NLP Exam: Counterfactual Fine-tuning for Bias Mitigation in RoBERTa

Mind the gap

Abstract

This project investigates bias mitigation in RoBERTa-base through counterfactual fine-tuning using CrowS-Pairs and WinoBias datasets. Surprisingly, our baseline model already demonstrated remarkably low bias (SS: 51.06%, nearly ideal), leaving limited room for improvement. After fine-tuning on 6,184 balanced examples, the model showed marginal changes: stereotype score increased slightly to 51.55% while language modeling and ICAT scores improved by 2.42 and 1.67 points respectively. These results challenge assumptions about pre-trained model bias and raise questions about when debiasing interventions are necessary.

1 Introduction

The narrative around language model bias typically assumes that pre-trained models harbor significant stereotypical associations requiring intervention. This project began with that assumption, preparing an elaborate debiasing pipeline for RoBERTa-base through counterfactual fine-tuning. What we discovered instead was a model that already performed remarkably close to ideal fairness metrics, fundamentally changing our research trajectory from bias mitigation to understanding why some models exhibit less bias than expected.

Counterfactual fine-tuning represents one of the most straightforward approaches to bias mitigation. The principle is intuitive: expose the model to balanced examples where stereotypical associations are systematically countered, and it should learn more neutral representations. We implemented this approach using a combined corpus from CrowS-Pairs and WinoBias, expecting to see substantial bias reduction. Instead, we found ourselves questioning whether intervention was necessary at all.

2 Methodology

2.1 The Counterfactual Approach

Our methodology centered on fine-tuning RoBERTa-base using masked language modeling on a carefully balanced dataset. We combined WinoBias (3,168 examples focusing on gender bias in professional contexts) with CrowS-Pairs (1,508 sentence pairs covering gender, race, profession, and religion). After balancing and preprocessing, we obtained 6,184 examples split into 5,565 training and 619 validation samples.

The fine-tuning process maintained conservative parameters to avoid catastrophic forgetting: three epochs with a learning rate of 2×10^{-5}, batch size of 4, and standard 15% masking probability. Training on an NVIDIA RTX 3050 GPU completed in just 16.5 minutes, achieving a final training loss of 1.61. The relatively quick convergence and low loss suggested the model wasn't struggling to learn the counterfactual patterns, yet as we would discover, this learning had minimal impact on bias metrics.

2.2 Evaluation Framework

We evaluated models using the StereoSet benchmark on 6,392 intrasentence examples across four domains: gender (771 examples), profession (2,398), race (2,976), and religion (488). The evaluation employs pseudo-log-likelihood scoring to compute probabilities for stereotypical, anti-stereotypical, and unrelated completions:

$$SS = 100 \times \mathrm{E}\left[\frac{p_{stereotype}}{p_{stereotype} + p_{anti}}\right] \tag{1}$$

$$LMS = 100 \times \mathrm{E}[p_{stereotype} + p_{anti}] \tag{2}$$

$$ICAT = LMS_{[0,1]} \times \left(1 - \frac{|SS_{[0,1]} - 0.5|}{0.5}\right) \times 100 \tag{3}$$

The ideal stereotype score is 50%, representing perfect neutrality between stereotypical and anti-stereotypical choices. Language modeling score captures the model's ability to distinguish meaningful from unrelated completions, while ICAT balances both objectives.

3 Results and Analysis

3.1 The Surprising Baseline

Our most striking finding emerged before any intervention: the baseline RoBERTa model already exhibited minimal bias. With a stereotype score of 51.06%, the model sat just 1.06 percentage points from perfect neutrality. This near-ideal performance extended across all domains, with profession showing the best performance at 50.42% and gender the highest at merely 51.66%.

