DATASCI 266

Exploring Modern LLM Gender Bias

Christine Sako

Yoko Morishita

Jordan Andersen

Agenda

01 Background

02 Methods: Counterfactual Data Augmentation

03 Methods: Debiased Embeddings

04 Methods: Iterative Nullspace Projection

05 Results

06 Discussion

Background

Motivation

- Modern models are trained on data that may reinforce common cultural and social stereotypes associated with certain professions
- This can lead to negative effects for downstream occupational or resume tasks

Literature

- Prior studies show Counterfactual Data Augmentation (CDA), Debiased Embeddings, and Iterative Null Space Projection (INLP) can mitigate against bias
- We extend these methods to current modern models to see how they perform

Methods: CDA

Counterfactual Data Augmentation

CDA Approach: Used SpaCy to swap gendered pronouns and PERSON entities, generating gender-flipped biographies and validating each swap with sentence-similarity checks to preserve meaning.

Dataset: Expanded the dataset from 257k to 515k by adding counterfactual biographies, creating a more gender-balanced distribution across professions.

Training: Trained a ModernBERT model on the CDA data to compare accuracy and bias with the baseline.

	hard_text	profession	gender	text_cf	semantic_sim
0	He s also the project lead of and major contr...	21	0	She s also the project lead of and major cont...	0.888533
1	She is able to assess, diagnose and treat mino...	13	1	He is able to assess, diagnose and treat minor...	0.863421
2	Prior to law school, Brittni graduated magna c...	2	1	Prior to law school, PERSON graduated magna cu...	0.792561
3	He regularly contributes to India's First Onli...	11	0	She regularly contributes to India's First Onl...	0.964250
4	He completed his medical degree at Northwester...	21	0	She completed her medical degree at Northweste...	0.964301

Methods: Debiased Embeddings

Debiased Embeddings

Debiasing Method: Computed a gender direction from gendered word pairs and removed that component from ModernBERT embeddings using projection-based debiasing.

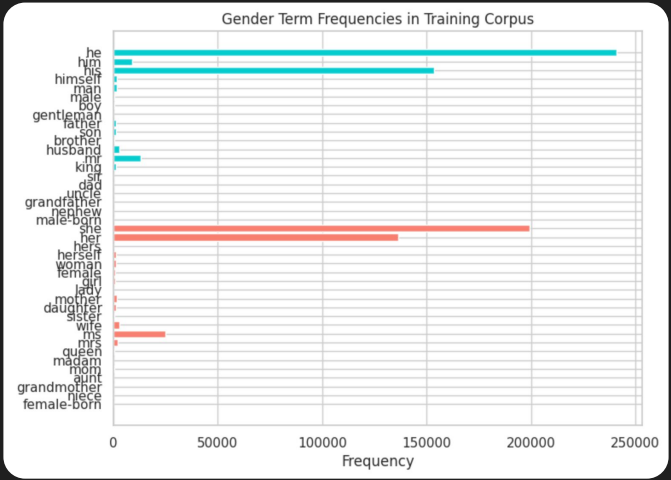
Embeddings & LoRA: Extracted 768-dimensional CLS embeddings and applied LoRA for efficient fine-tuning on occupation classification.

Model Comparison: Evaluated four setups (Baseline, Debiased, LoRA, and LoRA Debiased) to measure how fine-tuning and debiasing jointly affect performance and residual gender bias.

Defining Gender Terms

```
# Defining and printing gender terms
male_terms = [
    "he", "him", "his", "himself", "man", "male", "boy", "gentleman", "father", "son",
    "brother", "husband", "mr", "king", "sir", "dad", "uncle", "grandfather", "nephew", "male-born"
]
female_terms = [
    "she", "her", "hers", "herself", "woman", "female", "girl", "lady", "mother", "daughter",
    "sister", "wife", "ms", "mrs", "queen", "madam", "mom", "aunt", "grandmother", "niece", "female-born"
]
```

