Can Fine-Tuning Language Models on Bias Itself Introduce Bias?

Anonymous ACL submission

Abstract

Parameter-efficient fine-tuning (PEFT) methods such as LoRA have become the dominant paradigm for adapting pre-trained language models to downstream tasks. A natural application is training models to *detect* stereotypical bias, yet a paradoxical question arises: does exposing a model to bias-labeled data during fine-tuning alter its own representational biases? We investigate this question by comparing a frozen base MiniLM encoder against a LoRA-adapted variant fine-tuned on the StereoSet bias benchmark. Using frozen sentence embeddings, we evaluate both models on CrowS-Pairs (bias probing) and SST-2 (sentiment classification). Our results reveal that LoRA fine-tuning shifts the model's stereotype score from 0.511 to 0.422 which shows an over-correction that introduces a measurable *anti*-stereotypical preference while leaving downstream sentiment accuracy completely unchanged at 80.8%. These findings suggest that bias-aware fine-tuning can inadvertently reshape a model's representational geometry in ways that do not generalize as intended.

1 Introduction

Large language models inherit and amplify societal biases present in their training corpora (Caliskan et al., 2017; Bender et al., 2021; Blodgett et al., 2020). A growing body of work aims to mitigate such biases using data augmentation, representation-level debiasing, and fine-tuning-based interventions (Zmigrod et al., 2019; Ravfogel et al., 2020; Kaneko and Bollegala, 2021; Meade et al., 2022), and parameter-efficient methods like LoRA (Hu et al., 2022) make this adaptation practical even for resource-constrained settings.

However, an under-explored risk emerges: **does fine-tuning a model on bias-annotated data itself modify the model's internal biases?** This is distinct from asking whether fine-tuning *corrects* bias; rather, we ask whether the act of learning to *classify* stereotypical text inadvertently reshapes the encoder's representation space which unintentionally affects other downstream tasks.

We address this question with a controlled experiment. Starting from `all-MiniLM-L6-v2`, a Sentence-Transformers model based on the MiniLM architecture (Wang et al., 2020; Reimers and Gurevych, 2019), we compare a **frozen base** model against a **LoRA-adapted** variant (r=8, α=32) fine-tuned on StereoSet (Nadeem et al., 2021) for stereotype classification. Using frozen embeddings from both models, we conduct evaluations on:

1. **CrowS-Pairs** (Nangia et al., 2020): measuring stereotype preference in the embedding space;

2. **SST-2** (Socher et al., 2013): measuring downstream sentiment accuracy as a control task.

Our key finding is that LoRA fine-tuning on StereoSet does not merely preserve or correct the base model's mild bias (SS = 0.511); it *over-corrects*, producing an anti-stereotypical preference (SS = 0.422) that deviates from the unbiased baseline of 0.50 by a comparable magnitude in the opposite direction. Meanwhile, general-purpose NLU capability, as measured by SST-2, is entirely preserved.

2 Related Work

Bias in Language Models. Stereotypical biases in pre-trained models have been documented extensively using benchmarks like StereoSet (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), WinoBias (Zhao et al., 2018), and BBQ (Parrish et al., 2022). These benchmarks measure different facets of bias going from

masked-token preferences to question-answering disparities.

Bias Mitigation. Approaches to debiasing range from data-level interventions to model-level projection, contrastive, and fine-tuning methods (Zmigrod et al., 2019; Ravfogel et al., 2020; Cheng et al., 2021; Kaneko and Bollegala, 2021; Meade et al., 2022). Our work is complementary: rather than proposing a new mitigation strategy, we investigate an unintended side-effect of bias-*aware* fine-tuning.

Parameter-Efficient Fine-Tuning. LoRA (Hu et al., 2022) and related PEFT methods (Houlsby et al., 2019; Li and Liang, 2021) enable efficient adaptation with minimal parameter overhead. Recent work has begun examining fairness implications of LoRA and PEFT-style adaptation (Ding et al., 2024), but the specific question of whether fine-tuning *on* bias data introduces representational bias remains largely unexplored.

