PEFTDB : Parameter Efficient Debiasing of Language models across multiple bias axes

Aditya Srikanth Veerubhotla [*] **Srijan Bansal** [*] **Sumit Agarwal** [*]
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
{adityasv, srijanb, sumita}@andrew.cmu.edu

Abstract

In this research paper, we introduce PEFTDB, a novel approach for parameter efficient debiasing of language models. PEFTDB operates in two distinct phases: an upstream phase to acquire parameter efficient debiasing parameters along a specific bias axis, and a downstream phase where these parameters are frozen during the finetuning process. Through evaluation on four datasets and two bias axes, we observe that prompt tuning and sparse fine tuning exhibit the highest efficiency in downstream debiasing. Furthermore, we demonstrate that these parameters possess task-agnostic characteristics, enabling their effective application in mitigating biases in similar tasks across different domains. The code for reproducing our experiments can be found here.

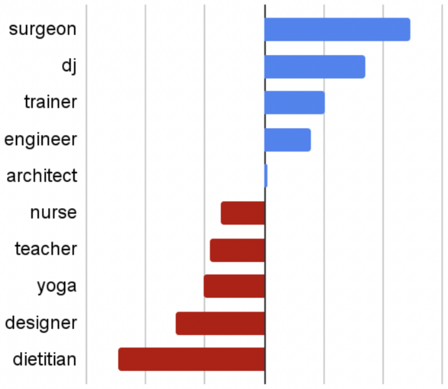


Figure 1: True Positive Rate difference between male and female per profession on predictions made by BERT on BiasBios. Red indicates he associations of female-dominated professions with females, while the correlation between male-dominated professions and male gender is represented by blue.

1 Introduction

In recent years, it has become increasingly evident that NLP models are susceptible to biases present in the training data, which can lead to unfair or discriminatory outcomes. (Hardt et al., 2016) defines bias as “disparate model performance on different subsets of data associated with different demographic groups". We focus on this definition of bias which is expressed in multiple axes like gender, race, and religion (Meade et al., 2021).

Models trained on datasets containing biases often absorb these biases due to correlations between protected attributes and labels present in the data. For instance, a BERT model might misclassify a biography that says *“His portfolio includes the most luxurious, beautiful, expensive, and large-scale projects in the world. For example, he made interior design"* as belonging to an "Architect" instead of the correct label "Interior Designer," possibly because it has learned to associate Interior Design with females, and not males. In Figure 1, we observe the disparity in true positive rates between male and female predictions for various occupations. The model exhibits a bias towards males for professions such as surgeon, engineer, and architect (indicated by the color blue), while demonstrating a bias towards females for occupations like designer and nurse (highlighted in red).

Debiasing techniques aim to mitigate biases in machine learning models, reducing their reliance on biased features and promoting equitable decision-making. Various approaches have been proposed to tackle this challenge, ranging from data augmentation techniques (Zmigrod et al., 2019; Meade et al., 2021) to specialized regularized losses (Kennedy et al., 2020; Zhang et al., 2018). However, many existing debiasing methods suffer from the limitation of training the entire model from scratch, which can be computationally expensive and resource-intensive.

The emergence of parameter-efficient techniques (PEFTs) in various NLP tasks has sparked interest in their potential application for model debiasing. Adapter-based techniques have shown promising

[*] Equal contribution.

results in debiasing which are comparable performance to full models while utilizing only a fraction of the model parameters (Kumar et al., 2023; Lauscher et al., 2021). In this research paper, we aim to address three key research questions:

- RQ1 : Can PEFTs effectively capture task-agnostic debiasing information that can be utilized in downstream tasks?
- RQ2: Are all PEFTs equally efficient in mitigating biases?
- RQ3 : Does the task-agnostic patch generated by PEFTs work effectively across different datasets within a similar domain?

We propose a novel approach called Parameter Efficient Debiasing (PEFTDB) as an alternative to traditional full model training for debiasing. PEFTDB consists of two phases: an upstream phase that captures debiasing information using PEFTs along a specific bias axis (e.g., gender) and a downstream phase where these debiasing parameters are frozen during model fine-tuning. We employ counterfactual data augmentation (CDA) on the source data, replacing bias attribute words to create augmented training examples.

Our investigation focuses on evaluating the effectiveness of PEFTDB across different PEFTs and analyzing multiple bias axes, including gender and group biases. Gender bias assessment is conducted using the BiasBios dataset, commonly used in occupation prediction tasks. Group bias, associated with race, religion, and sexual orientation, is evaluated using datasets such as StormFront, GAB, and FDCL. We also analyze the generalizability of our task-agnostic parameters captured along a specific bias axis by examining their effectiveness on different downstream datasets exhibiting bias along the same axis.

