Exploring Modern LLM Gender Bias

Christine Sako, Yoko Morishita, Jordan Andersen

2025-12-03

1 Abstract

This study examines methods to mitigate gender bias in ModernBERT, a transformer model trained on web-scraped text that often reflects societal stereotypes. Such biases can lead NLP systems to associate certain occupations with specific genders, reinforcing discrimination in applications like resume screening. We evaluate three complementary approaches: Counterfactual Data Augmentation (CDA) for data balancing, embedding-level debiasing via vector projection (Debiased Embeddings), and Iterative Nullspace Projection (INLP). Our experiments show that CDA modestly improves F1-Macro from 0.814 to 0.815 and accuracy from 85.9% to 86.1% while reducing gender bias. Using Debiased Embeddings removed gender-direction cosine similarity in all variants, while leaving accuracy mostly unchanged (ModernBERT: 0.7698 to 0.7699, LoRA: 0.8512 to 0.8509). INLP yields the strongest bias reduction with true positive rate gap RMS falling from 0.173 to 0.067, but lowers accuracy from 77.0% to 70.3%.

2 Introduction

A common source of training data for NLP models is web-scraped text, which often includes content from websites, forums, blogs, social media, and other online sources. This text is largely human-generated and provides valuable information that helps capture natural language patterns such as grammar, syntax, contextual relationships and diverse styles. Unfortunately, this rich source of data can also reflect systemic biases in society such as stereotypes, discrimination, and unequal representation, which can influence the behavior of NLP models and propagate these biases in downstream applications.

Gender bias in NLP systems used in resume screening and occupational classification is a prime example of this effect. Models trained on biased data may reinforce occupational stereotypes, systematically associating certain professions with men or women. Traditional group-level accuracy metrics often miss subtle causal effects of gender, necessitating methods that correct for biases present in the training data. Generating semantically consistent counterfactual text, removing gender information from embeddings via projection-based debiasing, and constraining classifier parameters through nullspace projection represent complementary strategies for mitigating gender bias and isolating its influence on model predictions. In this paper, we will explore these described methods and contribute to the literature on gender debiasing by testing them on modern, state-of-the-art models, such as ModernBERT.

To address gender bias in occupation classification, we evaluate three complementary methods that progressively intervene at the data, representation, and embedding levels. Starting from CDA to balance training data, we then analyze projection-based Debiased Embeddings combined with LoRA fine-tuning to directly mitigate gender signals in embeddings while preserving task performance. Finally, we explore INLP as a more aggressive technique that removes linearly decodable gender information at a potential cost to accuracy, revealing tradeoffs between fairness and predictive power.

3 Background

3.1 CDA

Lu et al. (2018) introduced CDA as a method to identify and mitigate gender bias in neural NLP models by creating matched sentence pairs that differ only in gendered terms. Their experiments demonstrated that CDA can significantly reduce model bias without sacrificing downstream task performance. Our work builds on this foundation by extending CDA to a more naturalistic dataset. Additionally, we apply CDA within the framework of modern transformer architectures, evaluating whether the debiasing benefits observed in earlier models like BERT and GPT-2 remain effective with newer, more efficient architectures and in occupation-specific classification tasks. Xie & Lukasiewicz (2023) expanded the exploration of debiasing by combining CDA with parameter-efficient tuning methods, including prompts, prefixes, and adapters. Similarly, Webster et al. (2021) provided key insights into the emergence and measurement of gendered correlations in pre-trained models, highlighting general mitigation strategies like CDA and dropout. Our study contributes to this growing body of work by situating CDA within a focused downstream classification task for profession classification and by testing whether CDA remains effective in mitigating gender bias at scale.

3.2 Projection-Based Debiased Embeddings

Bolukbasi et al. (2016) has highlighted that word embeddings meant to capture semantic relationships from large text corpora can capture societal biases, specifically stereotypical associations. To mitigate this, they proposed a “debiasing” algorithm that identifies a gender subspace in the embedding space and removes gender projections from words that should be neutral (e.g. doctor, teacher, etc.), while simultaneously equalizing explicitly gendered pairs (e.g. king-queen, brother-sister, etc.). This approach preserves important semantic relationships while reducing biased gender associations. Later work by Gonen & Goldberg (2019) and Zhao et al. (2018) has shown that gender information can persist through this process implicitly. Even so, the framework remains a useful and interpretable baseline for understanding and quantifying bias in static embeddings. Applying it in our context allows us to establish a clear point of comparison for evaluating how gender associations manifest in biographical data and how subsequent debiasing strategies might improve upon this foundation.

3.3 Iterative Nullspace Projection

Ravfogel et al. (2020) introduced INLP, noting that gender information is encoded across multiple directions in the embedding space rather than a single direction. INLP is a post hoc method for removing linearly decodable information about a designated attribute from neural representations by repeatedly training linear classifiers to predict the attribute and projecting the representations onto the intersection of their null spaces. The authors show that INLP substantially reduces attribute leakage from BERT encodes, with tradeoffs that can include modest degradation in task accuracy depending on the strength of the removed signal and evaluation setup. We assess the efficacy of INLP on gender bias mitigation with state-of-the-art encoders, particularly focusing on ModernBERT.