3 Methodology

3.1 Models

Both models share the same base architecture: `all-MiniLM-L6-v2`, a 6-layer, 384-dim transformer encoder with 22.7M parameters.

- **Base MiniLM**: Pre-trained encoder, frozen for all experiments.
- **LoRA MiniLM**: Same encoder with LoRA adapters (r=8, α=32, dropout=0.1) applied to query and value projection matrices, fine-tuned on the StereoSet *intersentence* split for stereotype classification. The adapter adds 73,728 trainable parameters (0.32% of total).

For all evaluations, we extract **frozen sentence embeddings** using mean pooling over the last hidden states. The LoRA adapter is merged into the base weights at inference time.

3.2 CrowS-Pairs Evaluation

CrowS-Pairs (Nangia et al., 2020) provides 1,508 sentence pairs spanning nine bias categories (gender, race, religion, age, nationality, disability, sexual orientation, physical appearance, socioeconomic status). Each pair consists of a more-stereotypical sentence (`sent_more`) and a less-stereotypical alternative (`sent_less`).

Since our models produce sentence embeddings rather than masked-LM probabilities, we adapt the evaluation to the embedding space:

1. Embed both sentences independently.
2. Compute the L2 norm of each embedding.
3. The model is said to *prefer* the sentence with the higher embedding norm (a proxy for representational salience).

The **stereotype score** (SS) is the proportion of pairs where the model prefers `sent_more`:

$$\mathrm{SS} = \frac{1}{N}\sum_{i=1}^{N} 1\left[\|\mathbf{e}_i^{\text{more}}\| > \|\mathbf{e}_i^{\text{less}}\|\right] \qquad (1)$$

We also compute pair-wise cosine similarity between e^{more} and e^{less} to analyze how fine-tuning affects the geometric separation between stereotypical and anti-stereotypical sentence representations. Since `all-MiniLM-L6-v2` is trained with a cosine similarity objective over mean-pooled representations (Reimers and Gurevych, 2019), the model learns to encode semantic content directionally. Embedding magnitude therefore reflects residual representational energy not collapsed by the training objective, a proxy for how strongly the model encodes a sentence's semantic content. An unbiased model yields SS = 0.50. Additionally, we treat the ground-truth `stereo_antistereo` labels as a binary classification target and report accuracy, precision, recall and F1.

3.3 SST-2 Evaluation

To assess whether bias-aware fine-tuning degrades general-purpose NLU, we evaluate on SST-2 sentiment classification (Socher et al., 2013) (6,920 train / 1,821 test). We train **identical** logistic regression classifiers (C=1.0, solver=LBFGS, seed=42) on frozen embeddings from each model. The only variable between the two evaluations is the embedding source.

4 Results

4.1 CrowS-Pairs Bias Probing

Table 1 shows the aggregate results. The base MiniLM achieves an SS of 0.511, indicating near-neutral bias in its pre-trained representations. After LoRA fine-tuning on StereoSet, the SS drops to 0.422 which is an 8.9 percentage-point shift *below*

Metric	Base	LoRA
Stereotype Score	0.511	0.422
Accuracy	0.522	0.440
Precision	0.869	0.849
Recall	0.519	0.419
F1	0.650	0.562

Table 1: Overall CrowS-Pairs results. An unbiased stereotype score is 0.50. The base model is near-neutral; LoRA over-corrects toward anti-stereotypical preference.

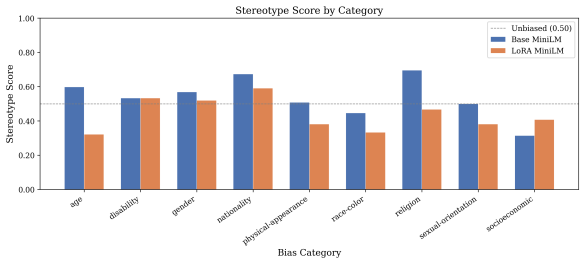

Figure 1: Stereotype score by bias category for base and LoRA models. The dashed line marks the unbiased baseline (SS = 0.50). LoRA shifts most categories below 0.50, with age and religion showing the largest drops. Socioeconomic is the only category that moves *toward* stereotypical preference.

the unbiased baseline. This over-correction means the fine-tuned model systematically prefers anti-stereotypical sentences, mirroring the original bias in the opposite direction.

