

Fairness-Aware Fine-Tuning of GPT-2 for Paraphrase Detection

Stanford CS224N Default Project

Deonna Owens

Department of Computer Science
Stanford University
deonnao@stanford.edu

Abstract

Pretrained language models achieve strong performance across many natural language processing tasks, yet their sensitivity to demographic cues remains a significant concern. In this work, we investigate whether fairness-aware training techniques can reduce identity-based prediction instability in GPT-2 while preserving task performance. We study this question in the context of paraphrase detection on the Quora Question Pairs dataset. Starting from a baseline GPT-2 model that achieves 89.55% development accuracy, we evaluate two fairness interventions: counterfactual data augmentation (CDA) and a consistency-based fairness regularization objective that penalizes prediction divergence between original and identity-swapped inputs. CDA substantially improves subgroup parity, reducing the subgroup accuracy gap from 3.65% to 2.90% while also lowering prediction flip rate. The fairness-regularized model achieves the strongest robustness to identity substitutions, reducing flip rate from 4.56% to 2.80% while maintaining baseline accuracy. These results demonstrate that fairness-aware training can meaningfully improve robustness to demographic perturbations without sacrificing predictive performance, while different interventions optimize distinct notions of fairness.

1 Key Information to include

- Mentor: Ahmed Ahmed

2 Introduction

Pretrained language models have become the dominant paradigm in natural language processing, achieving strong performance across a wide range of tasks including classification, question answering, and text generation. Models such as GPT-2 learn rich contextual representations from large-scale unlabeled corpora and can be adapted to downstream tasks through fine-tuning [1]. The availability of widely adopted libraries such as the HuggingFace Transformers framework has further enabled the rapid deployment and experimentation with these models in practical applications [2].

Despite their strong empirical performance, pretrained language models often exhibit sensitivity to demographic cues present in training data. In particular, models may associate identity-related tokens with task labels, leading to inconsistent predictions when inputs differ only in demographic identifiers such as gendered names or pronouns. Consequently, semantically equivalent inputs that vary only in demographic attributes can produce different model outputs, raising concerns about the reliability and fairness of NLP systems deployed in real-world applications.

Prior work has explored several approaches for mitigating demographic bias in language models. These include data-level interventions such as counterfactual data augmentation, as well as model-level approaches such as reweighting or adversarial training objectives that aim to reduce the influence of demographic information in learned representations. Counterfactual data augmentation (CDA), in

particular, generates identity-swapped variants of training examples to reduce reliance on demographic signals [3]. By exposing the model to semantically equivalent inputs that differ only in demographic attributes, CDA encourages the model to base its predictions on task-relevant information rather than identity-specific cues. However, data augmentation alone does not explicitly enforce prediction consistency when identity substitutions occur at inference time, and models may still produce divergent predictions for counterfactual inputs.

In this work, we study whether fairness-aware training techniques can improve the robustness of GPT-2 to identity-based perturbations in the context of paraphrase detection. We fine-tune GPT-2 on the Quora Question Pairs dataset and evaluate two fairness interventions: counterfactual data augmentation and a proposed consistency-based fairness regularization objective. The regularization objective introduces an additional training signal that penalizes differences between the model’s predictions on original inputs and their identity-swapped counterparts, encouraging stable predictions when demographic identifiers are altered. Our experiments analyze the trade-offs between predictive performance and demographic robustness, measuring both standard classification accuracy and fairness metrics that quantify prediction instability under identity substitutions.

3 Related Work

3.1 Pretrained Language Models

Large pretrained language models have become a standard foundation for downstream NLP tasks. GPT-2 demonstrated that decoder-only transformers trained on large-scale web corpora can learn transferable linguistic representations that support strong downstream adaptation through fine-tuning [1]. In our setting, GPT-2 serves as the base model through which we study the interaction between fine-tuning and fairness-aware objectives.

3.2 Bias in NLP Systems

A substantial body of work has shown that NLP models can encode and amplify social biases present in training data. Early work on static word embeddings demonstrated that gender stereotypes are often represented geometrically in embedding space, motivating methods that explicitly remove bias directions from learned representations [4]. Subsequent work showed that such bias also affects downstream structured prediction tasks. For example, Zhao et al. introduced WinoBias and showed that coreference systems systematically favor stereotypical associations, while Dixon et al. studied unintended bias in text classification and emphasized the importance of evaluating performance across identity terms and demographic subgroups [5, 6]. Together, these studies suggest that fairness concerns arise not only at the representation level, but also in end-task behavior.