4 Data

Our data comes from the Bias in Bios dataset (De-Arteaga et al. (2019)), sourced from Laboratoire Hubert Curien (2023) a collection of professional biographies scraped from the web. Each biography (‘hard_text‘) is paired with one of twenty-eight labels to indicate the true profession of the individual (‘profession‘), and a label indicating their gender (‘gender‘: 0: male, 1: female). The dataset has predefined splits as follows: (Training: 257,478, Validation: 39,642, Test: 99,069) and the training dataset shows a modest gender imbalance (53.9% male, 46.1% female) across professions. Distribution of the 28 professions, gender within each profession, and text length in the training dataset can be seen in **Figure A.1**, **Figure A.2**, and **Figure A.3** in the appendix, respectively.

5 Methods

Our methodology compares three main approaches to reduce gender bias in ModernBERT. This progression reflects increasing degrees of embedding intervention designed to test bias mitigation efficacy and accuracy tradeoffs, with LoRA applied only in the projection-based method to investigate the interaction of fine-tuning with bias representation.

5.1 CDA

To mitigate gender bias in profession classification tasks, we employ Counterfactual Data Augmentation (CDA). This approach builds on work from Lu et al. (2018), where the goal is to reduce the model's reliance on gendered cues by ensuring that it sees an equal number of professions for each of the gender subclasses in the data.

Step 1: Baseline A ModernBert model was trained on the original Bias in Bios dataset to determine baseline accuracy and F1 macro scores. Additional accuracy and F1 metrics were tracked along the gender subclasses for comparison with the later model trained on the CDA dataset.

Step 2: Counterfactual Data Augmentation Counterfactual Data Augmentation - SpaCy was used to generate the counterfactual dataset. The algorithm leveraged the entity labeling method and dependency parsing tool within the SpaCy library to identify the main noun subjects and pronouns. Identified pronouns were mapped to a gender-swapped dictionary and interchanged with their respective opposites. Main subjects were identified and replaced with 'PERSON' to mitigate model bias resulting from traditionally gendered names. To preserve semantic accuracy, the augmentation pipeline included quality control checks using sentence similarity metrics from the sentence transformers model 'all-MiniLM-L6-v2'. Most counterfactuals maintained high similarity scores (above 70%), and quality checks via sampling indicated that CDA effectively preserved the informational content of the biographies while swapping gender signals. The resulting dataset contains 514,956 samples to use for training, compared to the original 257,478. The distribution of semantic similarity scores between the original and augmented bios can be seen in Figure A.4 in the appendix.

Step 3: Training the CDA Model The counterfactual dataset was then used to train a separate ModernBert model and evaluate changes in accuracy and F1 metrics overall and between the subclasses.

5.2 Debiased Embeddings

The Debiased Embeddings approach taken in this project is an extension of the projection-based debiasing technique introduced by Bolukbasi et al. (2016) to contextual embeddings generated by the transformer-based model, ModernBERT. The objective is to remove gender information from embeddings while preserving task-relevant semantic content. To efficiently fine-tune the model for occupation classification, we additionally incorporate Low-Rank Adaption (LoRA) for parameter-efficient (PEFT) updates for 2 of our 4 model variants for comparison. The Debiased Embeddings method can be broken down into 5 steps:

Step 1: Computing the Gender Direction To identify the gender encoding subspace, we extract contextualized embeddings for gendered word pairs (e.g. he-she, father-mother, etc.) appearing in neutral contexts. For each pair, we compute the difference vector between the masculine and feminine forms and average these difference vectors across all word pairs and contexts to obtain a single, stable gender direction vector (g). This direction captures the dominant axis of gender variation in the embedding space and serves as the basis for projection-based debiasing.

Step 2: Direct Projection Debiasing Each embedding vector (e) is debiased by applying an operation that eliminates the gender-associated component, effectively removing the gender-associated component:

$$e_{debiased} = e - (e * g)g$$

Step 3: Embedding Extraction CLS token embeddings are extracted from the ModernBERT encoder for the train, validation, and test splits. This procedure was applied to both the baseline and modernBERT models, ensuring downstream evaluations compare representations produced under identical conditions. Each embedding has a dimensionality of 768, and the resulting matrices are stored for downstream classification and fairness evaluation

Step 4: Fine-Tuning with LoRA LoRA adapters are attached to ModernBERT and fine-tuned to enable efficient, task-specific adaptation to our profession classification objective. This method injects low-rank trainable matrices into attention layers, significantly reducing the number of trainable parameters while simultaneously retaining full-model performance. Training was conducted for 3 epochs, with a batch size of 15, a learning rate of 5e-4, and accuracy as the evaluation metric. The resulting LoRA fine-tuned model provides embeddings that reflect both task learning and potential amplification or attenuation of bias through adaptation.

Step 5: To systematically examine the interaction between fine-tuning with LoRA and debiased embeddings, we evaluate four different ModernBERT-based configurations:

1. *ModernBERT Baseline*: Embeddings directly from the pretrained ModernBERT base model, representing unaltered contextual representations.

2. *ModernBERT Debiased*: Embeddings processed with gender debiasing applied, removing variance along the gender subspace.

3. *LoRA Baseline*: Embeddings from the ModernBERT model, fine-tuned using LoRA on the profession classification task, capturing the effects of parameter-efficient adaptation on bias structure.

4. *LoRA Debiased*: LoRA embeddings with gender debiasing applied, allowing examination of whether post-hoc embedding debiasing remains effective after fine-tuning.