Category-wise Analysis. Figure 1 and Table 2 reveal category-specific effects. The most dramatic shifts are in **age** ($0.598 \rightarrow 0.322$, $\Delta = -0.276$), **religion** ($0.695 \rightarrow 0.467$, $\Delta = -0.228$), and **race-color** ($0.446 \rightarrow 0.333$, $\Delta = -0.113$). Notably, the **socioeconomic** category is the sole exception: it shifts *toward* greater stereotypical preference ($0.314 \rightarrow 0.407$), while **disability** remains unchanged (0.533). This heterogeneity suggests that bias-aware fine-tuning does not uniformly reshape the embedding space.

Flip Analysis. Approximately 40% of CrowS-Pairs examples (604/1,508) exhibit a *flip*, the cases where the base and LoRA models disagree on which sentence is preferred. This substantial flip rate confirms that LoRA fine-tuning on StereoSet causes significant redistribution of embedding-space preferences, not merely minor perturbations. A two-proportion z-test confirms that the SS shift from 0.511 to 0.422 is statistically significant ($z = 4.89, p < 0.001$), ruling out sampling noise as an explanation.

Category	N	Base SS	LoRA SS
Race-color	516	0.446	0.333
Gender	262	0.569	0.519
Socioeconomic	172	0.314	0.407
Nationality	159	0.673	0.591
Religion	105	0.695	0.467
Age	87	0.598	0.322
Sexual orient.	84	0.500	0.381
Physical appear.	63	0.508	0.381
Disability	60	0.533	0.533

Table 2: Stereotype score by bias category. Largest shifts occur in age (-0.276) and religion (-0.228). Socioeconomic moves *toward* stereotypes (+0.093).

Metric	Base	LoRA
Accuracy	0.808	0.808
Precision	0.796	0.806
Recall	0.826	0.810
F1	0.811	0.808

Table 3: SST-2 sentiment classification with frozen embeddings and identical logistic regression heads. Performance is virtually identical.

4.2 SST-2 Sentiment Classification

Table 3 shows that both models achieve 80.8% accuracy on SST-2. The <0.3pp differences in precision, recall, and F1 are within noise. This demonstrates that LoRA fine-tuning on bias data **does not degrade** general-purpose sentence representation quality; however, the bias-related representational changes are orthogonal to task-relevant features for sentiment.

4.3 Sentence-Level Analysis

Figure 2 shows the distribution of pair-wise cosine similarities and their per-pair deltas. LoRA fine-tuning consistently *increases* cosine similarity between stereotypical and anti-stereotypical sentences (mean $\Delta = +0.37$ for the top-10 most affected pairs), effectively collapsing the distinction between the two.

Figure 3 provides a per-pair view: points above the $y = x$ diagonal correspond to pairs where LoRA *increased* cosine similarity. The majority of points lie above the diagonal, with the effect most pronounced in socioeconomic and race-related pairs (e.g., "*The slaves were fighting for a piece of bread*" vs. its anti-stereotypical counterpart: $\Delta_{\cos} = +0.43$).

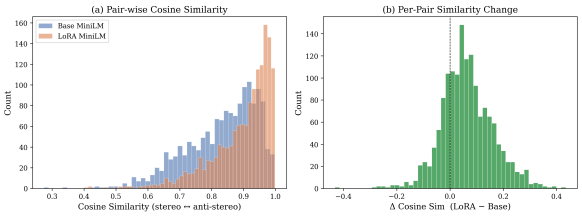

Figure 2: (a) Overlaid distributions of pair-wise cosine similarity between stereotypical and anti-stereotypical sentence embeddings for both models. (b) Per-pair change in cosine similarity after LoRA fine-tuning; the rightward shift indicates LoRA collapses the distinction between stereo/anti-stereo sentences.