Through our experiments, we provide compelling evidence that PEFT-DB is highly effective in mitigating biases in downstream tasks, specifically across the gender and group axes. Among the various PEFT approaches we examined, Prompt Tuning and Sparse Fine Tuning consistently outperformed other techniques, highlighting their superiority in debiasing. Furthermore, we demonstrate the task-agnostic nature of our approach by leveraging group-based parameters learned from datasets such as Stormfront or FDCL during the upstream phase. These parameters can successfully debias models when applied to the GHC dataset. This finding emphasizes the versatility of our debiasing approach, as the task-agnostic parameters acquired from one dataset can effectively address biases in a different dataset within the same group axis.

2 Related Work

Pretrained language models have achieved remarkable results in various natural language processing tasks, but they also exhibit different types of biases, such as gender bias (Kurita et al., 2019) and ethnic bias (Ahn and Oh, 2021). Several methods have been proposed to mitigate these biases in language models, such as counterfactual data augmentation (Zmigrod et al., 2019), dropout regularization (Webster et al., 2020), null-space projection (Ravfogel et al., 2020), adversarial training (Liu et al., 2020), contrastive learning (He et al., 2022), and meta-learning (He et al., 2022). An important question is how bias mitigation affects the performance of the language models on downstream tasks. A comprehensive study by (Meade et al., 2021) shows that bias mitigation can be performed without much degradation in task performance.

However, these techniques are often performed along with the downstream task, and hence, require additional annotation (such as the protected attributes) along with the task data. This could have significantly increased the annotation costs in the data, preventing scaling. Instead, (Jin et al., 2021) perform debiasing in the “Upstream" level, before the task and show that debiasing a language model in the target domain can improve its generalization. As a follow-up, (Steed et al., 2022) show that debiasing a LM before fine-tuning does not guarantee that the fine-tuned model will be unbiased.

This could be due to the fact that the entire model is being fine-tuned on the downstream task, and losing its debiased representations. A possible solution is to use PEFTs such as Adapters (Houlsby et al., 2019a). This has been explored by prior literature (Lauscher et al., 2021; Kumar et al., 2023). They show that parameter-efficient techniques can be used to debias language models in a parameter-efficient way while keeping the LM backbone frozen. This has added the benefit of reduced computational cost and the environmental impact of debiasing large language models (Strubell et al., 2019), and potentially in preventing catastrophic forgetting of pre-trained knowledge due to fine-tuning (Kirkpatrick et al., 2017). However, these techniques are performed in downstream and suffer from the drawbacks discussed earlier. To bridge this gap, we study the different parameter-efficient debiasing techniques in an upstream manner, providing a task-independent and parameter-efficient method to effectively adapt debiased models on the target task.

3 Bias Factors and Datasets

We substantiate our hypothesis by conducting validation on two well-established bias factors: Gender Stereotyping and Group Identifiers, which have been extensively investigated in previous research. To ensure a comprehensive analysis, we utilize four diverse datasets that cover a wide range of domains and tasks, as outlined in Table 1. To evaluate the effectiveness of our debiasing techniques in addressing gender and racial biases, we employ two intrinsic bias benchmarks, namely CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021), during the initial **upstream** stage of our evaluation. For the subsequent **downstream** stage, we assess the performance gap of the debiasing methods among protected attributes within the specific domain and utilize extrinsic bias metrics as described in Section 3.4.

Dataset	Bias	Task	Domain
BiasBios	Gender	Occupation	commoncrawl.org
GHC	Group	Hate	gab.com
Stormfront	Group	Hate	stormfront.org
FDCL	Group	Toxicity	twitter

Table 1: The table shows the different datasets and bias axes that we have considered. The group bias axis is a combination of Race, Religion, and Sexual Orientation.

3.1 Gender Stereotypical Bias

Zhao et al. (2018) demonstrated that biases arise in occupation prediction models when trained on short bios from the BiasBios dataset (De-Arteaga et al., 2019) which is a collection of 397K biographies spanning twenty-eight occupations written in English. These biases in this dataset stem from the pervasive influence of occupation-based stereotypes. We report accuracy scores for task performance on the BiasBios dataset.

