Towards Understanding Task-agnostic Debiasing Through the Lenses of Intrinsic Bias and Forgetfulness

Anonymous ACL submission

Abstract

While task-agnostic debiasing provides notable generalizability and reduced reliance on downstream data, its impact on language modeling ability and the risk of relearning social biases from downstream task-specific data remain as the two most significant challenges when debiasing Pretrained Language Models (PLMs). The impact on language modeling ability can be alleviated given a high-quality and long-contextualized debiasing corpus, but there remains a deficiency in understanding the specifics of relearning biases. We empirically ascertain that the effectiveness of task-agnostic debiasing hinges on the quantitative bias level of both the task-specific data used for downstream applications and the debiased model. We empirically show that the lower bound of the bias level of the downstream fine-tuned model *is* the bias level of the debiased model, in most practical cases. To gain more in-depth understanding about how the parameters of PLMs change during fine-tuning due to the *forgetting* issue of PLMs, we propose a novel framework which can **Pro**pagate **Social**ly-fair Debiasing to Downstream Fine-**tuning**, ProSocialTuning. Our proposed framework can push the fine-tuned model to approach the bias lower bound during downstream fine-tuning, indicating that the ineffectiveness of debiasing can be alleviated by overcoming the forgetting issue through regularizing successfully debiased attention heads based on the PLMs' bias levels from stages of pretraining and debiasing[1].

1 Introduction

Social fairness of PLMs has recently drawn intense critical attention, particularly due to the widespread deployment of PLM-based systems (Bender et al., 2021; Zhuo et al., 2023; Ouyang et al., 2022). Social biases embedded in PLMs can drive PLM-based systems to generate stereotypical content with respect to underrepresented demographic groups, raising serious issues of social fairness (Elsafoury and Abercrombie, 2023). Therefore the process of debiasing PLMs to better align them with social values of fairness is a key procedure before deploying PLMs for public access (Sun et al., 2019).

To illustrate the unintended behavior of social bias, a popular example is: *The **surgeon** asked the nurse a question, **he** ...*; *The **nurse** asked the surgeon a question, **she** ...*. Given the occupation token, *surgeon*, in the context of "*The surgeon asked the nurse a question*", PLMs are more likely to make a generation decision to assign the binary gender token he, instead of she, by referring to the occupational token. This indicates that PLMs predict surgeons as male with a higher probability than surgeons as female, presenting an example of gender bias (Bordia and Bowman, 2019; Lu et al., 2020). Intrinsically, PLMs amplify the statistical bias in the pretraining corpus where the concurrence between *surgeon* and he is much larger than that between *surgeon* and she (Liang et al., 2021). Despite various studies highlighting social bias issues (Bordia and Bowman, 2019; Nozza et al., 2022; Smith et al., 2022), the effectiveness of debiasing for downstream applications has been a debate (Kaneko et al., 2022; Jeoung and Diesner, 2022; Jin et al., 2021).

When it comes to debiasing, the language modeling abilities (Meade et al., 2022) and relearning social biases (Kaneko et al., 2022) are the two main concerns limiting the effectiveness of debiasing. Considering counterfactual data augmentation (CDA) (Webster et al., 2020) as an instance of debiasing, the lower quality of the debiasing corpora compared to the pretraining corpora negatively impacts the language modeling ability, therefore degrading downstream performance. Earlier studies have arrived at varying conclusions regarding the

[1] Unless explicitly stated otherwise, debiasing in this paper refers to task-agnostic debiasing.

effectiveness of debiasing in reducing social bias in fine-tuned tasks. Webster et al. (2020) and Jeoung and Diesner (2022) claim that a debiased model can help with downstream tasks, but Kaneko et al. (2022) empirically demonstrates that fine-tuning a debiased model for downstream tasks can lead to significantly biased models (He et al., 2022; Zhou et al., 2023a). However, an in-depth understanding of this ineffectiveness is still under-studied.

This paper focuses on the relearning of social bias challenge and proposes a framework to alleviate this problem via an in-depth understanding of how PLMs' parameters change during debiasing and fine-tuning. We empirically indicate that debiased PLMs are sensitive to bias in downstream data through a comprehensive analysis of the bias score of the fine-tuned model given various bias levels [2] in downstream data. Our observations indicate that: (1) the bias level of the debiased PLMs is the lower bound for any fine-tuned PLMs for practical cases, and (2) relearning social biases derives from the forgetting issue of PLMs (Kirkpatrick et al., 2017; Zhao et al., 2023).When fine-tuning occurs in downstream tasks exhibiting higher bias levels, the resultant model tends to display greater bias compared to the initial debiased model. Through meticulous control of bias levels within downstream tasks, we can conclude that the effectiveness of task-agnostic debiasing is dependent on the bias level of both the debiased PLMs and the downstream data.

To thoroughly understand how the attention heads of a PLM change, and how those changes are associated with social biases and downstream generalization, we propose ProSocialTuning. Specifically, we implement a generalization importance estimation method based on PAC-Bayes training, which indicates parameters' importance by learning parameter-wise noise variance through minimizing a variant of a PAC-Bayes bound in a post-training manner (Liu et al., 2023a; Louizos et al., 2018). A higher noise variance indicates less importance to generalization. In the downstream fine-tuning stage, we apply regularization to successfully debiased attention heads, guided by their importance to downstream generalization.

In Section 2 we introduce relevant works. Section 3 introduces our first main contribution: use of the bias level as a lower bound. Section 4 presents the necessary mathematical and algorithmic background context for our second main contribution: our novel framework, ProSocialTuning. The remaining sections detail ProSocialTuning and its experimental evaluation.

[2]We define bias level as the intrinsic/extrinsic bias score of the target PLM before/after fine-tuned with downstream data.