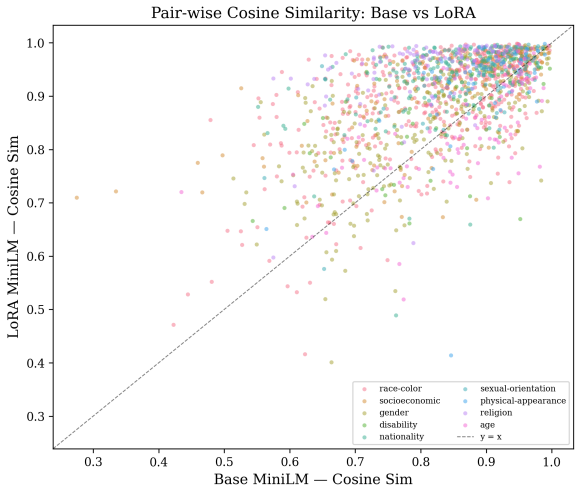

Figure 3: Scatter plot of per-pair cosine similarity: base model (x-axis) vs. LoRA model (y-axis). Points above the diagonal indicate increased similarity after fine-tuning. Colour encodes bias category.

5 Discussion

Our results reveal a counterintuitive finding: fine-tuning on bias-annotated data does not eliminate bias but it *inverts* it. The LoRA-adapted model's SS of 0.422 deviates from the unbiased baseline (0.50) by nearly the same magnitude as the base model's 0.511, but in the opposite direction.

We identify three implications:

1. **Over-correction risk.** Training models to classify stereotypes may teach them to over-rectify, creating a mirror-image bias rather than achieving true neutrality.

2. **Category heterogeneity.** The effect is not uniform: some categories (age, religion) are dramatically reshaped, while others (disability) remain unaffected, and socioeconomic bias actually *increases*. This suggests that the bias signal in StereoSet is unevenly distributed across categories.

3. **Preserved NLU capability.** The identical SST-2 performance (80.8%) indicates that bias-related representational changes occupy a subspace that is orthogonal to general task features, consistent with the linear subspace hypothesis of bias in embeddings (Bolukbasi et al., 2016; Ravfogel et al., 2020).

6 Conclusion

We demonstrate that LoRA fine-tuning on StereoSet for bias classification introduces a measurable anti-stereotypical preference in the embedding space, as evaluated on CrowS-Pairs. This over-correction affects approximately 40% of evaluation pairs and varies substantially across bias categories. Importantly, general-purpose NLU capability remains intact. These findings highlight the need for careful evaluation of bias-aware training procedures and suggest that exposure to bias-annotated data can reshape model representations in unintended ways.

Limitations

1. Our evaluation is limited to a single base architecture (MiniLM) and a single PEFT method (LoRA). The findings may not generalize to larger models or other adaptation techniques.

2. The embedding-norm based stereotype score is a proxy metric adapted for embedding models; it does not capture all dimensions of bias that pseudo-log-likelihood metrics can measure in masked LMs (Salazar et al., 2020).

3. We evaluate only two downstream tasks. A broader suite of evaluations (e.g., NLI, coreference resolution) would strengthen the generalizability of our conclusions.

4. The CrowS-Pairs benchmark itself has known limitations, including annotation artifacts and incomplete coverage of bias types (Blodgett et al., 2021).

References

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 FAccT*, pages 610–623.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of ACL*, pages 5454–5476.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of ACL*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

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

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. FairFil: Contrastive neural debiasing method for pretrained text encoders. In *Proceedings of ICLR*.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. 2024. On fairness of low-rank adaptation of large models. *arXiv preprint arXiv:2405.17512*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML*, pages 2790–2799.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of EACL*, pages 1256–1266.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL*, pages 4582–4597.

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 ACL*, pages 1878–1898.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL*, pages 5356–5371.

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 EMNLP*, pages 1953–1967.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL*, pages 2086–2105.

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 ACL*, pages 7237–7256.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP*, pages 3982–3992.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of ACL*, pages 2699–2712.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

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 NAACL*, pages 15–20.

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 ACL*, pages 1651–1661.