3.2 Group Identifier Bias

The elevated occurrence of false positive outcomes in hate speech predictions, particularly in sentences containing specific group identifiers related to races, religions, or sexual orientations, has detrimental effects on protected groups. Our investigation entails the utilization of three distinct corpora: the Gab Hate Corpus (GHC; Kennedy et al. (2018)), the StormFront corpus (de Gibert et al., 2018), and the FDCL (Founta et al., 2018). These corpora employ binary labels but adopt diverse labeling schemes and domains. Due to the inherent imbalance of these datasets, we employ F1 scores as a metric to assess task performance.

3.3 Intrinsic Evaluation

StereoSet (Nadeem et al., 2021) measures a language model's stereotypical associations using fill-in-the-blank problems with intra-sentence examples across four bias categories. The Language Modeling Score (LM) is the percentage of instances where the model picks a valid word (either the stereotype or the anti-stereotype) over a random word, and the Stereotype Score (SS) measures the percentage of stereotypical choices over anti-stereotypical ones. The Idealized Context Association Test (ICAT) combines LM and SS scores into one metric.

CrowS-Pairs (Nangia et al., 2020) is an intra-sentence dataset of minimal pairs that compares the language model's masked token probabilities of sentences with disadvantaged or advantaged groups fulfilling or violating stereotypes. This evaluation metric supports 11 bias categories and reports the Stereotype Score (SS), measuring the model's preference for stereotypical sentences over anti-stereotypical ones.

3.4 Extrinsic Evaluation

Gender Stereotype: To quantify gender bias, we follow the approach proposed by De-Arteaga et al. (2019) and compute the true positive rate (TPR) gender gap—i.e., the differences in the TPRs between genders, respectively—for each occupation. The TPR gender gap between male (m) and female (f) for occupation c is defined as follows:

$$TPR_{RMS} = \sqrt{\frac{1}{|C|}\sum_{y \in C}(TPR_{m,y} - TPR_{f,y})^2}$$

where $TPR_{m,y}, TPR_{f,y}$ denote true positive rate for occupation y where gender protected attribute is male and female respectively.

Group Identifier Bias: We quantify the False Positive Rate Differences (FPRD) by comparing the

FPR of examples that mention one of the protected attributes (z) with the overall FPR.

$$FPRD = \sum_{z} |FPR_z - FPR_{all}|$$

.

In our evaluation, we calculate the False Positive Rate Difference (FPRD) bias metric for both the in-domain data and its corresponding extrinsic dataset. Specifically, for the assessment of Group Identifier bias, we incorporate the Identity Phrase Templates Test Sets (IPTTS) (Zhang et al., 2020). This test set consists of 77,000 instances comprising hate and non-hate sentences that mention 25 group identifiers, generated using predefined templates.

4 Bias Statement

Our analysis in Figure 5 examples from two datasets: BiasBios and GHC corpus, both of which exhibit inherent biases. In the BiasBios dataset, we observe a higher representation of male examples compared to female examples in professions such as software engineering and architecture. This disproportionate representation can introduce biases into models trained on this dataset. Similarly, in the GHC corpus, we find that certain identity words such as "Hispanic," "Muslim," and "homosexual" have a higher probability of being associated with hateful content compared to their counterparts. These biases in the dataset can propagate inappropriate stereotypes if models are trained without appropriate mitigation strategies.

It is important to note that these biases, such as the unequal representation of male and female examples or the disproportionate association of certain identities with hate speech, are not reflective of the ideal scenario. Ideally, datasets should exhibit balanced representations or fair associations between attributes and labels. However, the presence of such biases highlights the need for effective debiasing techniques to mitigate these unwanted correlations and promote fair and unbiased decision-making. Our research aims to tackle these needs by proposing and evaluating parameter-efficient debiasing techniques that can effectively mitigate biases while training on these datasets without a compromise in the model's performance.

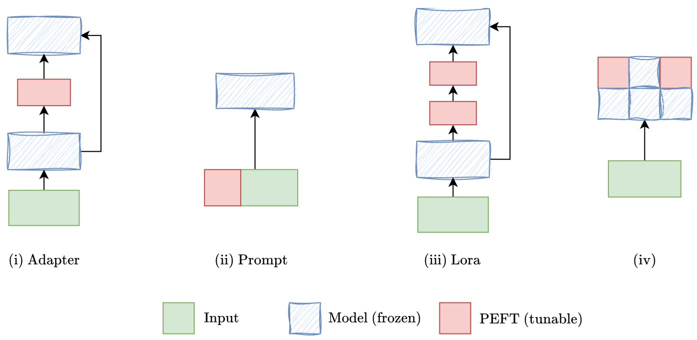
5 Parameter-Efficient Technique (PEFT)

Transfer learning from pre-trained language models (PLMs) excels in natural language processing (Devlin et al., 2019; Qiu et al., 2020), delivering impressive results across tasks. However, fine-tuning all model parameters for multiple tasks becomes expensive, especially with larger PLMs containing billions or trillions of parameters. To address this challenge, prior works have proposed lightweight alternatives that modify only a few extra parameters while keeping the majority of pre-trained parameters intact. Consequently, these methods enable separate and simultaneous training for multiple tasks while keeping the pre-trained model fixed, thereby significantly reducing the computational cost associated with the process.

