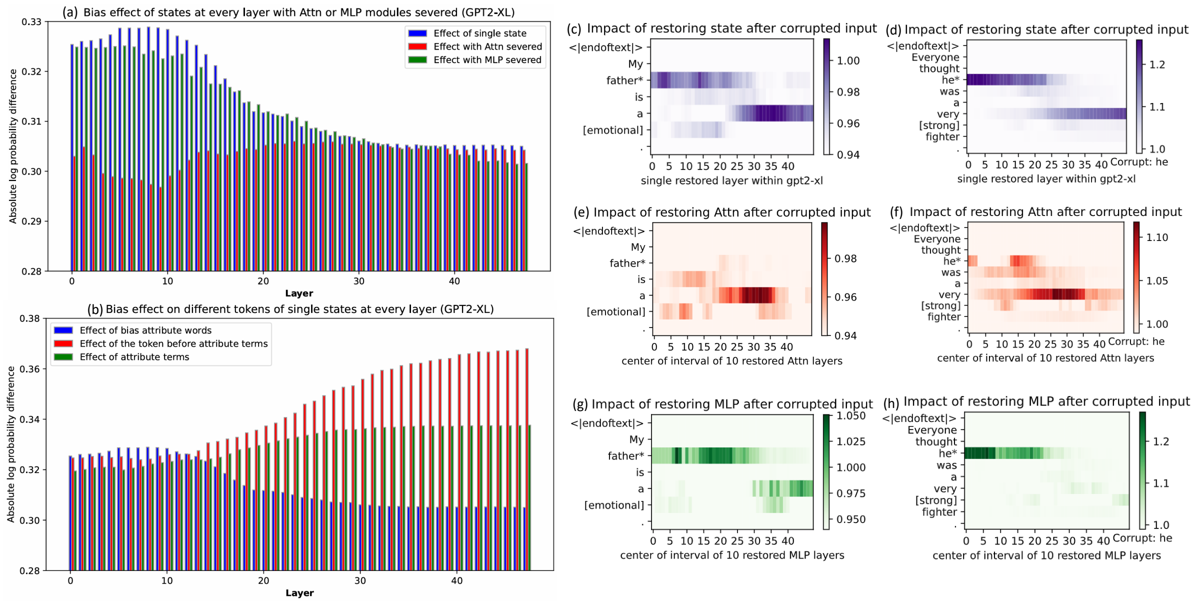

Figure 5: Gender bias tracing on GPT2-XL. (a) Comparing bias effect with and without severing Attn or MLP. (b) Comparing bias effect on different token positions. The bias impact on output probability is mapped for the effect of (c-d) each hidden state on the context, (e-f) only MLP activations, and (g-h) only attention activations. * marks the corrupted bias attribute words and [] refers to the attribute terms in (c-h).

subsequent layers. The reason is that since the lower layers capture the text patterns (Geva et al., 2021), bias patterns in the pre-trained corpus, such as cooccurrence with stereotyped terms, are memorized in the early layers. Figure 5 (b) also shows that bias attribute words have the most effects at the early layers. Meanwhile, it indicates that attribute terms and the token before it associated with bias at the upper layers, especially for the token before attribute terms because semantic information is usually modeled in the top layers, and the token probability is most influenced by the previous one in a causal language model. Two cases in Figure 5 (c-h) illustrate the aforementioned observations well. Besides, Figure 5 (e-f) manifests that attention from the bias token to attribute tokens shows a strong relation with bias, which results from the causal effect of the bias token.

A.3 Tracing Data Construction

We begin with utilizing SPARQL to query the instance of gender, race, and religion, obtaining a variety of words targeted to specific bias. These words are the source collection of bias attribute words. Based on the collection, we then adopt simple string matching to extract bias attribute words from the context sentence x of each sample s in the dataset. As a result, we can trace the activations of these bias attribute words in language models.

A.4 Bias Tracing with RoBERTa-large

Figure 6 shows the bias effects of RoBERTa-large. Different from GPT2-XL, Transformer blocks, attention layers, and MLPs follow the same trend in bias effects without causal effects. According to Figure 6 (a), the strong association is located in the early layers, and the impacts become less and less from the bottom layer to the top layer because bias patterns are captured in these beginning layers, the same as GPT2-XL. Figure 6 (b) also illustrates that bias words have the most bias effects in the bottom layers and the attribute terms containing the semantic information of bias influence the modeling at the upper layers.

B Baselines

CDA (Counterfactual Data Augmentation) retrain a pre-trained language model. It generates and incorporates data that represents what could have happened under different conditions. By altering aspects of data related to biased attributes, such as changing gender or race in a dataset, a counterfactual data set is created to create a more balanced training environment for models.