This four-way design enables controlled comparison across two axes: 1) pretraining vs. fine-tuning and 2) non-debiased vs. debiased embeddings. By analyzing performance and fairness metrics across these combinations, we can assess the extent to which gender information persists or reemerges under different model specifications.

5.3 Iterative Nullspace Projection

The goal of the INLP method is to systematically remove linearly encoded gender information from text embeddings while preserving task-relevant semantic structure.

First, ModernBERT was utilized to encode biographical texts into CLS token embeddings, providing dense vector representations of the input data. A logistic regression classifier was first trained to predict occupations as a baseline.

INLP was then applied through 300 iterations of training linear gender classifiers. In each iteration i, a linear gender classifier was trained on the current representation to identify a gender direction w_i . A projection matrix P_i was computed to project the embeddings onto the nullspace orthogonal to w_i and the cumulative projection matrix was updated as $P_{current} * P_i$. After 300 iterations, the final matrix P represents the composition of all individual nullspace projections, effectively reducing linear gender correlations across embedding dimensions.

The debiased embeddings, obtained as $X_{debiased} = X * P$ were used to retrain the logistic regression model for occupation classification, which was compared with the baseline.

6 Results and Discussion

The following Results section presents empirical findings comparing three distinct bias mitigation techniques across fairness and accuracy metrics, revealing how each approach balances the tradeoff between reducing gender bias and maintaining occupation classification performance. These insights inform practical decisions on choosing debiasing strategies tailored to specific deployment requirements.

6.1 CDA

The ModernBERT CDA model demonstrated slight but meaningful improvements over the baseline. The CDA model achieved an F1-Macro of 0.8006 and accuracy of 86.16%, compared to the baseline's F1-Macro of 0.803 and accuracy of 85.9%. More importantly, evaluation of the CDA model revealed significant decreases in the true positive rate gap for males and females, especially in traditionally gendered professions such as teacher, yoga instructor, paralegal, and interior designer, suggesting that CDA was able to improve the model's predictions for certain categories without sacrificing overall accuracy. Despite this achievement, the model still seems to struggle with decreasing the true positive rate gap in certain professions, and even causes a gap increase in others, such as architects and poets. This suggests that it still relies on learned-biases in addition to the counterfactual data to make predictions, and may require stronger debiasing methods to further mitigate the gender gaps. **Figures A.5 and A.6** in the appendix visually demonstrate these reductions and remaining gaps in true positive rates from the baseline to the CDA model.

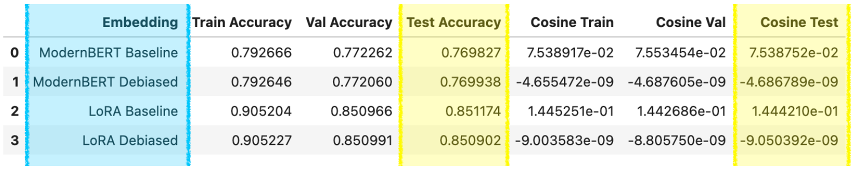
6.2 Debiased Embeddings

We evaluated four embedding conditions to assess how gender-projection debiasing interacts with model architecture and downstream task training: (1) *ModernBERT Baseline*, (2) *ModernBERT Debiased*, (3) *LoRA Baseline*, and (4) *LoRA Debiased*. Results were examined across three dimensions: profession classification accuracy, cosine similarity with the learned gender direction (i.e. amount of bias), and the differences observed between baseline and debiased variants.

When comparing the embeddings of the *ModernBERT Baseline* and *ModernBERT Debiased* models, we find that the baseline ModernBERT embeddings exhibit a measurable alignment with the gender direction (mean cosine ≈ 0.075 across all splits), indicating that gender information is encoded in the representations even for non-gendered inputs. After applying gender-projection debiasing, cosine similarities effectively disappear, confirming that the debiasing procedure successfully removes gender-direction components in practice. Importantly, debiasing the embeddings does not degrade performance, and profession classification accuracy remains virtually unchanged (Baseline Test: 0.7698, Debiased Test: 0.7699). This suggests that ModernBERT's gender-associated subspace is largely orthogonal to the dimensions needed for occupational classification. Thus, projection-based debiasing appears to eliminate gender signals without degrading task-relevant representational structure.

When comparing the embeddings of the *LoRA Baseline* and *LoRA Debiased*, we find that LoRA fine-tuning dramatically improves classification performance (Test accuracy: 0.8512), demonstrating that task-specific adaptation significantly enriches the linear separability of profession embeddings. However, this adaptation substantially amplifies the magnitude of the gender direction, as cosine similarities increase to approximately 0.144 across all splits, which is nearly double that of the baseline ModernBERT value. This suggests that fine-tuning for profession prediction inadvertently increases reliance on gender-correlated representational components. Applying the same projection-based debiasing to the LoRA embeddings again removes the gender component entirely. Crucially, accuracy remains effectively unchanged (LoRA Baseline: 0.8512, LoRA Debiased: 0.8509). Thus, even in a setting where fine-tuning intensifies gender encoding, geometric projection remains an effective mitigation strategy that preserves downstream performance. **Figure 1** displays validation-to-test performance changes and accuracy–debiasing tradeoffs for all four embedding variants, with test set results highlighted in yellow.