2 Related Works

The **effectiveness** of a separate step of debiasing before downstream fine-tuning has been explored in recent studies. Kaneko et al. (2022) implement comprehensive studies on the intrinsic bias of PLMs and extrinsic bias of fine-tuned PLMs in downstream applications, in terms of gender bias. The experimental results show a debiasing step is less effective for downstream tasks, against the conclusion of debiasing transferability in Jin et al. (2021). Goldfarb-Tarrant et al. (2021) indicates the intrinsic bias evaluation metric is not correlated to application bias. A similar conclusion is presented in Steed et al. (2022), in which the authors investigate the bias transfer hypothesis and prove that debiasing cannot help mitigate bias in fine-tuned tasks. Zhou et al. (2023b) proposed causal-Debias to solve the ineffectiveness of debiasing but their assumption about causal factors is too strong and cannot generalize to other datasets well.

PAC-Bayes Training is a training algorithm that is different from the conventional empirical risk minimization, optimizing a machine learning model by minimizing a generalization error bound (PAC-Bayes bound). McAllester (1998) trained a shallow network through minimizing a non-vacuous PAC-Bayes bound and achieved good performance. The PAC-Bayes with BackProp proposed by Rivasplata et al. (2019) trains shallow probabilistic networks and certifies their risk by PAC-training on the MNIST dataset. Liu et al. (2023a) proposes PAC-tuning to leverage PAC-Bayes training for fine-tuning PLMs in the significantly challenging context of high dimensional parameters and small size of the training dataset. The PAC-tuning is an extension to Zhang et al. (2023) which introduced an Auto-tune method based on PAC-Bayes training, by optimizing both prior and posterior variance of model's parameters, and proposing a new PAC-Bayes bound for unbounded classification loss.

3 Bias Lower Bound

In this section, we present the first major contribution of this work: that the bias level, i.e., the level

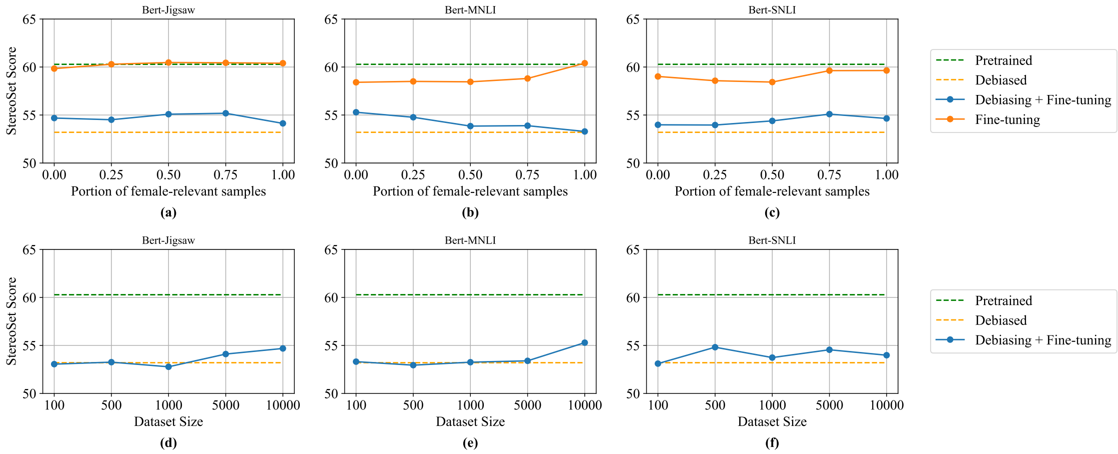

Figure 1: StereoSet Scores of BERT models when bias level and training dataset size Vary. The (**StereoSet**) intrinsic bias scores of the pretrained, debiased, and fine-tuned models are assessed concerning different bias levels and training dataset sizes present in specific datasets for downstream tasks. The fine-tuned model is based on the debiased one and fine-tuning indicates fine-tuning the pretrained model with task-specific data. Models are considered to be less biased when closer to 50.

of a specific type of bias (e.g., gender bias) of a debiased model can be leveraged as a lower bound for optimizing the fine-tuning of PLMs, given a biased fine-tuning dataset. With this, we aim to close the debate about the ineffectiveness of debiasing via experiments highlighting extreme cases.

We began by investigating the correlation between the effectiveness of debiasing and the bias levels in the debiased model and downstream tasks, in the context of the gender bias task. To do so, for different datasets, we compare the bias score of fine-tuned models, as measured by the StereoSet Score [3], with respect to: (1) proportions of female gender-relevant samples, as defined by the gender word list in Zhao et al. (2018), and (2) dataset sizes, as shown in Figure 1. Given a debiased model, we manipulate the bias levels in the training set and report the bias score of the fine-tuned model with respect to various bias levels. We use three datasets for analysis: MultiNLI (Williams et al., 2018) from the GLUE benchmark, the Jigsaw Unintended Bias in Toxicity Classification [4], and the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015). To experiment with dataset sizes, we randomly sample data from the training dataset wherein no sentences contain female-relevant words. We consider varying dataset sizes of 100, 500, 1000, 5000, and 10000 instances to analyze the impact of different training dataset sizes.

To vary the bias levels with respect to gender-relevant samples across PLMs, we rebalance samples containing words relevant to the female gender in our training dataset. Then we construct a training dataset with 10,000 samples and change the amount of samples with the pre-defined female-relevant words. In our experiments, we systematically varied the proportion of sentences containing female gender words, setting it at 0.0, 0.25, 0.5, 0.75, and 1.0. Subsequently, we calculated the average bias score across three different seeds for each of these proportion settings. To validate the effects of debiasing on the language modeling ability, we conducted experiments to gauge the language modeling score[5]. As shown in Appendix Figure 3, the Pearson product-moment correlation coefficients between bias score and language modeling score is less than 1. Thus, we can focus on the effects of the bias levels of the data and models, as those are the most straightforward factors in practical scenarios.