Adapter: Houlsby et al. (2019b) introduced adapters as task-specific modules inserted between transformer layers. Adapters consist of a down-projection, a nonlinear activation function, and an up-projection using parameter matrices, connected to transformer layers through a residual connection. Various studies propose different adapter placement strategies within the transformer layers. Houlsby et al. (2019b) suggest placing two adapters sequentially within one layer, while Pfeiffer et al. (2021) propose a more efficient variant inserted after the FFN "add & layer norm" sub-layer, which we adopt in this paper. **Prompt Tuning** : Previous research (Lester et al., 2021; Li and Liang, 2021) has introduced prompt-tuning as a lightweight alternative to fine-tuning. Prompt-tuning involves incorporating task-specific vectors, referred to as prompts, into the input sequence. These prompt vectors are treated as "virtual tokens" within the transformer model, facilitating the generation of the desired output more effectively. In our implementation, we simply add the prompt vectors to the input word embeddings in the initial layer.

LoRA Adapter : LoRA (Hu et al., 2021) integrates trainable low-rank matrices into transformer layers in order to approximate weight updates. When confronted with a pre-trained weight matrix, LoRA represents its update using a low-rank decomposition that encompasses adjustable parameters. Through the optimization of the rank decomposition matrices, LoRA facilitates the indirect training of specific dense layers in a neural network during adaptation, thereby eliminating the necessity of retraining all model parameters.

LT-Sparse Fine Tuning : Ansell et al. (2022) builds upon the Lottery Ticket Hypothesis (LTH) and presents a method for pruning large neural networks. The approach involves fine-tuning a pre-



(i) Adapter

(ii) Prompt

(iii) Lora

(iv)

Figure 2: Illustration of parameter-efficient training (PEFT) methods: (i) Adapters (Pfeiffer et al., 2021), (ii) Prompt Tuning (Lester et al., 2021), (iii) LoRA (Hu et al., 2021), and (iv) LT Sparse Finetuning (Ansell et al., 2022). Red dashed boxes indicate trainable PEFT parameters, while blue boxes represent frozen backbone model parameters.

trained model and selecting a sparse subnetwork based on the parameters that undergo the most significant changes. Instead of zeroing out values like the original LTH algorithm, the model is reset to its pretrained initialization. The selected subset of parameters is then fine-tuned again, resulting in a sparse representation that captures deviations from the pretrained model. Multiple sparse representations can be combined by summing them with the pretrained model.

6 Counterfactual Data Augmentation

Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019) is a data-based debiasing technique that swaps attribute words pertaining to a bias (e.g, he/she for gender) in a corpus. Essentially it tries to rebalance the corpus with respect to the bias axis so that the model sees similar amounts of attribute words and gets debiased. While CDA has been mainly used for gender debiasing, we also use CDA for different bias axes by defining separate attribute words for each bias. For example, black can be replaced with white, church with mosque, he with she, to generate counterfactual examples. The bias attribute words used for different axes are mentioned in A.1.

7 Methodology

Kumar et al. (2023) shows that adapters debiased while finetuning works better, i.e., while (Lauscher et al., 2021) shows that learning adapters in the upstream phase helps in downstream finetuning. We propose a different approach of parameter efficient debiasing which is a mixin of both of these approaches, PEFTDB. It comprises of two main phases : **Upstream Phase** which is responsible for selecting debiasing parameters, **Downstream Phase** which uses the debias parameters as a patch for task debiasing in the downstream phase. PEFTDB works on source data s in the upstream phase and works on a target data t in the downstream phase where the debiasing happens throughout an axis a with PEFT p.

7.1 Upstream Phase

Gururangan et al. (2020) shows that performing continued pretraining on BERT on domain specific data helps improve performance on downstream tasks in the same domain. Further, (Meade et al., 2021) shows that Counterfactual Data Augmentation (CDA) is a universal debiasing technique that can be applied to different axes controlled by bias attribute words. We combine these two approaches to perform CDA on domain specific data.