3.3 Bias Mitigation Methods

Prior mitigation strategies in NLP can be broadly grouped into representation-level, data-level, and objective-level approaches. Representation-level methods attempt to remove protected-attribute information from embeddings or hidden states [4]. Data-level methods instead modify the training distribution, often through counterfactual augmentation. Maudslay et al. proposed name-based counterfactual data substitution to reduce gender bias by generating identity-swapped examples while preserving semantic content [3]. Relatedly, Zhao et al. showed that data augmentation can reduce gender bias in coreference systems when paired with debiasing techniques [5]. Objective-level approaches incorporate fairness directly into optimization; for instance, Zhang et al. use adversarial learning to discourage predictive models from encoding protected attributes [7]. Our work is most closely related to the latter two lines of research. We study whether combining counterfactual data augmentation with a prediction-consistency regularization objective can reduce identity-based prediction instability during fine-tuning. This combination is motivated by the limitation that augmentation alone does not explicitly enforce consistent predictions at inference time for identity-swapped inputs.

4 Approach

Our goal is to study whether fairness-aware training techniques can improve the robustness of pretrained language models to demographic perturbations while preserving task performance. Our primary experiments focus on paraphrase detection using GPT-2, where we evaluate two fairness interventions: counterfactual data augmentation and a consistency-based fairness regularization objective. These methods aim to reduce prediction instability when demographic identifiers are altered while maintaining strong paraphrase classification performance.

4.1 Model Architecture

All experiments use the GPT-2 base (124M parameters) architecture [1]. GPT-2 is a decoder-only transformer consisting of stacked self-attention layers followed by position-wise feed-forward networks. The configuration used in this work contains 12 transformer layers with hidden dimension 768 and 12 attention heads. The model operates on byte-pair encoded tokens drawn from a vocabulary of size 50,257 and supports sequences up to length 1,024.

The transformer follows the pre-normalization architecture in which layer normalization precedes both the attention and feed-forward blocks. Dropout with rate 0.1 is applied to both hidden representations and attention weights, and the model uses GELU nonlinearities. Input token embeddings are tied with the output projection matrix. The same architecture is used across all experiments.

4.2 Paraphrase Detection

We fine-tune GPT-2 on the Quora Question Pairs dataset. Each input pair (s_1, s_2) is converted into a prompt-style representation:

Question 1: "s1"
Question 2: "s2"
Are these questions asking the same thing?

Let $f_\theta(x)$ denote the GPT-2 model applied to an input sequence x . We define $h_{final} = f_\theta(x)$ as the final hidden representation corresponding to the last token in the prompt. A linear classification head is applied to this representation:

$$y = Wh_{final} + b,$$

where $W \in \mathbb{R}^{768 \times 2}$ produces logits for the binary paraphrase labels. The model is trained using cross-entropy loss.

4.3 Baseline Model

Our baseline model consists of GPT-2 fine-tuned for paraphrase detection using the task formulation described above. The model receives the prompt-formatted question pair and predicts the paraphrase label using a linear classification head applied to the final hidden representation h_{final} . Training is performed using standard cross-entropy loss without counterfactual data augmentation or fairness regularization. This baseline represents a conventional fine-tuning setup and serves as the reference point for evaluating the impact of fairness-aware training interventions.

4.4 Counterfactual Data Augmentation

To reduce reliance on demographic cues during training, we apply counterfactual data augmentation (CDA) following the approach of [3]. CDA generates semantically equivalent training examples by replacing identity-related tokens with counterfactual alternatives while preserving the semantic content of the original input.

Our augmentation procedure includes three categories of substitutions summarized in Table 1. During training, identity substitutions are applied with probability 0.5 to generate counterfactual examples.