SentenceDebias (Liang et al., 2020) first estimates the demographic bias subspace by encoding sentences containing bias attribute words or their counterfactuals into sentence representations and using principle component analysis (Abdi and Williams, 2010) to define the bias subspace as the first K principle components. and then debias sentence representations by subtracting their projection onto the bias subspace.

Self-Debias (Schick et al., 2021) first prompts a model to generate toxic text, such as encouraging a model to discriminate based on gender. Then, the model can generate a non-discriminative continuation, during which the probabilities of tokens that were prominent in the toxic generation are deliberately scaled down.

INLP (Ravfogel et al., 2020) introduces Iterative Null-space Projection (INLP), a method that reduces bias in word embeddings by iteratively projecting them onto the null space of bias terms using a linear classifier. This method constructs a projection matrix to project input onto the null space of the linear classifier, continuously updating both the classifier and the projection matrix.

C Gender Counterfactual Test Set

We utilize the method mentioned in Appendix A.3 to extract gender attribute words in gender bias samples. Then these gender attribute words are reversed into their counter facts manually. The labels "stereotype" and "anti-stereotype" are exchanged for each sentence. For instance, after reverse, the stereotyped context in Figure 1 is "Boys tend to be more determined than girls." and the anti-stereotyped context is "Boys tend to be more soft than girls.".

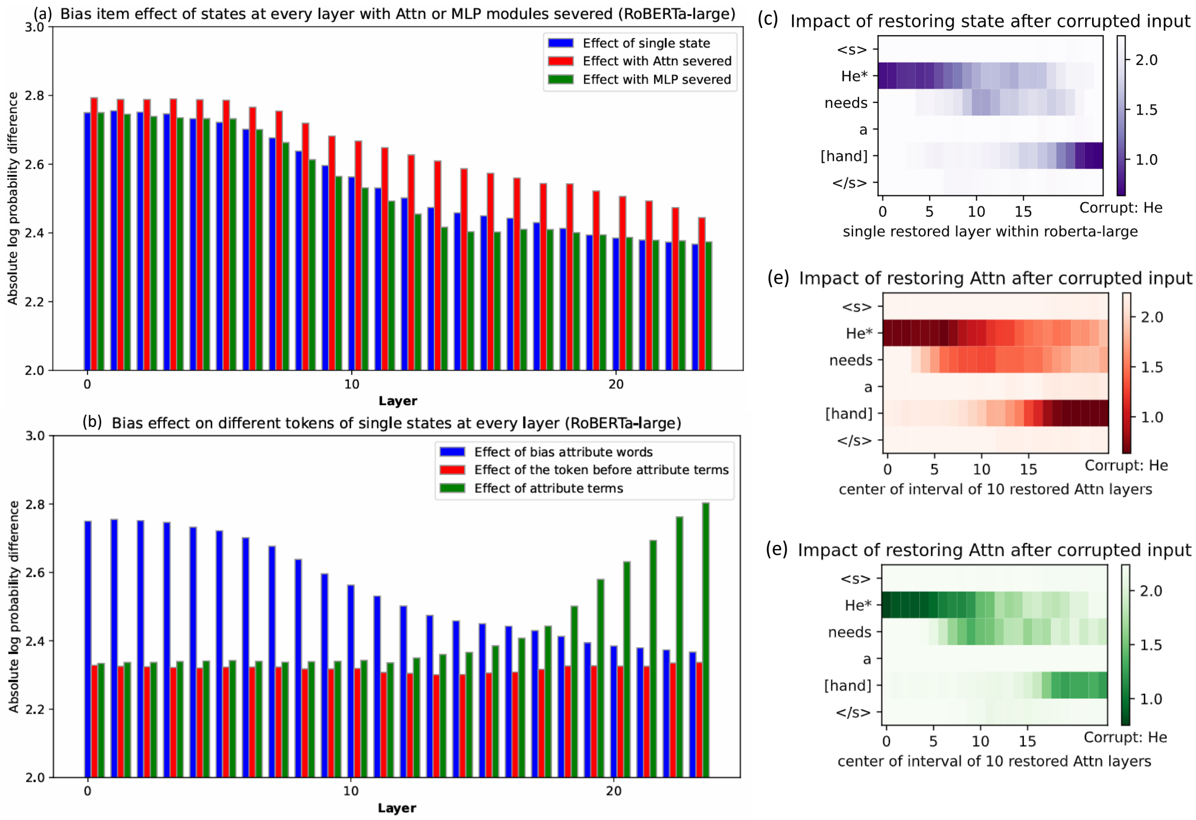

Figure 6: Gender bias tracing with RoBERTa-large.