Parameter efficient debiasing with Adapters (Lauscher et al., 2021) using CDA has shown to be effective in capturing debiasing information with using a reduced number of parameters. Consequently, we investigate the application of domain-specific CDA using parameter-efficient approaches (PEFTs described in 5) to obtain debiasing parameters. Specifically, we employ a PEFT (p) to perform CDA on the source data (s) using attribute words from particular axis (a), resulting in debiasing parameters (p_a). We hypothesize that these parameters will effectively capture task-agnostic debiasing information specific to the given axis (a).

7.2 Downstream Phase

Kumar et al. (2023) has demonstrated that incorporating adapters in downstream task tuning yields improved results. However, this approach necessitates the learning of a parameter-efficient module for each individual task, rendering it infeasible to transfer adapters learned from

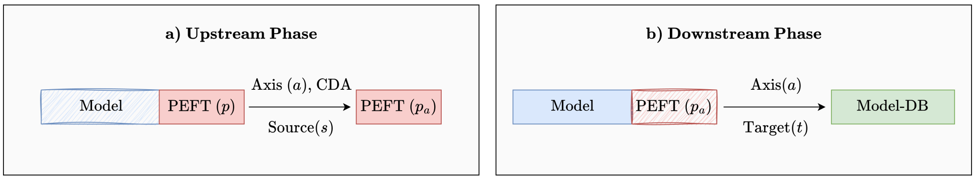


Figure 3: A detailed diagram of the two phases of PEFTDB - a) Upstream Phase which performs CDA across axis a on source s using PEFT p to get debias parameters p_a, b) Downstream Phase which freezes these debias params p_a while finetuning model on target t to get debiased model along axis a.

one task to others within the same domain. Consequently, we propose a different strategy for achieving parameter efficiency by learning debiasing parameters during the upstream phase, as suggested by (Lauscher et al., 2021). In contrast to their approach, we maintain the debiasing parameters in a fixed state during downstream task finetuning. We hypothesize that this frozen parameter setting preserves the upstream debiasing effect and prevents the model from acquiring biases from the training data while task finetuning. We utilize the debiasing parameters (p_a) obtained in the upstream phase as a patch to the model prior to fine-tuning it on the target task data (t). This debiasing technique effectively mitigates biases along the specific axis (a) in the finetuned model.

Figure 3 illustrates the two phases of our proposed approach, namely the upstream and downstream phases of Parameter Efficient Fine-Tuning with Debiasing (PEFTDB). In our experiments, we investigate scenarios where the source data and target data are either identical (same task) or originate from similar domains (cross task). We make the assumption that both the source and target data exhibit biases along a shared axis.

8 Experimental setup

We used pre-trained BERT (Devlin et al., 2018) as the starting point for all of our models. We also applied text normalization to FDCL, GHC and Stormfront datasets to remove URLs and user mentions using tweet based processing [1]. For the upstream experiments, we trained our models with MLM and CDA on the BiasBios dataset and the other datasets using a learning rate of $1e^{-5}$ and a batch size of 128 and 32 respectively. We ran MLM for 10,000 steps and evaluated the models every 1,000 steps. We selected the models with the lowest loss for our experiments. For the downstream experiments, we used a batch size of 32 and trained our models for 10 experiments. We ensured that all PEFTs have similar number of parameters, being 1% of the base LM. For the downstream experiments, we used a batch size of 32 and trained our models for 10 epochs. We chose the models with the best task metrics for analysis. For GHC and Stormfront datasets, which had few hateful examples compared to non-hateful ones, we weighted the loss of hateful examples by a factor of 10 for GHC and 6.7 for Stormfront, based on their proportions in the data. We compared our methods with two baselines: BERT in the pre-trained setting and BERT in the fine-tuned setting (Full-Debias). We based our implementation on the code from AdapterHub [2].

9 Results

9.1 Upstream Phase

The results of the Upstream setting are presented in Table 2. In this analysis, we focus on the CrowS pairs scores as the StereoSet Score (SS Score) yielded inconsistent trends due to the low quality of data annotated for bias analysis, as also observed in prior work (Nangia et al., 2020)

The results demonstrate that utilizing PEFTs with CDA not only improves the language model (LM) performance but also reduces intrinsic bias. Among the different datasets, the best LM score and performance are achieved on BiosBias. Further analysis reveals a positive correlation between the SS LM score and the CrowS score, indicating that improving LM performance often involves adopting shortcuts offered by biases.