Taken together, these results highlight a key tradeoff in bias mitigation. Pretrained models (e.g. ModernBERT) encode gender information moderately, but removal of this component is both possible and does not degrade performance. Fine-tuned models (e.g. LoRA) tend to amplify gender information, likely due to



	Embedding	Train Accuracy	Val Accuracy	Test Accuracy	Cosine Train	Cosine Val	Cosine Test
0	ModernBERT Baseline	0.792666	0.772262	0.769827	7.538917e-02	7.553454e-02	7.538752e-02
1	ModernBERT Debiased	0.792646	0.772060	0.769938	-4.655472e-09	-4.687605e-09	-4.686789e-09
2	LoRA Baseline	0.905204	0.850966	0.851174	1.445251e-01	1.442686e-01	1.444210e-01
3	LoRA Debiased	0.905227	0.850991	0.850902	-9.003583e-09	-8.805750e-09	-9.050392e-09

Figure 1: Table of Accuracy/Debiasing Tradeoff Between Variants

correlations embedded in occupation labels or dataset structure. However, in both conditions, linear gender projection debiasing fully erases representational gender bias without harming predictive accuracy. Thus, projection-based debiasing is a robust method across model variants and training environments. At the same time, the substantial increase in gender direction magnitude after LoRA fine-tuning underscores the importance of applying debiasing *after* task-adaptation, rather than relying solely on pretrained representations. The distributional impact of bias for each profession in both models (as measured by the cosine similarity between profession-specific embeddings and the gender direction) can be found in the Appendix as **Figure A.7**.

6.3 Iterative Nullspace Projection

The results show that INLP effectively mitigates gender-related disparities, as measured by the True Positive Rate (TPR) gap root mean square (RMS) error, although with an expected trade-off in predictive accuracy. For ModernBERT, classification accuracy decreased from 77.0% to 70.3% post-INLP, while TPR Gap RMS reduced from 0.173 to 0.067. This pattern confirms INLP's capability to attenuate linearly encoded gender signals within embeddings, yielding fairer model outputs. **Figure A.8** clearly indicates that INLP removed the correlation between TPR and the relative proportion of women in profession y.

Compared to Ravfogel et al. (2020), ModernBERT begins and ends with lower gender bias than BERT, achieving a smaller final TPR Gap RMS (0.067 versus 0.095) despite slightly lower accuracy. This suggests that ModernBERT's embeddings are inherently less biased and respond better to debiasing interventions. However, balancing fairness and accuracy remains a key challenge.

6.4 Summary

CDA improved F1-Macro from 0.814 to 0.815 and accuracy from 85.9% to 86.1%, with bias-sensitive true positive rate gains in several professions. Projection-based debiasing removed gender alignment in embeddings (cosine similarity dropped from ~ 0.075 to near zero) while keeping accuracy stable around 77%, and LoRA fine-tuning improved accuracy further to 85.1% but amplified gender signals, which were effectively removed by debiasing without accuracy loss. INLP reduced the true positive rate gap RMS from 0.173 to 0.067 but lowered accuracy from 77.0% to 70.3%, demonstrating the accuracy–fairness tradeoff. Interestingly, the Cosine similarity to the gender direction is reduced to near zero by both Debiased Embeddings and INLP, with Debiased Embeddings yielding the smallest residual bias. This suggests that methods explicitly tuned to a single gender axis can outperform broader debiasing on that particular metric. (Reference table: **Figure A.9**)

7 Conclusion

We evaluated three gender-debiasing approaches for ModernBERT: CDA during data preparation, and debiasing/INLP in post-training embedding adjustment. CDA and debiased embeddings reduced bias while maintaining or improving accuracy, showing that data balancing and directional debiasing can mitigate spurious gender correlations without harming task performance. LoRA further preserved accuracy while suppressing the principal gender direction. INLP achieved stronger bias reduction but at a higher accuracy cost, likely due to overlap between gender and occupation signals.

For future research, we propose to explore hybrid strategies that combine methods from CDA, Debiased Embeddings, INLP, and LoRA to enhance both fairness and performance, with configurations tailored to specific task goals (e.g., simpler models, higher accuracy, or broader debiasing). Another direction is to develop bias-aware fine-tuning and adversarial debiasing techniques that jointly optimize accuracy and fairness during model training rather than through post-hoc adjustments.

8 Contributions

J.A. was responsible for the Counterfactual Data Augmentation (CDA) models. C.S. was responsible for the Projection-Based Debiasing models. Y.M. was responsible for the Iterative Nullspace Projection models.

Github Repo:
https://github.com/myoko-codes/266_gender_bias_in_bios

Presentation: https://docs.google.com/presentation/d/1a0wS5Q8CFQsyWThnXCH6g2O3eOq1R5xNBYWA49E1Gso/edit?usp=sharing