Substitution Type	Count	Examples
Gendered name pairs	60	<i>James</i> $\leftrightarrow$ <i>Mary</i>
Ethnicity-associated name pairs	20	<i>Connor</i> $\leftrightarrow$ <i>Jamal</i>
Pronoun / gendered-term swaps	22	<i>he</i> $\leftrightarrow$ <i>she</i> , <i>king</i> $\leftrightarrow$ <i>queen</i> , <i>mr</i> $\leftrightarrow$ <i>mrs</i>

Table 1: Identity substitutions used for counterfactual data augmentation.

4.5 Consistency-Based Fairness Regularization

While CDA exposes the model to identity-swapped examples, it does not explicitly enforce consistent predictions between original and counterfactual inputs. To encourage prediction stability, we introduce a fairness regularization objective that penalizes divergence between predictions on the original example and its counterfactual example.

Let p_{orig} denote the predicted probability distribution for the original example and p_{cf} the prediction for its counterfactual example. We define the fairness loss as the symmetric KL divergence

$$L_{fairness} = \frac{1}{2} (KL(p_{cf} \parallel p_{orig}) + KL(p_{orig} \parallel p_{cf})).$$

The final training objective is

$$L_{total} = L_{task} + \lambda L_{fairness},$$

where L_{task} is the standard classification loss and λ controls the strength of the fairness constraint. In our experiments we set $\lambda = 0.5$, which provided a stable balance between task performance and fairness regularization in preliminary validation experiments.

5 Experiments

We evaluate GPT-2 across three tasks: paraphrase detection, sentiment classification, and Shakespearean sonnet generation. Our primary experiments study fairness-aware training for paraphrase detection, while the remaining tasks provide additional evaluation of the model’s general fine-tuning behavior within the CS224N project framework.

5.1 Datasets

Quora Question Pairs (QQP). We evaluate paraphrase detection using the Quora Question Pairs (QQP) dataset.¹ The dataset contains pairs of questions annotated according to whether they are semantically equivalent. The training set contains 283,011 pairs, the development set contains 40,430 pairs, and the test set contains 80,860 pairs with hidden labels used for leaderboard evaluation.

Stanford Sentiment Treebank (SST). Sentiment classification experiments use the Stanford Sentiment Treebank [8], which contains movie review sentences annotated with fine-grained sentiment labels across five classes.

Counterfactual IMDB (CFIMDB). We additionally evaluate sentiment classification using the Counterfactual IMDB dataset [9], which introduces counterfactual identity substitutions to test robustness to demographic cues.

Shakespeare Sonnets. For generative modeling we fine-tune GPT-2 on a corpus of Shakespearean sonnets provided as part of the CS224N project framework.

5.2 Evaluation Metrics

For classification tasks we report accuracy. For paraphrase detection we additionally evaluate two fairness metrics designed to measure model sensitivity to demographic perturbations.

¹<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

Subgroup Accuracy Gap. Let $\text{Acc}(g)$ denote classification accuracy for subgroup g . The subgroup accuracy gap is defined as

$$\max_{i,j} |\text{Acc}(g_i) - \text{Acc}(g_j)|.$$

Subgroups correspond to identity categories including *male*, *female*, *neutral*, and *name-only*. Only subgroups containing at least ten examples are considered.

Prediction Flip Rate. Prediction flip rate measures the fraction of examples whose predicted label changes when identity tokens are deterministically swapped to produce a counterfactual example. Lower flip rates indicate greater prediction stability under demographic perturbations.

For sonnet generation we evaluate text quality using the CHRF metric, which measures character-level n -gram overlap between generated sequences and reference text.

5.3 Training Configuration

All paraphrase detection experiments use the GPT-2 base model (124M parameters). Models are optimized using AdamW [10] with coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$. Training is performed using mixed-precision automatic mixed precision (AMP) on GPU with random seed 11711. Across all runs we use a batch size of 16 with four gradient accumulation steps, producing an effective batch size of 64. Models are trained for 10 epochs with learning rate 10^{-5} . Training each paraphrase detection model required approximately 2–3 hours on a single NVIDIA GPU using mixed-precision training.