According to Figure 1, the fine-tuned model indicates more bias than the debiased one in most cases, implying the ineffectiveness of debiasing. This is further verified by the lower bias score of the fine-tuned model based on the pretrained model versus the pretrained model (Figure 1(b)-(c)). These findings indicate that the bias level in the downstream task is *less than* that of the pretrained model. Changing the bias levels in training data results in varying fluctuations of bias scores among fine-tuned models across the three evaluated benchmark tasks. The bias score gap between

[3] In this work, the intrinsic *bias score* is the StereoSet Score (Nadeem et al., 2021a).

[4] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[5] The language modeling score evaluates the baseline performance of PLMs in language modeling tasks. An ideal model would have a score of 100.

the fine-tuned model based on the pretrained model versus the debiased model is attributed to the disparity of their language modeling abilities. Given the experimental results regarding varying dataset sizes (Figure 1(d)-(f)), it is obvious that fewer training samples result in lower bias scores. Therefore we can conclude that the bias levels of the downstream tasks are highly relevant to the debiasing effectiveness.

Remarkably, debiasing + fine-tuning displays the highest bias scores of around 55 across various bias levels and tasks. Conversely, fine-tuning has a peak bias score closely aligned with the bias score of the pretrained model. Moreover, the lowest bias scores exhibited by debiasing + fine-tuning with differing dataset sizes are strikingly akin to the bias score of the debiased model. However the bias score of debiasing + fine-tuning should be higher than the debiased model, considering downstream tasks are generally rather biased in practical scenarios. Consequently, the efficacy of task-agnostic debiasing hinges upon both the bias level present in the downstream task data and the debiased model. The debiased model sets a definitive lower bound for the bias levels of the fine-tuned model after debiasing, as long as social bias exists within the downstream task data (Gaci et al., 2022b). Inspired by this conclusion, in Section 5, we prove that we can approach the lower bound of the bias level by regularization over the debiased model itself, without any additional debiasing methods or annotated datasets, given highly biased downstream tasks.

4 Background

In this section, we present the mathematical and algorithmic context necessary for understanding our ProSocialTuning framework. Assume a PLM f, consisting of L layers and K attention heads per layer, is parameterized by θ with attention weights as θ^A. The k^{th} attention head in the l-th layer $a_{l,k}$ is parameterized by $\theta^A_{l,k}$. We denote CMA($f, \mathcal{D}_{\text{cma}}$) as the Causal Mediation Analysis to the attention heads of f with dataset $\mathcal{D}_{\text{cma}}$, and denote CDA($f, \mathcal{D}_{\text{cda}}$) as debiasing of PLM f with the counterfactual data augmentation dataset $\mathcal{D}_{\text{cda}}$. For each training sample x_i and its label y_i, we denote the cross-entropy loss as $l(x_i, y_i; \theta)$.

4.1 Bias-inducing Attention Shift

Based on the conclusion of Section 3 that the bias level of the debiased PLMs performs the lower bound for downstream fine-tuning as long as there exists bias in the downstream task, we investigated how the bias-inducing effects of PLMs change throughout the pipeline of pretraining, debiasing, and fine-tuning, given the well-known forgetting issue of PLMs (Kirkpatrick et al., 2017). Our emphasis on the attention heads of PLMs stems from their deterministic nature in associating tokens during the inference process, as well as their utilization in previous debiasing works (Attanasio et al., 2022; Zayed et al., 2023; Gaci et al., 2022a).

Causal Mediation Analysis (CMA) is widely used in the social sciences fields. Imai et al. (2010) and Vig et al. (2020) first proposed localizing social bias-inducing network components using CMA. The rationale behind CMA is to measure the effect of a target network component concerning the anti-stereotypical and stereotypical outputs of PLMs, according to the interventions over the input prompt u. For analyzing gender bias, an example intervention is modifying the gender-relevant word.

Specifically, given the prompt u_{nurse} = "*The **nurse** is great,* __", the anti-stereotypical candidate word is $[he]$ and the stereotypical word is $[she]$. The prediction probability of $[he]$ given the prompt u_{nurse} is $p_\theta([he]|u_{\text{nurse}})$; by swapping the word ***nurse*** into ***man***, then the probability of *he* is $p_\theta([he]|u_{\text{man}})$. The effects of intervention in u to the output via $a_{l,k}$ is defined as:

$$e_{a_{l,k}} = \frac{p_\theta([he]|u_{\text{man}})}{p_\theta([she]|u_{\text{man}})} / \frac{p_\theta([he]|u_{\text{nurse}})}{p_\theta([she]|u_{\text{nurse}})} - 1$$

CMA measures how the prediction probability gap between anti-stereotypical predictions and stereotypical predictions is different from the ground-truth probability gap, considering the effect of $a_{l,k}$. By applying CMA, the distributions of bias-inducing effects of attention heads are shown in Figure 2. The effect distributions of attention heads within the pretrained model, debiased model, and fine-tuned models are rather different even though those fine-tuned models are all based on the same debiased model. For example, an attention head $a_{4,9}$ has higher bias-inducing effects in the pretrained model becomes less effective in all fine-tuned models, and not all attention heads are debiased, to some extent, in the debiased model. This strong inconsistency, termed as **bias-inducing attention shift**, is attributed to the forgetting issue of PLMs. The conclusion, from Section 3, that the effectiveness of debiasing is partially dependent on the bias level of the debiased model, motivates us