A Appendix

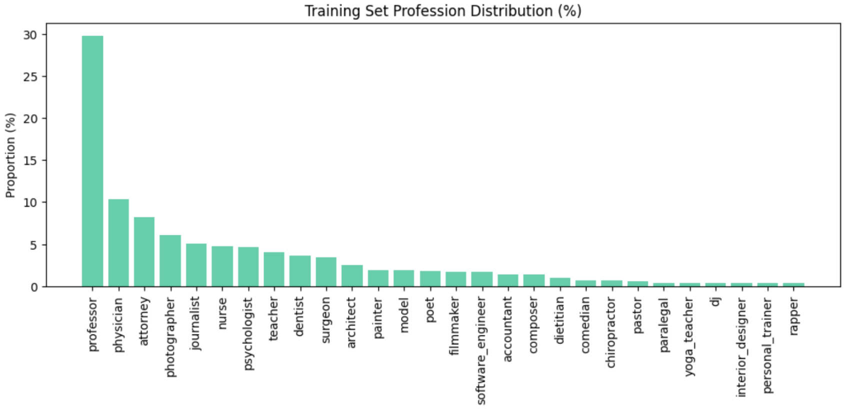


Figure A.1: Distribution of Profession (‘profession‘) in the Training Dataset

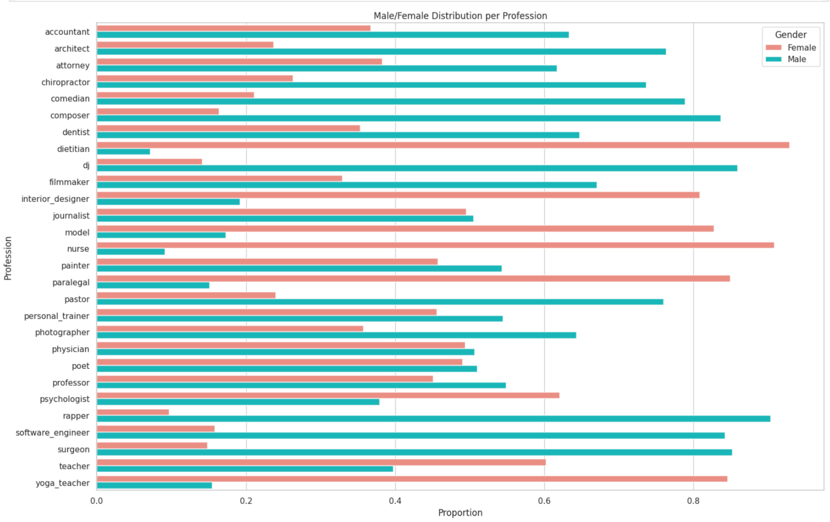


Figure A.2: Distribution of Gender (‘gender‘) in within Profession (‘profession‘) in the Training Dataset

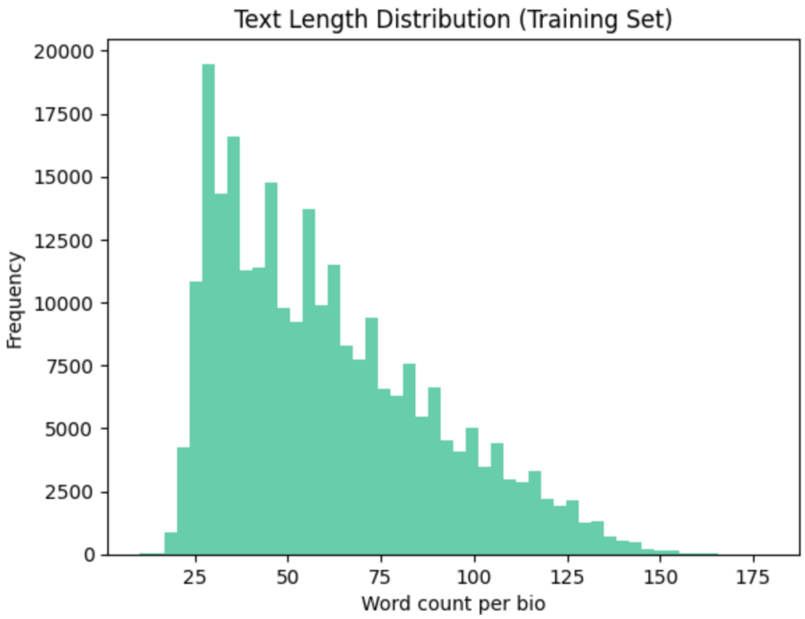


Figure A.3: Distribution of Text Length of hard text in the Training Dataset

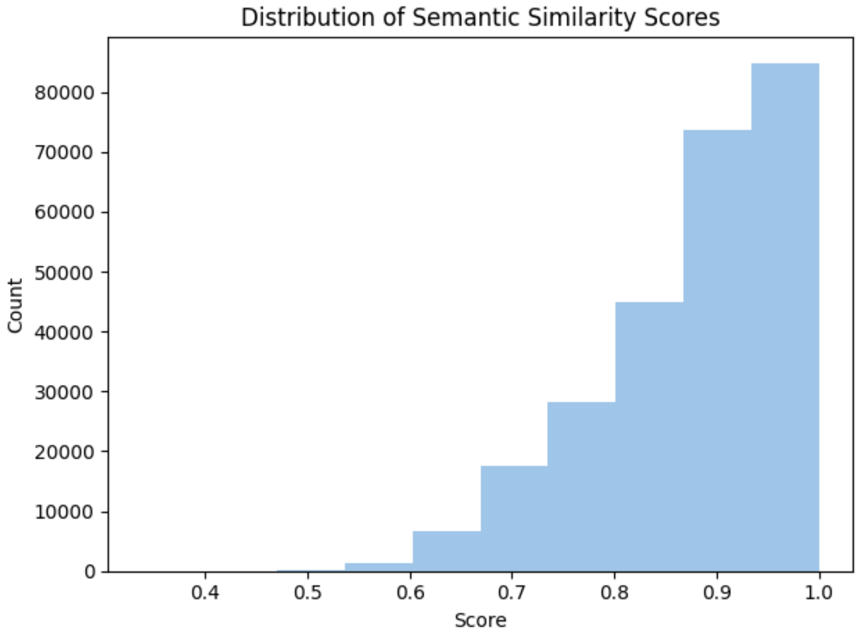


Figure A.4: Distribution of semantic similarity scores between original bios and augmented bios