Gender Term Frequency



Methods: INLP

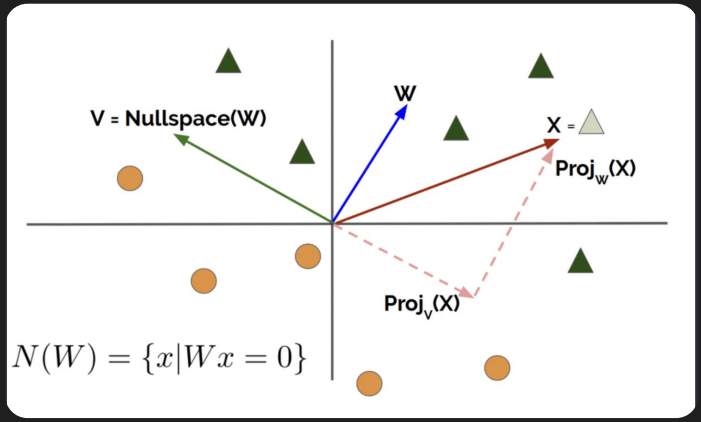
Iterative Null Space Projection

INLP Debiasing: Performed 300 iterations of linear gender classifiers, projecting embeddings into each nullspace to progressively remove linearly encoded gender signals.

Embedding & Baseline: Encoded biographies with ModernBERT CLS embeddings and trained a baseline logistic regression occupation classifier.

Evaluation: Applied the final projection to create debiased embeddings and retrained the classifier to compare against the baseline.

INLP concept in 2D:
X projected onto null space of W, which is trained to lineary distinguish male and female



https://shauli-ravfogel.netlify.app/post/inlp/

Results

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

Observations

CDA Slightly improved accuracy and maintained F1 macro

- CDA significantly reduced the gap in true positive rates between males and females for most professions.
- Some professions (e.g., model, surgeon) saw little improvement, indicating persistent learned biases

Debiased Embeddings removed gender alignment without changing accuracy

- LoRA fine-tuning increased accuracy but amplified gender encoding (cosine ~0.144).
- Post-LoRA debiasing eliminated gender signals while maintaining accuracy (0.8512 to 0.8509).

INLP reduced gender disparities but accuracy dropped

- ModernBERT started with lower bias than BERT and ended with a lower TPR Gap RMS (0.067 vs. 0.095).

Discussion

Key Findings

- **CDA & Debiased Embeddings: r**educe gender bias with minimal accuracy loss
- **INLP:** achieve stronger fairness gains at a steeper accuracy cost, indicating fairness–performance tradeoff
- **LoRA Fine-Tuning:** preserves high accuracy while suppressing the principal gender direction

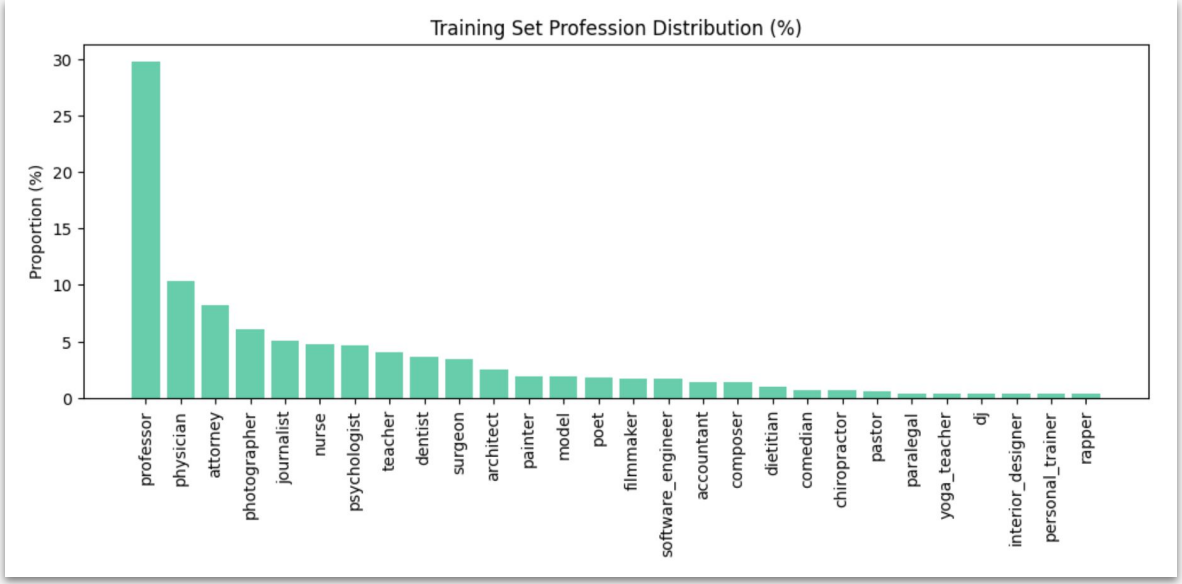
Future Directions

- **Combine multiple methods:** balances strong debiasing with high accuracy, for different deployment priorities, eg.:
  - Accuracy-critical
  - Fairness-critical
  - Complexity of bias (broader/multiple direction bias)
- Bias aware fine-tuning: integrates bias monitoring and adaptive debiasing during training (not just post-hoc)