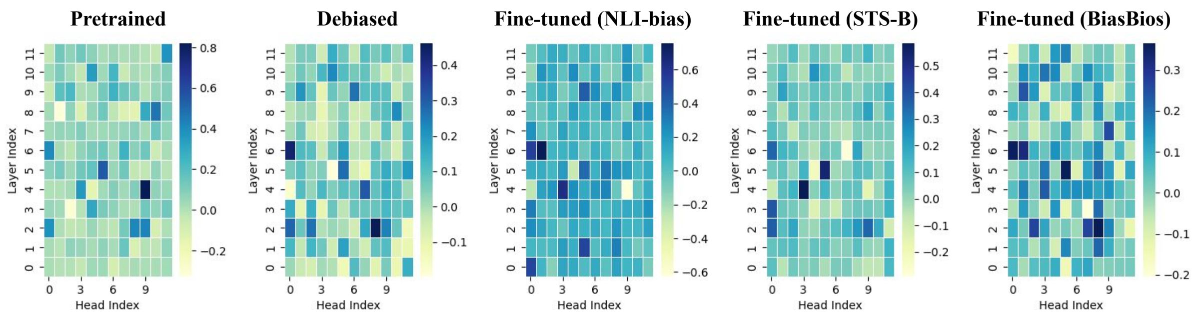

Figure 2: Visualization of CMA Effects of Attention Heads. From left to right, these figures show the effect of CMA on attention heads in the pretrained BERT-base model, debiased BERT-base model, and fine-tuned BERT-base model on benchmarks of NLI-bias, STS-B, and BiasBios respectively. The default random seed is 1. The fine-tuned model is based on the debiased model.

to regularize successfully debiased attention heads to enhance the effectiveness of debiasing.

4.2 PAC-Bayes Training

The idea of PAC-Bayes training arises from minimizing the PAC-Bayes upper bound over the generalization (test) error:

$$\overbrace{\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}\ell(x,y;\theta)}^{\text{Generalization Error}}$$

$$\leq \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\theta\sim\mathcal{Q}}\ell(x_i,y_i;\theta)}_{L_{\text{train}}} + \underbrace{\sqrt{\frac{\log\frac{1}{\delta}+\mathrm{KL}(\mathcal{Q}||\mathcal{P})}{2m}}}_{L_{\text{PAC}}}$$

PAC-Bayes bounds are probabilistic bounds that hold with high probabilities, i.e., $1-\delta(\delta>0)$, and for any neural network type. They characterize the generalization error of a trained model f_θ. Here, m is the number of training samples, $\mathcal{Q}$ and $\mathcal{P}$ are arbitrary pairs of posterior and prior distributions of θ, KL is the Kullback–Leibler divergence measuring the distance between two distributions, and $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ is the training data distribution and test data distribution, respectively.

PAC-Bayes training is a framework for understanding and improving generalization by directly minimizing a generalization upper bound. One difficulty in leveraging PAC-Bayes training for PLMs and any other deterministic models is to estimate $\mathcal{Q}$ and $\mathcal{P}$. A popular solution is to fix $\mathcal{P}$ and inject Gaussian noise to the trained parameters θ in the course of training, and estimate the Gaussian noise variance (Zhang et al., 2023; Liu et al., 2023a). Therefore the L_{train} term can be rewritten as $L_{\text{train}} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\text{diag}(q))}\ell(x_i,y_i,\theta+\epsilon)$ where $q\in\mathbb{R}^{|\theta|}$. L_{train} becomes increasingly larger as the injected noise variance q rises, indicating L_{train} is an increasing function with respect to q. Once convergence has been achieved by minimizing $L_{\text{train}}+L_{\text{PAC}}$, the learned noise ϵ can be utilized to reflect how important each parameter is to the final performance. Parameters associated with larger noise variance are less important than those with a smaller noise variance. This is because injecting larger noise into those parameters does not influence training error (L_{train}). A similar idea of Gaussian noise injection has been used in sparse Bayesian learning (Tipping, 2001). Sønderby et al. (2016) implements dropout through multiplying the outputs of neurons by Gaussian random noise. Molchanov et al. (2017) proposes a sparse variational dropout method to learn a costumed dropout rate per parameter via variational inference, and approximate the KL-divergence term by having a Gaussian posterior and a log-uniform prior over model weights.

5 ProSocialTuning

Using the analysis of Section 3 and bias-inducing attention shift (Section 4.1), ProSocialTuning shows that we can propagate debiasing efforts to downstream fine-tuning by only remembering the successfully debiased attention heads. This framework offers insight into understanding the resurgence of social bias in downstream applications.

5.1 Algorithm of ProSocialTuning

Algorithm 1 describes the pipeline of ProSocialTuning. Given a pretrained language model f_0, CMA is employed to get the bias-inducing effects of all attention heads ($\mathcal{B}^0$). We denote $\mathcal{B}^0_{l,k}$ as the bias-inducing effect of the k^{th} attention head in the l^{th} layer. After that f_0 is aligned with human values of social fairness through counterfactual data augmentation (Webster et al., 2020). The aligned model f_A

Algorithm 1: ProSocialTuning

Input: Pretrained Language Model f_0, Causal Mediation Analysis dataset $\mathcal{D}_{\text{cma}}$, counterfactual data augmentation dataset $\mathcal{D}_{\text{cda}}$, downstream dataset $\mathcal{D}_{\text{task}}$, regularization coefficient γ
Output: A fine-tuned model f_T
$\mathcal{B}^0$ = CMA(f_0, $\mathcal{D}_{\text{cma}}$) ▷ *causal mediation analysis*
f_A = CDA(f_0,$\mathcal{D}_{\text{cda}}$) ▷ *counterfactual data augmentation*
$\mathcal{B}^a$ = CMA(f_A, $\mathcal{D}_{\text{cma}}$) ▷ *causal mediation analysis*
Fine-tune f_A to convergence and produce $f_A^{'}$
Estimate generalization importance a^G by minimizing the objective of $\mathcal{E}_{\text{gen}}$ ▷ Section 5.2
Fine-tune f_A with the objective of $\mathcal{E}_{\text{tuning}}$ and produce f_T ▷ Section 5.3

is passed into CMA to get the bias-inducing effects of attention heads as $\mathcal{B}^a$. By comparing $\mathcal{B}^0$ and $\mathcal{B}^a$, we can determine which attention heads are debiased. ProSocialTuning propagates the learned fairness to downstream fine-tuning tasks by regularization over those successfully aligned attention heads, as further described below.