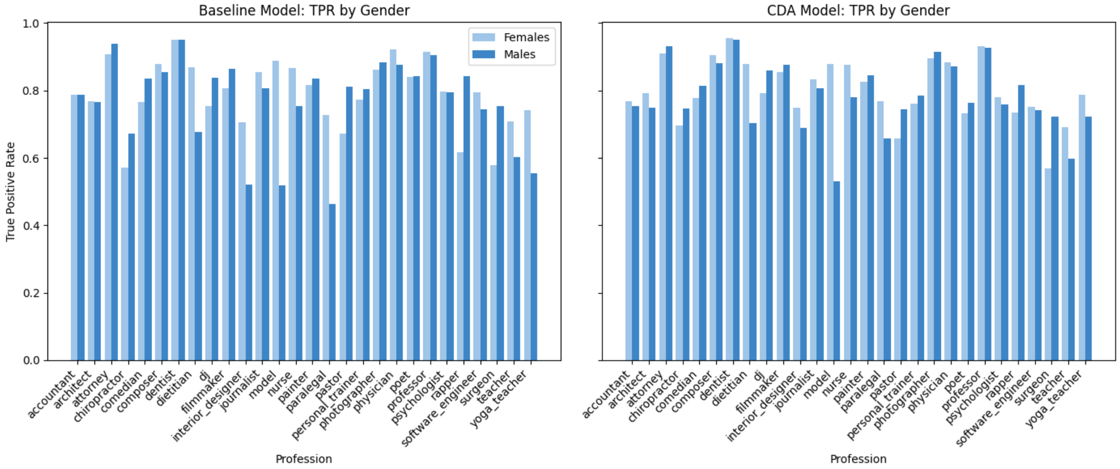


Figure A.5: True positive rates by gender for baseline and CDA models

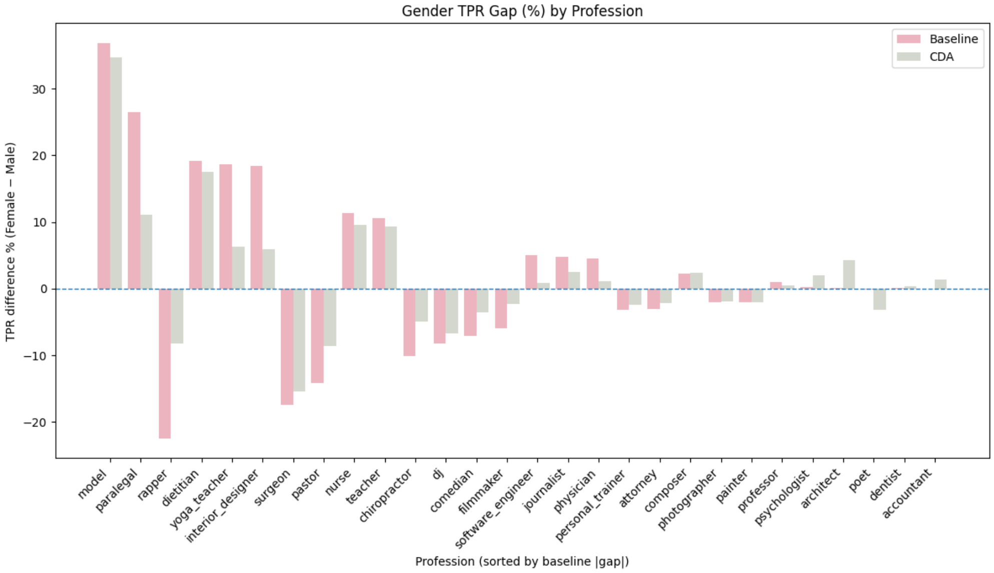


Figure A.6: Change in true positive rate gap (females - males) from baseline to CDA model