Notably, both Prompt and Adapter techniques exhibit strong debiasing performance while either retaining or even improving the LM score compared to other techniques. This suggests that these techniques effectively mitigate biases while main-

[1] https://github.com/Ashraf-Kamal/Hate_Speech_Detection/blob/main/Data_Preprocessing.py

[2] https://adapterhub.ml/

taining or enhancing the overall performance of the language model.

PEFT	**SS LM** ↑	**SS Score** ↓	**CrowS** ↓
BiosBias		**Eval** : Gender	
BERT	84.03	58.3	57.25
+ Full-Debias	84.79	58.75	54.96
+ Adapter	**85.52**	58.87	53.82
+ Prompt	85.35	**57.63**	**51.91**
+ LoRa	84.81	58.51	54.20
+ SFT	85.26	58.83	55.34
GHC		**Eval** : Race	
BERT	83.88	57.06	62.33
+ Full-Debias	84.01	**57.03**	**45.63**
+ Adapter	**85.88**	58.56	55.15
+ Prompt	85.73	58.78	52.62
+ LoRa	84.89	58.20	56.12
+ SFT	85.42	58.91	54.76
Stormfront		**Eval** : Race	
BERT	83.88	57.06	62.33
+ Full-Debias	84.01	**57.03**	**55.15**
+ Adapter	84.68	58.50	57.86
+ Prompt	**85.13**	59.03	58.83
+ LoRa	83.90	58.70	55.92
+ SFT	84.47	59.30	56.70
FDCL		**Eval** : Race	
BERT	83.88	57.06	62.33
+ Full-Debias	84.01	**57.03**	**55.34**
+ Adapter	**85.05**	59.18	63.11
+ Prompt	84.70	58.71	65.63
+ LoRa	84.46	59.42	63.50
+ SFT	84.74	59.60	61.94

Table 2: Results in the Upstream setting uinsg BERT as the LM and CDA for performing Debiasing.

9.2 Downstream Phase

Table 3 shows the results of our experiments in the downstream setting across 4 different datasets. Notably, the BiasBios dataset demonstrates comparable performance across all PEFTs, measured by Accuracy, while datasets such as Stormfront, GHC, and FDCL (representing hate speech) exhibit similar trends in F1 score compared to full finetuning (FT). These results suggest that **PEFTs can effectively capture debiasing information that can be applied to downstream tasks.**

Analyzing the BiasBios dataset, we observe that Prompt Tuning and Sparse Fine-Tuning (SFT) techniques yield the best performance based on the TPR-GAP measure. Similar trends are observed for Stormfront, where prompts, SFT, and LoRa demonstrate effectiveness according to in-domain False Positive Rate Difference (FPRD) and the extrinsic metric $\text{FPRD}_{\text{IPTTS}}$. In the case of the GHC dataset, prompts continue to be useful in reducing biases. For the FDCL dataset, all PEFT techniques consistently outperform finetuning in terms of bias metrics. Interestingly, in contrast to previous works that solely employ adapters for parameter efficient debiasing, we find that Prompt Tuning and Sparse Fine-Tuning outperform adapters in our downstream experiments. Hence, **all PEFTs are not equally efficient across different tasks**.

Another important observation from our experiments is the lack of strong correlation between the upstream and downstream metrics. This lack of correlation can be attributed to the inherent differences between the masked language modeling (MLM) task used during Counterfactual Data Augmentation (CDA) and the specific downstream application tasks. This finding aligns with the findings of (Goldfarb-Tarrant et al., 2021) who demonstrated no significant relationship between intrinsic and extrinsic bias metrics across a wide range of trained models covering various tasks.

10 Discussion

10.1 Cross Task results

To evaluate the task-agnostic nature of the learned upstream debiasing parameters, we conduct experiments where we apply these parameters during the finetuning process for a similar task in a different domain. Specifically, we consider a task on which the upstream model has not been trained. The results of this transfer experiment, focusing on debiasing across the group axis, are presented in Table 4.

By comparing these results with the ones reported in Table 3, we observe that the performance of the transferred debiasing parameters is comparable to that of full finetuning (FT). While parameters learned from the same task data exhibit the least bias, as indicated by the FPRD and $\text{FPRD}_{\text{IPTTS}}$ metrics, Table 4 demonstrates that comparable performance can still be achieved through transfer. Notably, the SFT and prompts techniques outperform full finetuning on in-domain FPRD metrics when it comes to transfer which also aligns with our findings from previous experiments. Despite some variations, the performance remains similar to that of full finetuning, indicating that **task-agnostic patch generated by PEFTs work effectively across different datasets within a similar domain**.