5.2 Generalization Importance Estimation

Specifically, to estimate the parameter-wise generalization importance, we propose a post-training method that first fine-tunes f_A to convergence, then estimates the injected noise variance associated with each parameter by minimizing $\mathcal{E}_{\text{gen}}$ (defined below). With the learned noise variance, we can calculate the parameter-wise generalization importance of a^G. Finally, the aligned model f_A is fine-tuned with the new objective function $\mathcal{E}_{\text{tuning}}$ (Section 5.3) over the downstream task dataset $\mathcal{D}_{\text{task}}$. Our proposed generalization importance estimation method is task-agnostic and less sensitive to hyper-parameters, enabling ubiquitous application of our proposed framework for downstream applications.

The L_{PAC} term in Section 4.2 can be simplified as $L_{\text{PAC}} = \text{KL}(\mathcal{Q}_q||\mathcal{P})$ if the prior distribution $\mathcal{P}$ is fixed and δ is omitted. The only learnable parameter is q, further reducing the computational complexity. The objective function for estimating generalization importance is: $\mathcal{E}_{\text{gen}} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\text{diag}(q))}\ell(x_i, y_i, \theta + \epsilon) + \lambda\text{KL}(\mathcal{Q}_q||\mathcal{P})$ where λ is the coefficient for the KL term. More details about our generalization estimation method are available in Appendix A.1.

Our method estimates generalization importance in a post-training manner, ensuring the estimation accuracy by referring to the performance of the converged model. ProSocialTuning enjoys computational benefits in contrast to other in-training approaches (Kwon et al., 2022). For the i^{th} parameter in θ, its generalization importance is calculated as $1/\exp(q_i)$. For the importance of each attention head, we summarize the importance associated with all parameters of the same attention head and take the summarized importance as the generalization importance measurement of that attention head. Appendix A.2 details our implementation of the generalization importance estimation.

5.3 Generalization-guided Regularization

Given the aligned model f_A debiased with counterfactual data augmentation, the attention heads' parameters of $\theta^{\text{cda}} \in \mathbb{R}^{|\theta^A|}$, detected bias-inducing effects of attention heads $\mathcal{B}^0 \in \mathbb{R}^{L\cdot K}$ and $\mathcal{B}^a \in \mathbb{R}^{L\cdot K}$, for f_0 and f_A respectively, as well as the generalization importance measurement $a^G \in \mathbb{R}^{L\cdot K}$, the objective function in downstream fine-tuning is: $\mathcal{E}_{\text{tuning}} = \frac{1}{m}\sum_{i=1}^{m}\ell(x_i, y_i; \theta) + \gamma\frac{1}{LK}\sum_{l,k}\frac{a_{lk}^G\cdot\mathbb{I}(\mathcal{B}_{lk}^a<\mathcal{B}_{lk}^0)}{\sum_{i,j}a_{ij}^G\cdot\mathbb{I}(\mathcal{B}_{ij}^a<\mathcal{B}_{ij}^0)}||\theta_{lk}^A - \theta_{lk}^{\text{cda}}||_2^2$ where γ is the regularization coefficient, and θ^{cda} is fixed. With the indicator function $\mathbb{I}(\mathcal{B}_{ij}^a < \mathcal{B}_{ij}^0)$ we only consider attention heads that have weaker effects for bias-induction in f_0 than their effects within f_A. The regularization coefficient γ is re-weighted according to the generalization importance of those attention heads. The generalization-guided regularization reflects the attention heads' sensitivity to downstream performance and helps balance the fairness-accuracy trade-off in downstream fine-tuning tasks.

6 Experiments

In this section, we introduce the experimental settings and results of ProSocialTuning, which indicate that an inability to address the forgetting issue in PLMs limits the effectiveness of debiasing.

6.1 Experimental Settings

In this paper, we take two masked language models BERT-base-uncased (Kenton and Toutanova, 2019) and RoBERTa-base (Liu et al., 2019) as our backbone models, and use the language modeling head

BERT-base	Accuracy (NLI-bias)	**Bias (NLI-bias)**	Accuracy (STS-B)	**Bias (STS-B)**	Accuracy (Biasbios)	**Bias (Biasbios)**
Vanilla-tuning	.795	**.021**	.507	**.197**	.722	**.018**
Debiased-tuning	.751	**.020**	.473	**.184**	.668	**.013**
EAR (Attanasio et al., 2022)	.796	**.013**	.509	**.233**	<u>.727</u>	**.017**
MABEL (He et al., 2022)	<u>.813</u>	**.030**	<u>.570</u>	**.181**	.694	**.028**
INLP (Ravfogel et al., 2020)	N/A	**N/A**	N/A	**N/A**	.714	**.038**
ProSocialTuning	.747	**<u>.012</u>**	.460	**<u>.169</u>**	.661	**<u>.003</u>**
<u>RoBERTa-base</u>	Accuracy (NLI-bias)	**Bias (NLI-bias)**	Accuracy (STS-B)	**Bias (STS-B)**	Accuracy (BiasBios)	**Bias (BiasBios)**
Vanilla-tuning	.859	**.021**	.578	**.330**	.691	**.030**
Debiased-tuning	.774	**.015**	.518	**.314**	.647	**.018**
EAR (Attanasio et al., 2022)	.859	**.040**	<u>.595</u>	**.333**	<u>.734</u>	**.026**
MABEL (He et al., 2022)	<u>.864</u>	**<u>.008</u>**	.591	**.304**	.718	**.029**
INLP (Ravfogel et al., 2020)	N/A	**N/A**	N/A	**N/A**	.693	**.016**
ProSocialTuning	.738	**.013**	.494	**<u>.280</u>**	.674	**<u>.008</u>**