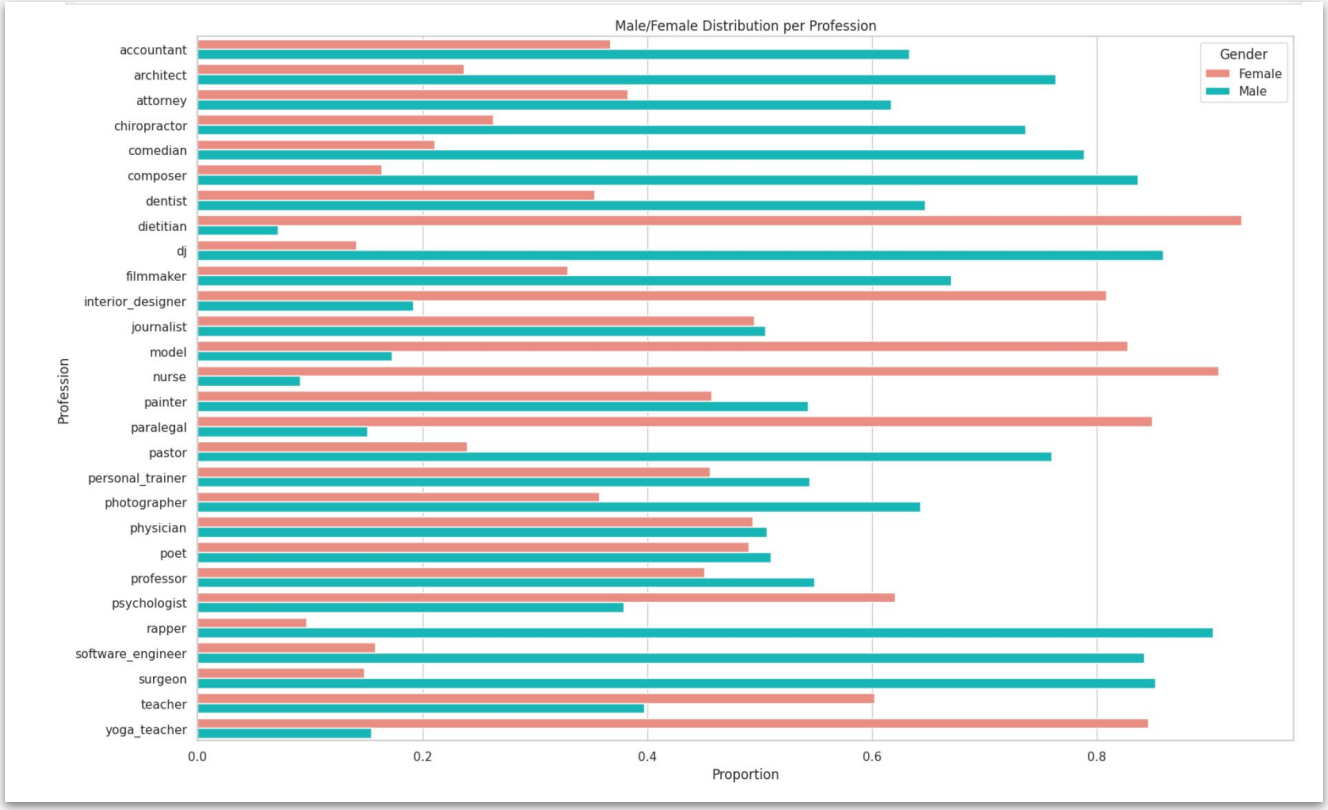
Thank you!