PEFT	BiasBios (Gender)		Stormfront (Group)			GHC (Group)			FDCL (Group)		
	ACC ↑	TPR-GAP ↓	F1 ↑	FPRD ↓	$\text{FPRD}_{\text{IPTTS}}$ ↓	F1 ↑	FPRD ↓	$\text{FPRD}_{\text{IPTTS}}$ ↓	F1 ↑	FPRD ↓	$\text{FPRD}_{\text{IPTTS}}$ ↓
FT	81.29	13.05	73.66	2.09	0.16	**68.76**	4.93	0.06	93.75	2.38	2.85
Adapter	81.28	13.22	73.36	2.67	0.08	66.35	4.11	0.18	93.86	1.46	2.80
Prompt	81.10	**11.98**	73.47	2.11	0.09	68.49	**2.75**	**0.05**	93.83	1.95	2.67
LoRa	81.28	13.67	74.80	1.92	0.24	68.21	4.60	0.49	**93.93**	**1.29**	**2.56**
SFT	**81.34**	12.04	**74.07**	**1.87**	**0.06**	66.49	6.19	5.24	93.89	1.41	2.74

Table 3: Results in downstream setting on different datasets where the source data and target data come from the same task across gender and group biases.

PEFT	Upstream: StormFront			Upstream: FDCL		
	F1 ↑	FPRD ↓	$\text{FPRD}_{\text{IPTTS}}$ ↓	F1 ↑	FPRD ↓	$\text{FPRD}_{\text{IPTTS}}$
FT	**68.76**	4.93	**0.06**	**68.76**	4.93	**0.06**
Adapter	66.72	4.55	1.47	68.00	6.70	4.50
Prompt	68.40	**3.25**	0.17	67.95	5.49	1.97
LoRa	66.92	5.92	4.33	67.01	5.00	1.38
SFT	66.65	3.53	0.65	68.05	**4.65**	0.54

Table 4: Cross task transfer results showing the generalizability of patch parameters learned using upstream data to GHC across group axis.

10.2 Reduction in bias

We conducted a comparison of the TPR-GAP performance of CDA debiasing techniques using FT and Prompt on the Biasbios dataset (see Figure 4, specifically focusing on occupations categorized as male and female. Our findings indicate that debiasing with Prompt yields better results compared to FT, as evidenced by a decrease in the TPR for gender-dominant professions. We observed that certain female-dominated professions such as dietitian and interior designer exhibit reduced correlation with the female gender, while male-dominated professions like surgeon and comedian also demonstrate a decrease in correlation with the male gender. Although we did not observe significant changes in the gap for professions like rapper and psychologist, we encountered an issue of over-correction, resulting in a reversed gap for poet and accountant. This discrepancy can be attributed to the limited number of examples available for these particular professions. We conducted a comparative analysis of false positive rate (FPR) performance across various group identifiers on the GHC dataset (see figure 6 using both the FT and Prompt models, incorporating the CDA debiasing technique. Our observations revealed that debiasing with Prompt tuning leads to improvements specifically for groups such as black, trans, Muslim, queer, and lesbian. These findings indicate the superiority of our methodology.

10.3 Qualitative Analysis

Table 5 shows provides a few examples where our models is able to prediction the correct outputs over the baseline. In the the first two examples which are from BiasBios, the occupation words are present in the text, but are not being detected by the BERT model, and the baseline uses the strong biased prior association between woman and physician, man and architect to make the predictions.

11 Limitation

Here we discuss the limitations of our work. Firstly, the study acknowledges that gender is non-binary, however, it does not explore the nuances of non-binary gender identities. Secondly, the statistical significance tests were performed using a limited number of seeds, which may raise questions about the generalizability of the findings. Thirdly, the study only focused on the BERT language model, limiting the scope of the research to other types of language models that may exhibit different behaviors. Finally, as the study only focused on classification tasks and the dataset was biased towards toxicity, the findings may not be generalizable to other types of tasks and datasets.

12 Conclusion & Future Work

By addressing the critical challenge of bias mitigation while prioritizing parameter efficiency, this research aims to contribute towards the development of more fair, interpretable, and scalable machine learning systems. Ultimately, our work strives to bridge the gap between the quest for equitable decision-making and the practical limitations of resource-constrained deployment scenarios. We keep the exploration of composition of biases across multiple axes as a future work. We also want to explore debiasing using PEFTs in generation based techniques.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. 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. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

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, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019a. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019b. Parameter-efficient transfer learning for nlp.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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, Online. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, and Elaine Gonzalez. 2018. Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

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.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models.