Table 1: Extrinsic Bias Evaluation on BERT-base and RoBERTa-base with Three Downstream Benchmarks: NLI-bias, BiasBios, and STS-B. Both accuracy and bias are reported; the optimal result is highlighted with <u>underline</u>. Please note: MABEL is pretrained with additional data augmented with SNLI and MNLI datasets, thus its accuracy on NLI-bias should be better than other methods. We did focus on propagating debiaisng from the debiased model to fine-tuned model, and the accuracy of ProSocialTuning is mainly determined by the steps of CDA.

of these backbone models. Masked PLMs are better suited for testing our technique than autoregressive models, e.g. the GPT family, for three main reasons. First, our solution is based on Causal Mediation Analysis and PAC-Bayes training, *both of which are model-agnostic*. Second, GPT-2 has been reported to be unstable for classification tasks (Radford et al., 2019; Liu et al., 2023b), which are used to test the effectiveness of our technique. Lastly, the strong correlation between social groups and labels on classification tasks makes them more challenging to debias than text generation tasks in terms of relearning social bias. This issue can more easily be mitigated for text generation tasks, such as those performed by the GPT family of models, by intervening the generation-time sampling (Yang et al., 2022). The latter two reasons further contribute to the difficulty in distinguishing the effects of debiasing methods from the unsatisfactory performance of an autoregressive model for this task.

For implementing mitigation of gender bias through counterfactual data augmentation, we follow Kaneko et al. (2022) to rebalance the debiasing corpus[6] with gender words from Zhao et al. (2018). We run 150 epochs for debiasing both backbone models. The StereoSet score (Nadeem et al., 2021b) is used as the intrinsic bias evaluation metric over Masked PLMs; we conduct extrinsic bias evaluation over fine-tuned PLMs with three tasks, e.g., STS-B (Cer et al., 2017), BiasBios (De-Arteaga et al., 2019), and NLI-bias (De-Arteaga et al., 2019). For NLI-bias we randomly sample 10,000 instances from the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) as training data and development data, and we generate 20,000 test samples with words related to male and 20,000 test samples with words related to female as defined by De-Arteaga et al. (2019). We sample 20,000 training samples from the training set for NLI-bias and BiasBios, but use all training data in STS-B. To implement causal mediation analysis, we re-use the Winograd-schema-style examples from Vig et al. (2020).

To validate the performance of ProSocialTuning, we implement experiments with the following models: (1) **Vanilla-tuning**: fine-tunes a model without any debiasing operations; (2) **Debiased-tuning**: fine-tunes a debiased model with downstream task-specific data, where the performance should be the upper bound with respect to that of ProSocialTuning; (3) **EAR** (Attanasio et al., 2022): attention-based debiasing method, which introduces a regularization term for minimizing the entropy of attention; (4) **MABEL** (He et al., 2022): enhances CDA by pretraining PLMs with natural language inference datasets, e.g., SNLI and MNLI, and is a supervised way to implement task-agnostic debiasing; and (5) **INLP** (Ravfogel et al., 2020): a task-dependent debiasing method, which removes gender information in sentence representations by projection. INLP iteratively trains linear classifiers that predict a certain undesired property and then exploits nullspace projection to make the classifiers oblivious to the undesired property. Details of the hyperparameters and implementations are available

[6]https://data.statmt.org/news-commentary/v15/

in Appendix A.2.

6.2 Main Results

Table 1 shows the extrinsic bias evaluation[7] results of the two backbone models of BERT-base and RoBERTa-base with three downstream fine-tuning datasets[8]. Table 2 indicates the intrinsic bias score of the model achieved with ProSocialTuning and the debiased model. Note that we do not pursue a SOTA debiasing method because our aim is to understand how the mechanism of forgetting causes the relearning of social bias during downstream fine-tuning. Regarding the accuracy of ProSocialTuning, it is determined by the performance of the debiased model. The low accuracy of ProSocialTuning can be straightforwardly resolved by taking a fusion strategy over the prediction of the debiased model and the original one (Liang et al., 2021), but this is not the focus of this paper. ProSocialTuning is proven effective at mitigating relearning social bias as long as its bias score is lower than that of the *Debiased-tuning* model.

Overall, ProSocialTuning achieves the best bias score for all downstream fine-tuning tasks, except the NLI-bias dataset with RoBERTa model, wherein MABEL outperforms other methods in both accuracy and bias. The bias score gap between ProSocialTuning and other methods is rather large for the task of BiasBios. This is because the causal mediation analysis is done with a corpus portraying gender occupation association but the association does not exist in other tasks. However, the downstream task-specific performance with CDA prohibits widespread usage owing to its negative impact on the language modeling ability.

In contrast to ProSocialTuning, other task-agnostic debiasing methods exhibit inconsistencies across diverse experimental setups. For instance, EAR demonstrates good accuracy and bias score improvements when applied to the BERT backbone model in the NLI-bias task. However, in certain scenarios, its bias score surpasses even that of the Vanilla-tuning method, as reported by Gaci et al. (2022b). Similarly, MABEL showcases increased bias compared to Vanilla-tuning in the STS-B task, highlighting the inefficiency of a purely task-agnostic debiasing approach devoid of interventions during downstream fine-tuning processes.