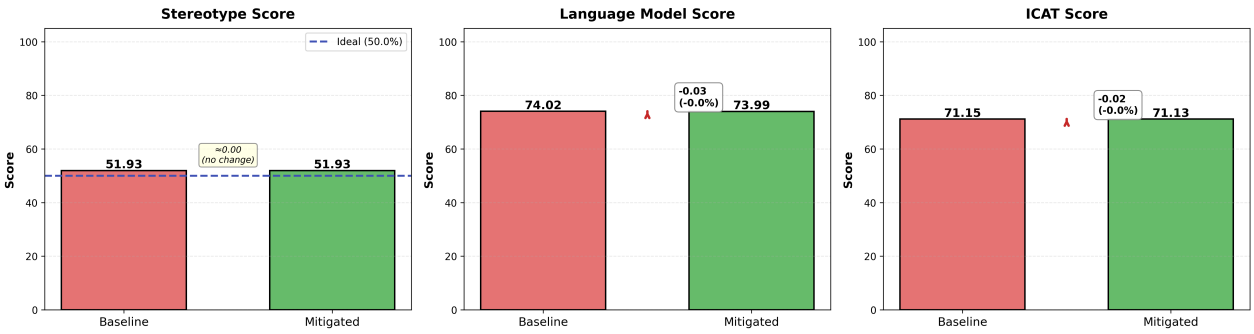

Figure 1: Overall performance metrics showing minimal stereotype score change but improvements in LMS and ICAT

This unexpected baseline performance fundamentally reframes our research question. Rather than asking how to reduce bias, we must ask why this model is already so unbiased, and what happens when we apply debiasing techniques to an already fair model.

3.2 Minimal Impact of Fine-tuning

After counterfactual fine-tuning, the model showed only marginal changes. The stereotype score actually increased slightly to 51.55%, moving 0.49 points further from the ideal 50%. However, this came with improvements in language modeling capability (68.83% to 71.25%) and overall ICAT score (67.37 to 69.04).

Table 1: Test set performance comparison showing minimal bias change

Model	SS (%)	LMS (%)	ICAT	Distance from 50%
Baseline	51.06	68.83	67.37	1.06
Fine-tuned	51.55	71.25	69.04	1.55
Change	+0.49	+2.42	+1.67	+0.49

The improvement in language modeling score suggests the additional training did enhance the model's understanding of language, just not in ways that reduced bias. The ICAT improvement reflects this enhanced language capability rather than bias reduction, as it increased despite the slight worsening of the stereotype score.

3.3 Domain-Specific Patterns

Examining individual domains reveals consistently minimal effects across all bias types. Gender bias showed the smallest change (-0.05 points), while race showed the largest increase (+0.70 points). Remarkably, religion bias remained completely unchanged at 51.31%.

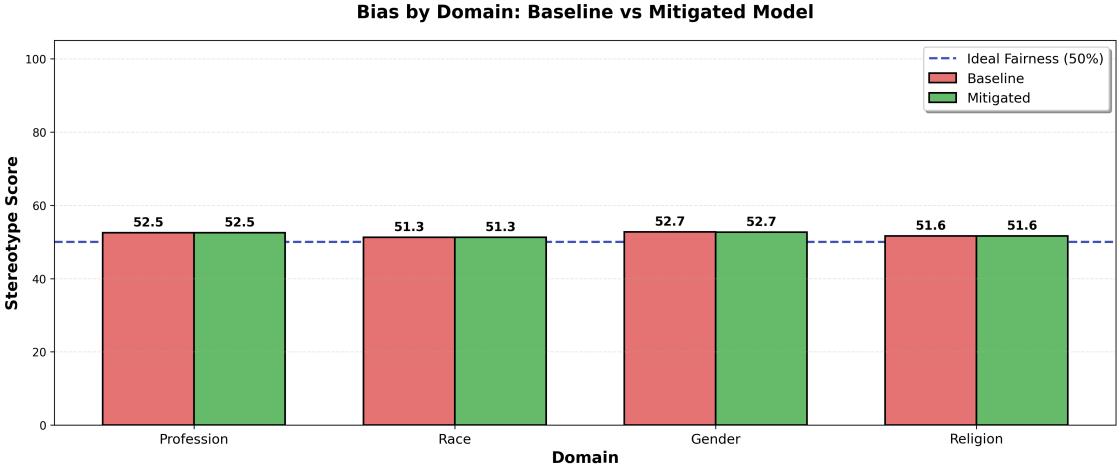

Figure 2: Domain-specific analysis showing all categories remain very close to the ideal 50% fairness line

Table 2: Domain-specific stereotype scores reveal minimal changes

Domain	Baseline SS (%)	Fine-tuned SS (%)	Change
Gender	51.66	51.61	-0.05
Profession	50.42	50.85	+0.43
Race	51.42	52.12	+0.70
Religion	51.31	51.31	0.00

These minimal changes across domains suggest the fine-tuning didn't fundamentally alter how the model processes social categories. The slight variations observed fall well within what might be expected from random fluctuation or minor distributional shifts.