Three training modes are evaluated. The baseline model corresponds to a previously trained GPT-2 paraphrase detection checkpoint that was trained for 10 epochs using standard cross-entropy loss without counterfactual augmentation or fairness regularization. This model is evaluated directly without additional training. For the augmentation setting, the model is trained from scratch using counterfactual data augmentation with 1,000 generated counterfactual pairs per epoch. Identity substitutions are applied during training. The fairness-regularized model is also trained from scratch using counterfactual data augmentation together with the fairness consistency objective. The fairness loss weight is set to $\lambda = 0.5$.

5.4 Results

Paraphrase Detection. Table 2 reports development-set results comparing the baseline model with two fairness-aware training strategies: counterfactual data augmentation (CDA) and CDA combined with fairness regularization.

Model	Dev Accuracy $\uparrow$	Subgroup Gap $\downarrow$	Flip Rate $\downarrow$
Baseline	89.55%	3.65%	4.56%
CDA	89.47%	2.90%	3.55%
CDA + Fairness Regularization	89.56%	4.32%	2.80%

Table 2: Paraphrase detection development results with fairness metrics.

Both fairness-aware training strategies improve prediction stability under identity perturbations relative to the baseline model. Counterfactual data augmentation reduces the prediction flip rate from 4.56% to 3.55% while also lowering the subgroup accuracy gap from 3.65% to 2.90%. This indicates that exposing the model to identity-swapped examples during training encourages the model to rely less on demographic cues.

The fairness-regularized model achieves the lowest prediction flip rate (2.80%), demonstrating the strongest robustness to counterfactual identity substitutions. Importantly, this model maintains essentially identical task performance to the baseline, achieving 89.56% development accuracy compared to 89.55% for the baseline model.

However, the fairness regularization objective increases the subgroup accuracy gap to 4.32%. This suggests that enforcing prediction consistency between original and counterfactual inputs does not necessarily improve parity across demographic subgroups.

Overall, these results illustrate a trade-off between different notions of fairness. Counterfactual data augmentation improves subgroup parity while also reducing prediction instability, whereas fairness regularization provides the strongest robustness to identity perturbations while preserving task accuracy.

On the official leaderboard, our paraphrase detection model achieves a test accuracy of **0.876**, indicating that the proposed fairness-aware training strategies maintain competitive performance on the core task.

Sentiment Classification. Sentiment classification results are shown in Table 3.

Dataset	Development Accuracy
SST (5-class)	50.9%
CFIMDB (binary)	98.4%

Table 3: Sentiment classification performance.

The model achieves strong performance on the binary CFIMDB dataset but lower accuracy on SST. This outcome is expected because SST requires fine-grained classification across five sentiment categories, making it a substantially more challenging task.

Sonnet Generation. For Shakespearean sonnet generation the model achieves a test CHRF score of **41.294** on the leaderboard. This score exceeds the project target threshold and indicates that the model successfully captures stylistic patterns characteristic of Shakespearean poetry. Generated outputs typically exhibit consistent rhyme structures and vocabulary patterns consistent with the sonnet form.

6 Analysis

We analyze the effects of our fairness interventions for GPT-2-based paraphrase detection using ablation studies, subgroup performance breakdowns, error analysis, and training dynamics.

6.1 Ablation Study

To isolate the contribution of each component, we compare three configurations: (1) a **baseline** model trained with standard fine-tuning, (2) **CDA-only**, which adds counterfactual data augmentation, and (3) **CDA + Fairness Regularization**, which adds a symmetric KL-divergence penalty ($\lambda = 0.5$). All models share the same architecture (GPT-2 base, 124M parameters), optimizer (AdamW), learning rate (1×10^{-5}), effective batch size (64), and seed (11711).

Epochs	Mode	Dev Acc.	Subgroup Gap ↓	Flip Rate ↓
2*5	CDA	0.8835	0.0433	0.0392
	CDA + Reg.	0.8856	0.0512	0.0296
10	Baseline	0.8955	0.0365	0.0456
10	CDA	0.8947	0.0290	0.0355
10	CDA + Reg.	0.8956	0.0432	0.0280

Table 4: Ablation results across training configurations. Subgroup Gap is the maximum accuracy difference between identity subgroups. Flip Rate measures the fraction of predictions that change under identity substitution.

Three patterns emerge from Table 4. First, CDA is most effective at reducing subgroup disparity, lowering the gap from 0.0365 to 0.0290 with minimal accuracy change. Second, fairness regularization primarily improves prediction stability, producing the lowest flip rate (0.0280). Third, longer training improves both accuracy and fairness metrics, indicating that the fairness interventions do not hinder convergence.