[7]More details about bias score calculation are available in Appendix A.3.

[8]All experiments are run with 3 seeds (1, 42, 100); reported performance scores are the average over three experiments.

The strong inconsistency of those baseline debiasing methods demonstrates debiasing performance cannot be propagated without solving the forgetting issue of PLMs. As a task-dependent debiasing method, INLP achieves rather good accuracy and debising performance given the RoBERTa model and the BiasBios dataset but it leads to a highly biased fine-tuned model with BERT. Since it requires the annotation of gender information of each sample, the experimental result is only available for the BiasBios dataset.

StereoSet Score	STS-B	NLI-bias	BiasBios
DEBIASED	53.20	53.20	53.20
Debiased-tuning	$54.53_{\uparrow 1.33}$	$54.94_{\uparrow 1.74}$	$54.78_{\uparrow 1.58}$
ProSocialTuning	$53.55_{\uparrow \mathbf{0.35}}$	$53.96_{\uparrow \mathbf{0.66}}$	$54.67_{\uparrow \mathbf{1.37}}$

Table 2: StereoSet Scores of Fine-tuned Models with Various Methods. DEBIASED reports the bias score of the debiased model using CDA. The closer the model's bias approaches 50, the lower its level of bias.

Table 2 shows the intrinsic bias score of fine-tuned BERT models with various methods. Given the bias score of the debiased model as 53.20, directly fine-tuning the debiased model results in an obvious increase of bias level. Furthermore, the increases associated with Debiased-tuning are over 1.0 after training with three datasets. In contrast, ProSocialTuning leads to a smaller increase of bias levels. For the downstream task of BiasBios, ProSocialTuning is close to Debiased-tuning; this is due to the higher bias level of the dataset by referring to the high bias core of Vanilla-tuning.

For more details about the ablation study, Appendix A.4 shows the results supporting the necessity of each component in ProSocialTuning.

7 Conclusion

This work addresses the ongoing debate surrounding the effectiveness of task-agnostic debiasing techniques for downstream tasks. Our research reveals a pivotal factor determining the effectiveness of debiasing: the bias level of the debiased model and the downstream task dataset. Specifically, the bias level of the debiased model serves as the lower bound for bias in fine-tuned tasks wherein social bias exists. To gain in-depth understanding of how forgetting changes PLMs' parameters, we introduce ProSocialTuning, a novel framework which mitigates the diminishing effectiveness by imposing regularization on attention heads that have already undergone successful debiasing.

8 Limitations

In this paper, we only consider two backbone models of BERT-base and Roberta-base due to hardware constraints. However, larger models are more vulnerable to social bias, therefore the analysis of bias level disparity should be done for larger PLMs. On the other hand, ProSocialTuning depends on the results of causal mediation analysis; specifically for this work, the prompts should be relevant to gender bias towards occupations in order to align causal mediation analysis with the downstream fine-tuning tasks of occupation prediction. For other downstream fine-tuning tasks such as STS-B and NLI-bias, the corpus for causal mediation analysis should be redesigned. Additionally, we omit the influence of the adapted classification layer in Section 3 by validating the intrinsic bias scores and language modeling ability. Given the smaller size of parameters, this omission of the adaptation layer is expected to be safe.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, Elena Baralis, et al. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119. Association for Computational Linguistics.

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*, pages 610–623.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Fatma Elsafoury and Gavin Abercrombie. 2023. On the origins of bias in nlp through the lens of the jim code. *arXiv preprint arXiv:2305.09281*.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022a. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yacine Gaci, Boualem Benattallah, Fabio Casati, and Khalid Benabdeslem. 2022b. Debiasing pretrained text encoders by paying attention to paying attention. In *2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702.

Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Sullam Jeoung and Jana Diesner. 2022. What changed? investigating debiasing methods using causal mediation analysis. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 255–265, Seattle, Washington. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Woosuk Kwon, Sehoon Kim, Michael W Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. 2022. A fast post-training pruning framework for transformers. *Advances in Neural Information Processing Systems*, 35:24101–24116.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Guangliang Liu, Zhiyu Xue, Xitong Zhang, Kristen Marie Johnson, and Rongrong Wang. 2023a. Pac-tuning: Fine-tuning pretrained language models with pac-driven perturbed gradient descent. *arXiv preprint arXiv:2310.17588*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023b. Gpt understands, too. *AI Open*.

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. *arXiv preprint arXiv:1907.11692*.

Christos Louizos, Max Welling, and Diederik P Kingma. 2018. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*.

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

David A McAllester. 1998. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234.

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)*, pages 1878–1898.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, pages 2498–2507. PMLR.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021a. 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.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021b. 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.

Debora Nozza, Federcio Bianchi, Dirk Hovy, et al. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvári. 2019. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*.

Eric Michael Smith, Melissa Hall Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. " i'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 3(2).

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Michael E Tipping. 2001. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. 2022. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. *arXiv preprint arXiv:2210.04492*.

Abdelrahman Zayed, Goncalo Mordido, Samira Shabanian, and Sarath Chandar. 2023. Should we attend more or less? modulating attention for fairness. *arXiv preprint arXiv:2305.13088*.

Xitong Zhang, Avrajit Ghosh, Guangliang Liu, and Rongrong Wang. 2023. Auto-tune: Pac-bayes optimization over prior and posterior for neural networks. *arXiv preprint arXiv:2305.19243*.

Jiachen Zhao, Zhun Deng, David Madras, James Zou, and Mengye Ren. 2023. Learning and forgetting unsafe examples in large language models. *arXiv preprint arXiv:2312.12736*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023a. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023b. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

A Appendix

Hyperparameters	Setting
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	1e-3
Learning rate for θ	5e-5
Learning rate for ω	1e-2
Maximum training epochs	25
Weight decay	0.01
Batch size	64

Table 3: Hyperparameter Settings for the AdamW Optimizer.