Appendix Plots

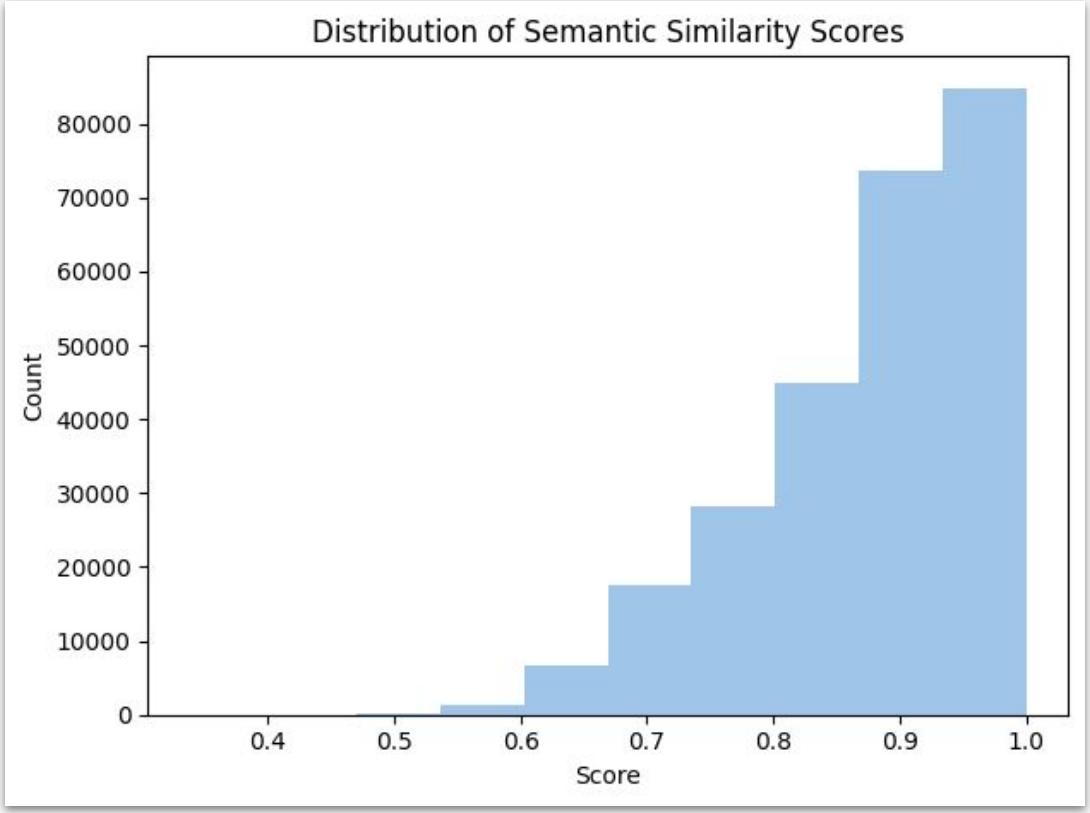
Distribution of Profession (`profession`) in the Training Dataset



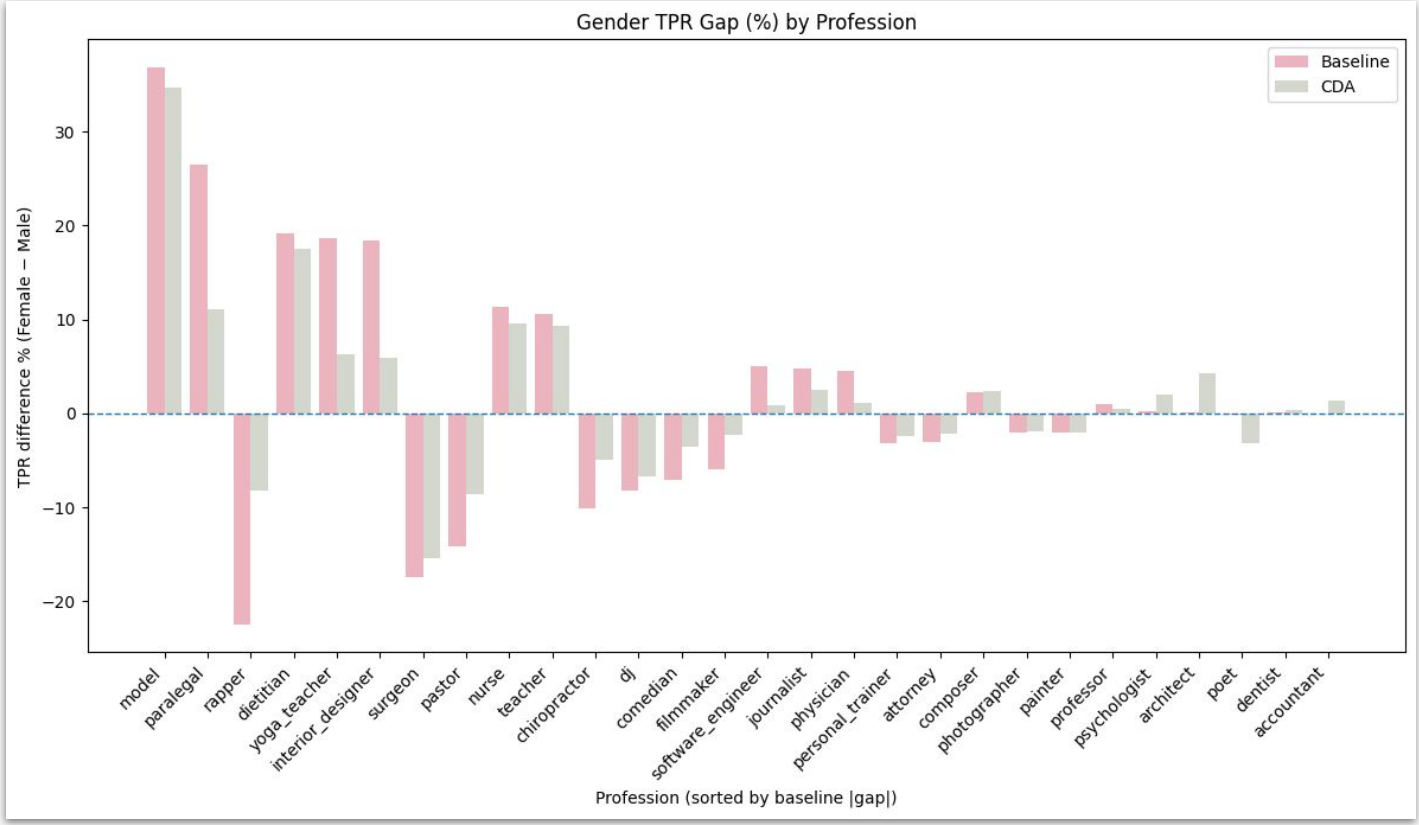
Distribution of Gender (`gender`) in within Profession (`profession`) in the Training Dataset