6.2 Subgroup Performance Breakdown

We group sentence pairs based on the presence of gendered pronouns, gendered nouns, and identity-related names. Pairs without identity terms are labeled “neutral.”

Subgroup	n	Baseline	CDA	CDA + Reg.
Female	1,751	0.8955	0.8972	0.8943
Male	1,833	0.9182	0.9133	0.9127
Name-only	734	0.9305	0.9223	0.9373
Neutral	36,678	0.8941	0.8933	0.8941
Male–Female Gap	—	0.0227	0.0161	0.0184

Table 5: Per-subgroup dev accuracy for the 10-epoch models.

Models consistently achieve higher accuracy on male-gendered pairs than female-gendered pairs. The baseline shows a 2.27 percentage point gap between male and female examples. CDA reduces this gap to 1.61 points, suggesting that gender-swapped training examples help regularize the model’s reliance on gendered cues.

The neutral subgroup accounts for roughly 90% of the development set and remains stable across configurations. This indicates that fairness interventions do not degrade performance on the majority of examples.

6.3 Prediction Stability

Prediction flip rate measures how often a model’s output changes when identity tokens are replaced with counterfactual alternatives such as “he” to “she” or “James” to “Mary.” Lower flip rates indicate that predictions rely less on identity-specific tokens.

The baseline model flips 171 out of 3,751 identity-containing examples (4.56%). CDA reduces this to 133 flips (3.55%), and CDA with fairness regularization further reduces it to 105 flips (2.80%). These results suggest that CDA improves robustness by diversifying the training distribution, while the regularization loss explicitly penalizes prediction inconsistency between original and counterfactual inputs.

However, the fairness-regularized model shows a higher subgroup accuracy gap than the CDA model. This difference highlights an important distinction between fairness metrics. The regularization objective operates at the instance level by aligning predictions between matched pairs, while subgroup accuracy measures aggregate performance across demographic groups.

6.4 Training Dynamics

We also examine how the task loss, fairness loss, and dev accuracy evolve during training for the CDA + Reg. model.

Epoch	Task Loss	Fairness Loss	Dev Acc.
0	0.438	0.064	0.837
2	0.281	0.034	0.874
4	0.227	0.039	0.886
7	0.164	0.049	0.896
9	0.130	0.055	0.896

Table 6: Training dynamics for the CDA + Fairness Regularization model.

As training progresses, task loss decreases and dev accuracy improves, while fairness loss gradually increases. As model predictions become more confident, small differences between original and counterfactual inputs produce larger KL divergence values. The fairness objective limits this divergence but does not eliminate it entirely.

6.5 Limitations

Our evaluation has several limitations. Identity subgroup classification relies on predefined lexicons and may miss implicit cues. Counterfactual augmentation is limited to 1,000 pairs, which likely underestimates its full potential. In addition, the neutral subgroup dominates the dataset, meaning overall accuracy is largely determined by identity-free examples. Finally, our swaps focus on binary gender categories and a limited set of names, which does not capture the full range of demographic attributes that may influence model behavior.

7 Conclusion

In this work we investigate whether fairness-aware training strategies can reduce identity-based prediction instability in pretrained language models while preserving task performance. Using GPT-2 for paraphrase detection on the Quora Question Pairs dataset, we evaluate two fairness interventions: counterfactual data augmentation (CDA) and a consistency-based fairness regularization objective. Our results show that fairness-aware training improves robustness to demographic perturbations without sacrificing predictive accuracy. CDA reduces the subgroup accuracy gap from 3.65% to 2.90%, while fairness regularization achieves the lowest prediction flip rate, reducing instability under identity substitutions from 4.56% to 2.80% while maintaining baseline performance. These findings suggest that different fairness interventions improve different aspects of demographic robustness. Future work could explore richer counterfactual generation strategies, alternative fairness objectives, and evaluation across additional demographic attributes and tasks.

References

- [1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.
- [3] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, 2019.
- [4] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- [5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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, 2018.
- [6] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.
- [7] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.
- [8] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

- [9] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- [10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.