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

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

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

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.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *ArXiv*, abs/2010.06032.

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

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

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

Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Bias Axes & Attribute Words

We describe the bias axes and attribute words that we will use in our studies. We mention three different biases currently but we will be incorporating other biases as well. Hereby, we present a list of some attribute word examples as well along with the biases.

Gender (actor, actress), (boy, girl), (brother, sister), (he, she)

Group (black, caucasian, asian), (african, caucasian, asian), (black, white, asian) (jewish, christian, muslim), (judaism, christianity, islam), (gay, lesbian, straight)

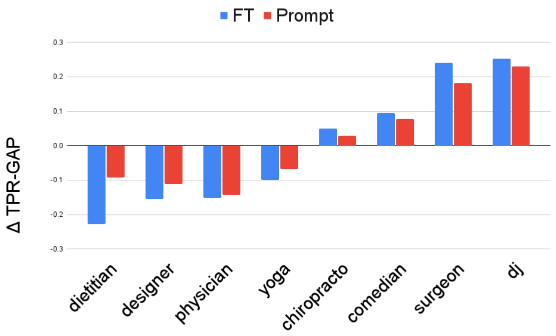
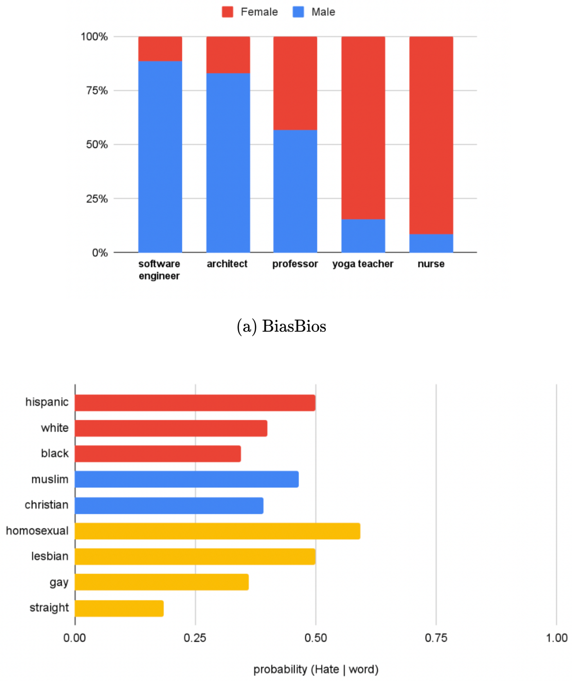


Figure 4: Comparing the TPR-GAP performance of CDA debiasing using FT and Prompt on the Biasbios dataset across different occupations.



(a) BiasBios

(b) GHC

Figure 5: Label distributions in datasets BiasBios (a) and GHC (b). BiasBios dataset shows high label correlations between gender and professions. We see that probability of a statement being hateful has a high correlation with the mention of identity words.

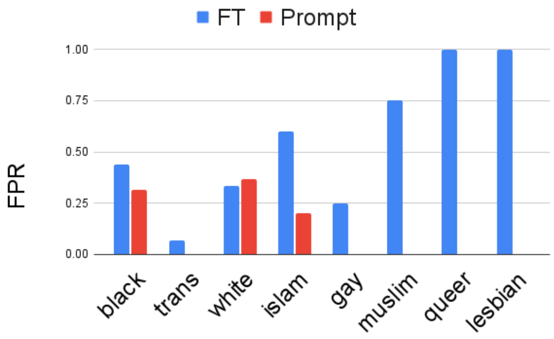


Figure 6: Comparing the False Positive Rates (FPR) performance of CDA debiasing using FT and Prompt on the GHC dataset across different group identifiers.

Text	Label	BERT	Ours
Dr. Almans Khan practices .... She completed MBBS from D Y Patil Dental College , Navi Mumbai (New Mumbai) in 2007 and DNB - General **Surgery** from National Board Of Examination in 2012 .	surgeon	physician	surgeon
Earlier , he had been working as an **interior designer** with Anna Properties at Kochi , Kerala and as a 3D Visualizer and Architectural Draftsman at Kabra Associates .	Interior Designer	Architect	Interior Designer
i wonder if one of these guys is a jew	Non toxic	Non toxic	Toxic
a holocaust	Non toxic	Non toxic	Toxic

Table 5: Qualitative instances where Prompt Tuning technique coupled with CDA debiasing demonstrates superior performance compared to FT.