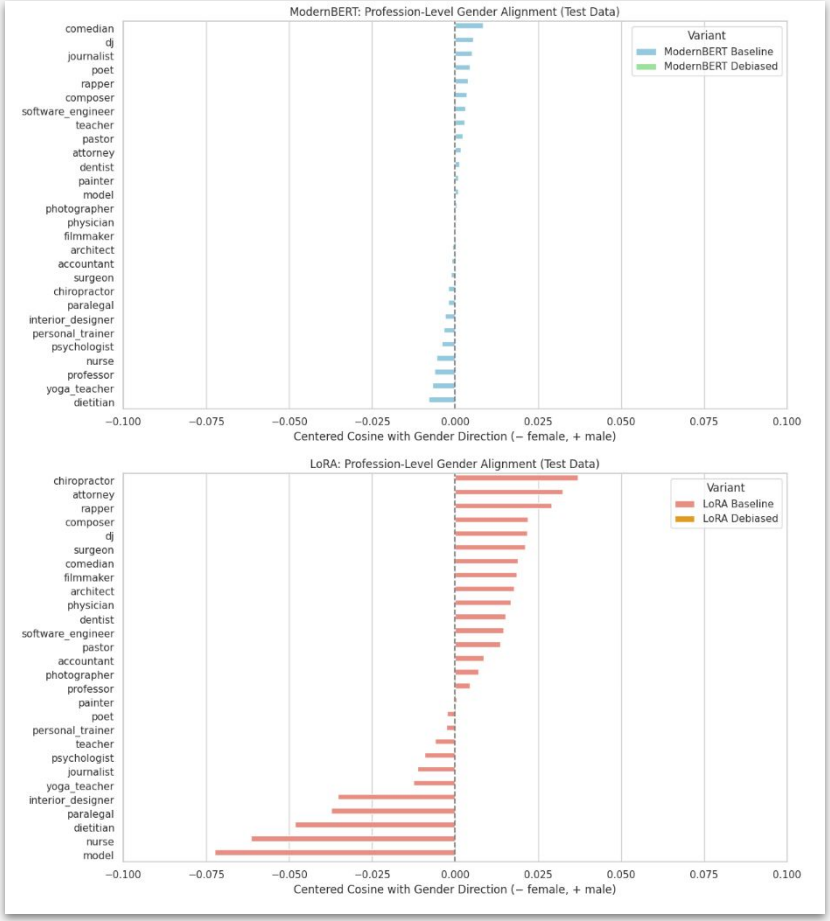
Distribution of semantic similarity scores for original and counterfactual texts