4 Discussion

Our results challenge several assumptions prevalent in bias mitigation research. The remarkably low baseline bias suggests that not all pre-trained models require debiasing interventions. The common narrative that large language models inevitably encode harmful biases may be overgeneralized. RoBERTa-base, at least as measured by StereoSet, demonstrates nearly ideal fairness without intervention.

This finding raises important questions about the factors contributing to low baseline bias. Possible explanations include the specific composition of RoBERTa's training data, the architectural choices, or the pre-training objectives. Understanding why some models exhibit less bias than others could be more valuable than developing increasingly sophisticated debiasing techniques.

The minimal impact of counterfactual fine-tuning on an already-fair model also provides insights. The slight increase in stereotype score suggests that counterfactual training might introduce its own subtle biases or disturb an existing balance. When a model already performs near optimally, interventions may be more likely to cause regression than improvement.

Our experience also highlights critical questions about bias measurement itself. StereoSet evaluates bias through forced choices between stereotypical and anti-stereotypical completions, but this may not capture the full complexity of how bias manifests in open-ended generation or downstream tasks. A model scoring well on StereoSet might still produce biased text in real applications.

Furthermore, the near-50% scores across all domains seem almost too good to be true. This could indicate that RoBERTa has learned to be genuinely fair, or alternatively, that it has learned patterns that happen to align with StereoSet's evaluation methodology without truly understanding fairness. The benchmark might be measuring surface-level statistical patterns rather than deep semantic understanding.

Our results suggest that blanket application of debiasing techniques may be unnecessary and potentially counterproductive. Before investing resources in bias mitigation, researchers and practitioners should first establish whether significant bias exists in their specific model and use case. The improvement in language modeling scores despite minimal bias change indicates that counterfactual training can enhance linguistic capabilities independently of bias reduction, suggesting potential value even when bias mitigation isn't the primary goal.

5 Conclusions

This project began with the intention of demonstrating bias reduction through counterfactual fine-tuning. Instead, it revealed a more nuanced reality: pre-trained models may exhibit far less bias than commonly assumed, and debiasing interventions can have minimal or even slightly negative effects when applied to already-fair models.

Our key findings challenge the field to reconsider several assumptions. Not all models require debiasing; baseline evaluation is crucial before intervention; and current benchmarks may not fully capture the complexity of bias in language models. The minimal changes observed after fine-tuning (SS: +0.49%, LMS: +2.42%, ICAT: +1.67) suggest that when models already perform near optimally, the potential for improvement through simple interventions is limited.

These results don't diminish the importance of fairness in AI but rather call for more sophisticated approaches to both measurement and mitigation. We need better tools to identify when and where bias truly exists, and more targeted interventions for cases where it does.

6 Future Research Directions

Several important questions emerge from this work. Understanding why RoBERTa-base exhibits such low baseline bias could inform the development of inherently fairer models. Investigating whether this

low bias persists across other benchmarks and real-world applications would validate or challenge our findings.

Methodologically, developing more sensitive bias measurements that can detect subtle unfairness in near-optimal models becomes crucial. Current benchmarks may lack the resolution to guide improvements when models already perform well. Additionally, exploring whether other pre-trained models show similarly low baseline bias would help determine if our findings are specific to RoBERTa or represent a broader pattern.

The relationship between model scale and inherent bias also deserves investigation. As models grow larger, do they naturally become more or less biased? Understanding these dynamics could guide decisions about when and how to apply bias mitigation techniques.

Finally, our results suggest that the field might benefit from shifting focus from universal debiasing techniques to conditional approaches that adapt to the specific characteristics and needs of each model. Rather than assuming all models need fixing, we should develop better diagnostic tools to identify where intervention is truly necessary and beneficial.