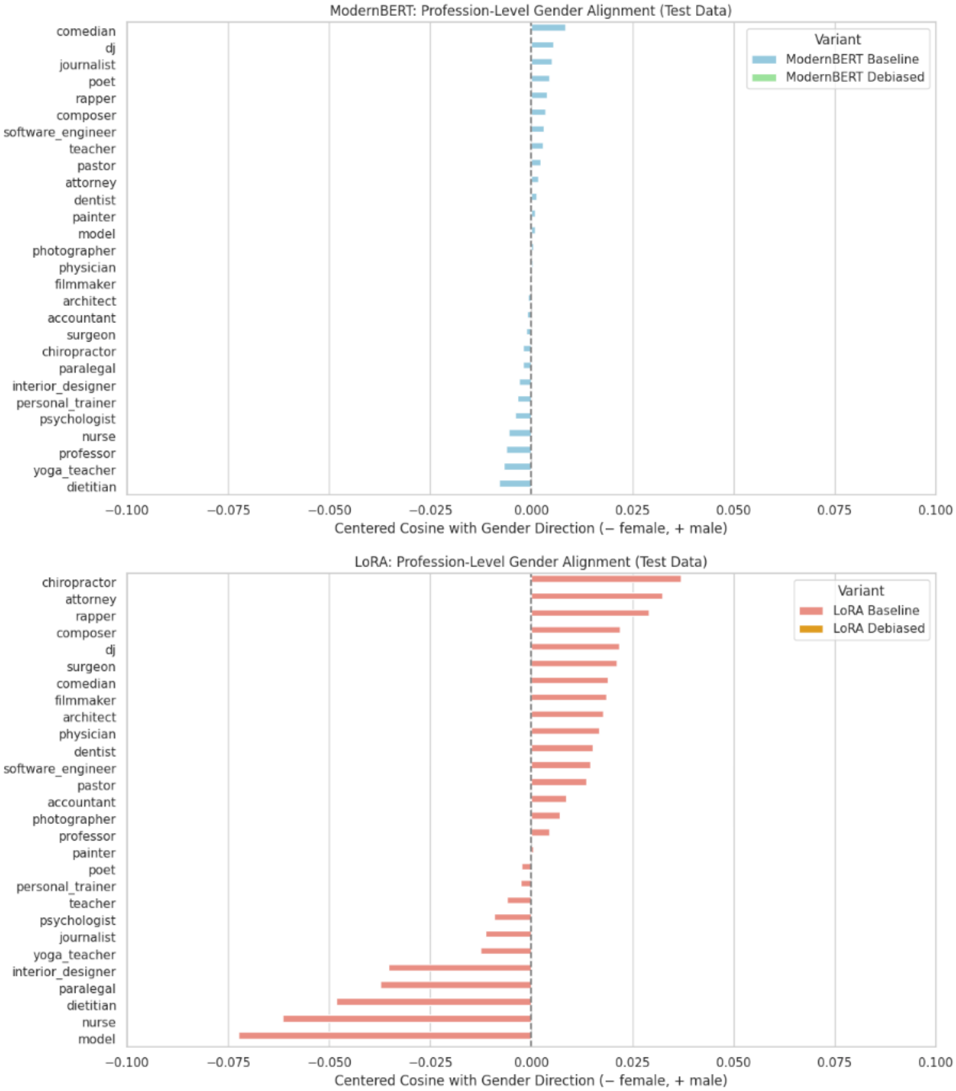


Figure A.7: Profession-Level Gender Alignment for ModernBERT and LoRA Embeddings

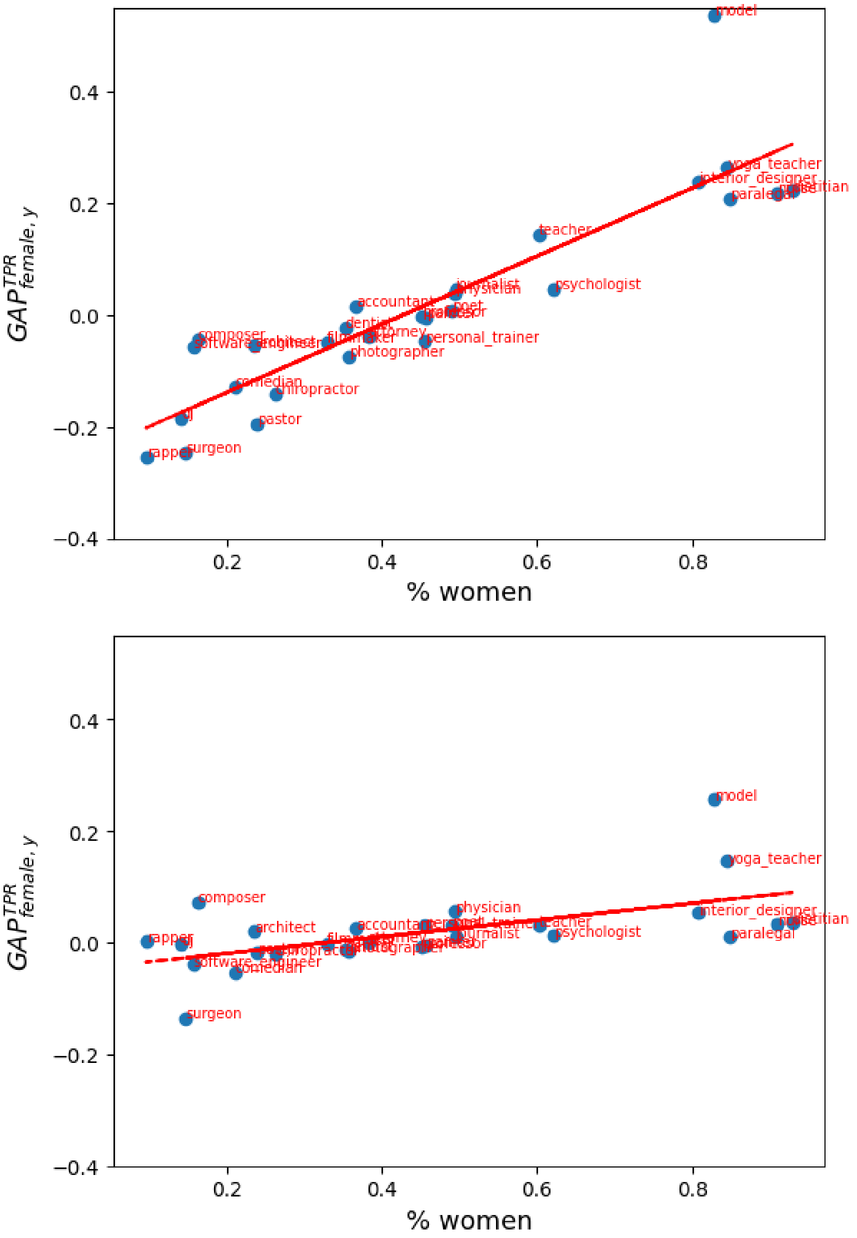
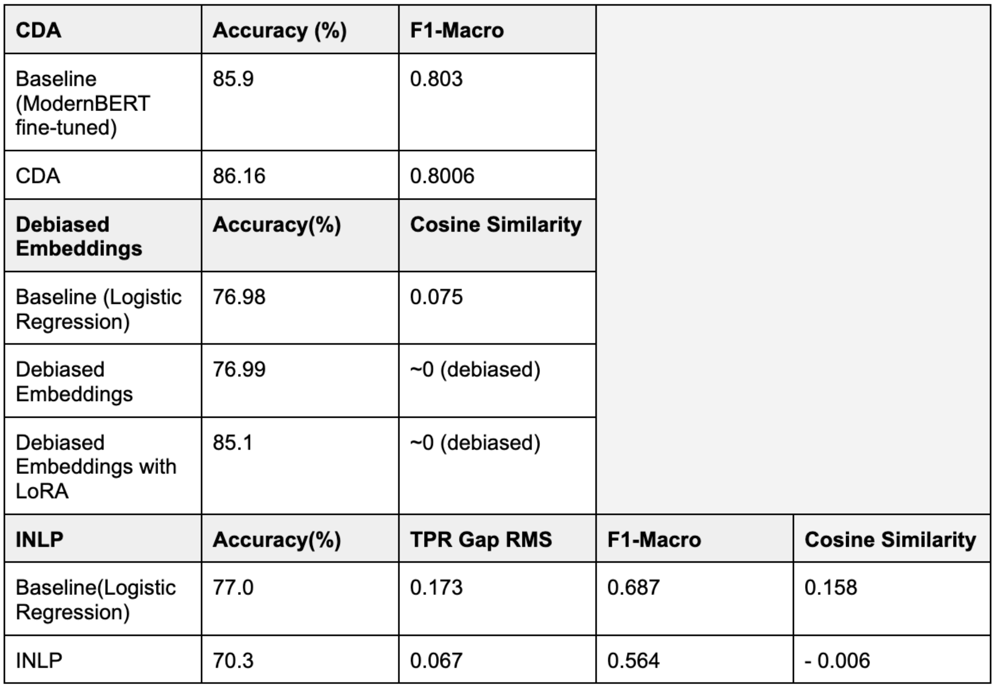


Figure A.8: Correlation between TPR (female, y) and relative proportion of women in profession y (Top: before INLP, Bottom: after INLP)



CDA	Accuracy (%)	F1-Macro		
Baseline (ModernBERT fine-tuned)	85.9	0.803		
CDA	86.16	0.8006		
Debiased Embeddings	**Accuracy(%)**	**Cosine Similarity**		
Baseline (Logistic Regression)	76.98	0.075		
Debiased Embeddings	76.99	~0 (debiased)		
Debiased Embeddings with LoRA	85.1	~0 (debiased)		
INLP	**Accuracy(%)**	**TPR Gap RMS**	**F1-Macro**	**Cosine Similarity**
Baseline(Logistic Regression)	77.0	0.173	0.687	0.158
INLP	70.3	0.067	0.564	- 0.006

Figure A.9: Summary of our three methods' results: CDA achieved the highest accuracy (86.16%). Compared with CDA, INLP reached only 0.564 of its F1-Macro, underscoring a substantial accuracy tradeoff. For cosine similarity, although the comparison is not perfectly one-to-one, INLP reduced gender alignment to near zero but still slightly higher than Debiased Embeddings. This is likely because Debiased Embeddings explicitly remove the same single gender direction used for cosine evaluation, while INLP spreads its effect across many dimensions and therefore may not zero out that specific direction as completely.

References

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*. https://arxiv.org/abs/1607.06520

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 120–128. https://doi.org/10.1145/3287560.3287572

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of NAACL*, 609–614. https://doi.org/10.18653/v1/N19-1061

Laboratoire Hubert Curien. (2023). *Bias in bios*. Available from Hugging Face. https://huggingface.co/datasets/LabHC/bias_in_bios

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender bias in neural natural language processing. *CoRR*, *abs/1807.11714*. https://arxiv.org/pdf/1807.11714

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *Proceedings of ACL*. https://aclanthology.org/2020.acl-main.647/

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2021). Measuring and reducing gendered correlations in pre-trained models. *arXiv Preprint*. https://arxiv.org/abs/2010.06032

Xie, Z., & Lukasiewicz, T. (2023). An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *Proceedings of ACL*. https://aclanthology.org/2023.acl-long.876/

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of NAACL*, 15–20. https://doi.org/10.18653/v1/N18-2003