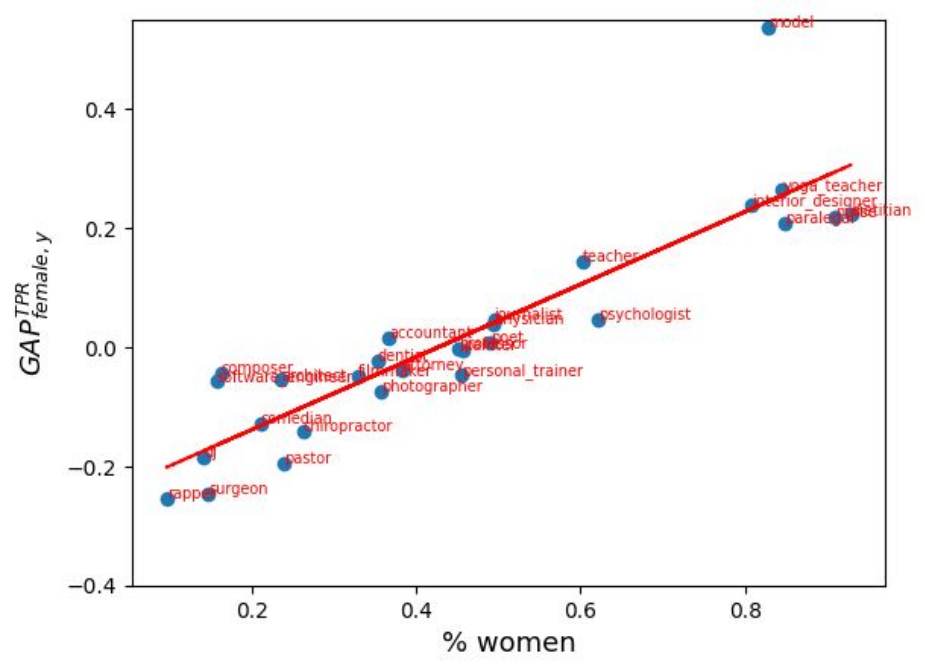
Change in true positive rate gap (females - males) from baseline to CDA model



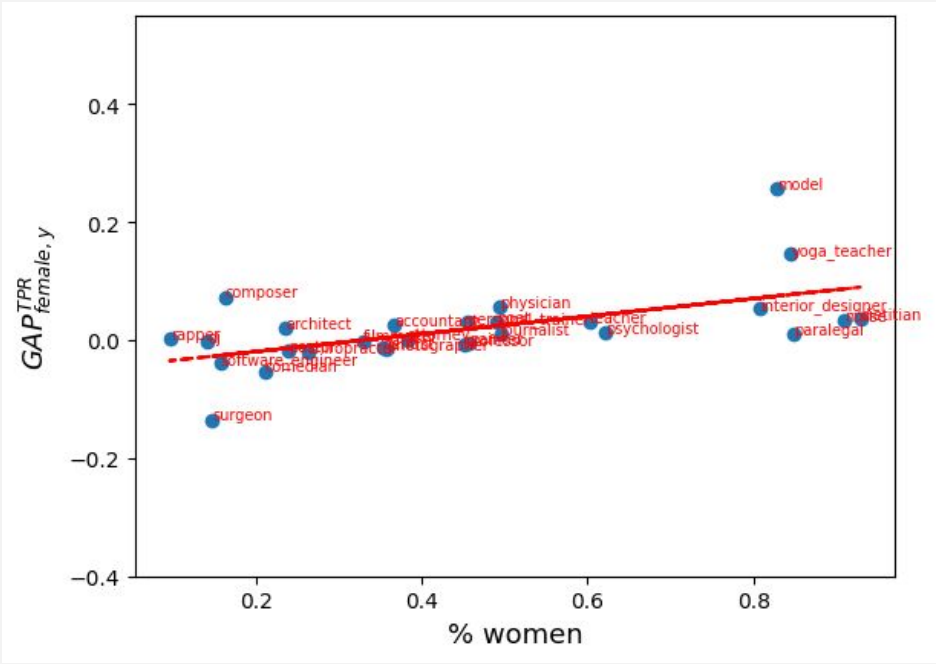
Profession-Level Gender Alignment for ModernBERT and LoRA Embeddings



Correlation between TPR (female, y) & relative proportion of women in profession y



Before INLP



After INLP