A.1 Details about Generalization Importance Estimation

In contrast to Molchanov et al. (2017), we fix $\mathcal{P}$ by a re-scaled parameter-wise logarithm prior where the prior noise variance is initialized as the absolute value of the parameter weights. Furthermore, fine-tuning a PLM-based classifier should assign different learning rates for the pretrained layers and the adapted classification layer, respectively. The difference in confidence w.r.t. pretrained layers and adaptation classification layers is also considered through leveraging a lower learning rate to update dimensions, in q, associated with pretrained layers and a higher learning rate for dimensions relevant to the adaptation layers.

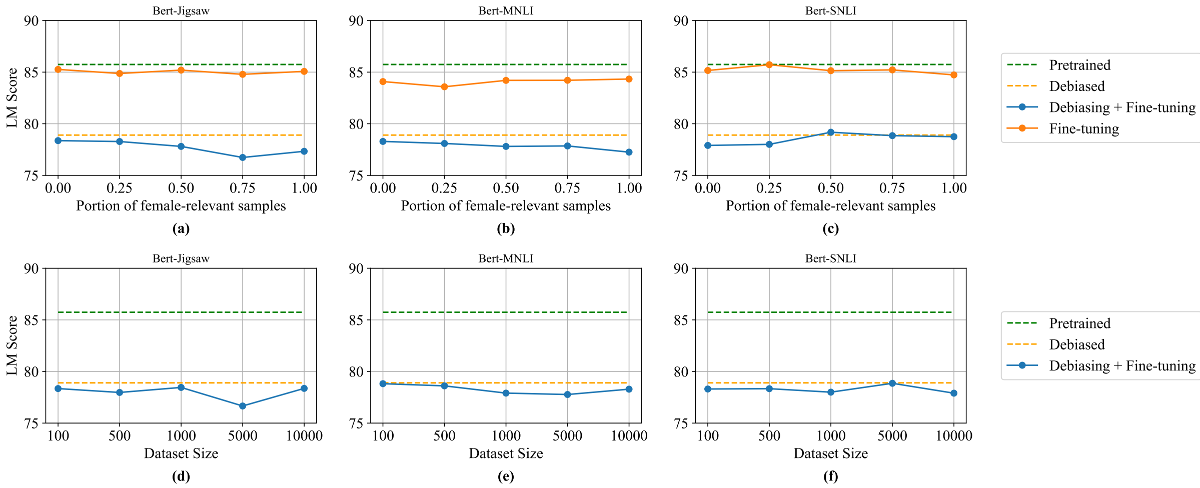

Figure 3: Language Modeling Scores. These figures present the language modeling scores of the pretrained, debiased, and fine-tuned models with respect to different bias levels and dataset sizes in downstream tasks.

A.2 Implementations

Figure 3 introduces the hyperparameters used for fine-tuning. We add an adapted layer of fully-connected forward neural network as the classification layer beyond a PLM. For all experiments except the CDA, we freeze the embedding layers of PLMs. For the generalization estimation driven by PAC-Bayes training, we first fine-tune models with 35 epochs to make them fit the task-specific data well. In the stage of generalization importance estimation, we initialize both the prior and posterior noise variance with $\log(0.001 \cdot |q_i|)$ where q_i is the i^{th} parameter of the final classification model. The noise parameter dimensions associated with the pretrained layers and classification layer are 0.01 and 0.1 respectively.

For the EAR method, we take regularization terms of 0.001, 0.01, 0.1, 1.0 and report the best downstream performance and bias scores. To implement MABEL, we directly leverage the open-source checkpoints[9] from HuggingFace as the debiased model and fine-tune it with downstream task-specific data. In the implementation of ProSocial-Tuning, we have the regularization γ hyperparameter space of 0.001, 0.01, 0.1, 1.0. For the INLP method, first, we fine-tune the classification model with 25 epochs to fit the data well and select the best model. Then, we iteratively train 300 linear SVM classifiers to fit the data concerning gender labels, and exploit nullspace projection to remove the gender information. Finally, we freeze the PLMs and train only the classification layers to fit the debiased representations.

A.3 Bias Score

Following Kaneko et al. (2022), we create the bias evaluation datasets w.r.t. different genders. For the BiasBios, we calculate the TPR score difference between male-relevant evaluation samples and female-relevant evaluation samples. For the NLI-bias dataset, we calculate the difference between the ratios w.r.t. classifying male-relevant evaluation samples to the label of neutral and w.r.t. classifying female-relevant evaluation samples to the label of neutral. For the STS-B dataset, we create parallel bias evaluation corpus w.r.t. genders, and we calculate ratio of how many parallel samples are predicted with the same label. Then we take the difference of this ratio to 1 as the bias score.

A.4 Ablation Study

Table 4 shows the experimental results of the ablation study, proving the necessity of generalization-guided regularization over successfully debiased attention heads. The generalization-guided regularization alleviates the negative impact on downstream task-specific performance and keeps those debiased attention heads to avoid relearning too many biases during downstream fine-tuning.

[9] https://huggingface.co/princeton-nlp/mabel-bert-base-uncased and https://huggingface.co/princeton-nlp/mabel-roberta-base

	STS-B Accuracy	STS-B Bias
Random Attention	.459	.216
Uniform Regularization	.455	.180
ProSocialTuning	.460	.177

Table 4: Ablation study for ProSocialTuning. We consider Random Attention to randomly pick up attention heads to regularize during downstream fine-tuning. For Uniform Regularization, we do not apply generalization-guided regularization but take uniform regularizations.