

A THEOREM PROOF

Lemma 1. For the majority group (G1), the performance of the model after FairLoRA fine-tuning is:

$$P(M_{\text{FairLoRA}}, G1) = (1 - \text{FPR}) \cdot P(M, G1) + \text{FPR} \cdot P(M_{\text{LoRA}}, G1)$$

Proof.

Definitions and Notations:

- M : the original model.
- M_{LoRA} : the model fine-tuned using LoRA.
- M_{FairLoRA} : the final model after applying FairLoRA fine-tuning.
- G1: the majority group.
- $P(M, G1)$: performance of model M on group G1.
- FPR: False Positive Rate when predicting G2 for samples from G1.

In the context of FairLoRA fine-tuning, the performance of the model on G1 depends on how samples from G1 are classified:

- *True Negatives (TN)*: samples from G1 correctly classified as G1.
- *False Positives (FP)*: samples from G1 incorrectly classified as G2.

Calculating the Performance:

Let N_1 be the total number of samples in G1.

- Number of True Negatives: $\text{TN} = (1 - \text{FPR}) \cdot N_1$.
- Number of False Positives: $\text{FP} = \text{FPR} \cdot N_1$.

For G1, the FairLoRA model uses:

- The original model M for True Negatives.
- The LoRA fine-tuned model M_{LoRA} for False Positives.

Thus, the total performance on G1 is the weighted average:

$$\begin{aligned} P(M_{\text{FairLoRA}}, G1) &= \frac{\text{Performance on TN} + \text{Performance on FP}}{N_1} \\ &= \frac{\text{TN} \cdot P(M, G1) + \text{FP} \cdot P(M_{\text{LoRA}}, G1)}{N_1} \\ &= \frac{[(1 - \text{FPR})N_1 P(M, G1) + \text{FPR}N_1 P(M_{\text{LoRA}}, G1)]}{N_1} \\ &= (1 - \text{FPR}) \cdot P(M, G1) + \text{FPR} \cdot P(M_{\text{LoRA}}, G1). \end{aligned}$$

Therefore, we have:

$$P(M_{\text{FairLoRA}}, G1) = (1 - \text{FPR}) \cdot P(M, G1) + \text{FPR} \cdot P(M_{\text{LoRA}}, G1).$$

This completes the proof. \square

Lemma 2. For the minority group (G2), the performance of the model after FairLoRA fine-tuning is:

$$P(M_{\text{FairLoRA}}, G2) = \text{TPR} \cdot P(M_{\text{LoRA}}, G2) + (1 - \text{TPR}) \cdot P(M, G2)$$

Proof.

Definitions and Notations:

- G2: the minority group.
- $P(M, G2)$: performance of model M on group G2.
- TPR: True Positive Rate when correctly predicting G2 for samples from G2.

For samples from G2, their classification can be:

- *True Positives (TP)*: samples from G2 correctly classified as G2.
- *False Negatives (FN)*: samples from G2 incorrectly classified as G1.

Calculating the Performance:

Let N_2 be the total number of samples in G2.

- Number of True Positives: $TP = TPR \cdot N_2$.
- Number of False Negatives: $FN = (1 - TPR) \cdot N_2$.

For G2, the FairLoRA model uses:

- The LoRA fine-tuned model M_{LoRA} for True Positives.
- The original model M for False Negatives.

Thus, the total performance on G2 is:

$$\begin{aligned}
 P(M_{FairLoRA}, G2) &= \frac{\text{Performance on TP} + \text{Performance on FN}}{N_2} \\
 &= \frac{TP \cdot P(M_{LoRA}, G2) + FN \cdot P(M, G2)}{N_2} \\
 &= \frac{[TPR N_2 P(M_{LoRA}, G2) + (1 - TPR) N_2 P(M, G2)]}{N_2} \\
 &= TPR \cdot P(M_{LoRA}, G2) + (1 - TPR) \cdot P(M, G2).
 \end{aligned}$$

Therefore, we have:

$$P(M_{FairLoRA}, G2) = TPR \cdot P(M_{LoRA}, G2) + (1 - TPR) \cdot P(M, G2).$$

This completes the proof. \square

Theorem 1. To ensure that FairLoRA does not degrade the overall performance of the model, the ratio of the true positive rate (TPR) to the false positive rate (FPR) must satisfy:

$$\frac{TPR}{FPR} \geq \frac{(1 - p)}{p} \cdot \frac{P(M, G1) - P(M_{LoRA}, G1)}{P(M_{LoRA}, G2) - P(M, G2)}$$

Proof.

Definitions and Notations:

- $p = \frac{N_2}{N_1 + N_2}$: proportion of samples from G2.
- $(1 - p)$: proportion of samples from G1.
- $\Delta P(G1)$: change in performance on G1.
- $\Delta P(G2)$: change in performance on G2.
- ΔP : overall change in performance.

Calculating the Change in Performance for G1:

From Theorem 1, the performance change on G1 is:

$$\begin{aligned}
 \Delta P(G1) &= P(M_{FairLoRA}, G1) - P(M, G1) \\
 &= [(1 - FPR)P(M, G1) + FPRP(M_{LoRA}, G1)] - P(M, G1) \\
 &= -FPR \cdot P(M, G1) + FPR \cdot P(M_{LoRA}, G1) \\
 &= FPR \cdot [P(M_{LoRA}, G1) - P(M, G1)].
 \end{aligned}$$

Calculating the Change in Performance for G2:

From Theorem 2, the performance change on G2 is:

$$\begin{aligned}\Delta P(G2) &= P(M_{\text{FairLoRA}}, G2) - P(M, G2) \\ &= [\text{TPR}P(M_{\text{LoRA}}, G2) + (1 - \text{TPR})P(M, G2)] - P(M, G2) \\ &= -\text{TPR} \cdot P(M, G2) + \text{TPR} \cdot P(M_{\text{LoRA}}, G2) \\ &= \text{TPR} \cdot [P(M_{\text{LoRA}}, G2) - P(M, G2)].\end{aligned}$$

Calculating the Overall Change in Performance:

The overall change is the weighted sum:

$$\Delta P = (1 - p) \cdot \Delta P(G1) + p \cdot \Delta P(G2).$$

Substituting the expressions for $\Delta P(G1)$ and $\Delta P(G2)$:

$$\Delta P = (1 - p) \cdot \text{FPR}[P(M_{\text{LoRA}}, G1) - P(M, G1)] + p \cdot \text{TPR}[P(M_{\text{LoRA}}, G2) - P(M, G2)].$$

Setting the Condition for No Performance Degradation:

To ensure the overall performance does not degrade ($\Delta P \geq 0$), we require:

$$(1 - p) \cdot \text{FPR}[P(M_{\text{LoRA}}, G1) - P(M, G1)] + p \cdot \text{TPR}[P(M_{\text{LoRA}}, G2) - P(M, G2)] \geq 0.$$

Assuming Performance Changes:

- Let $\Delta P_{G1} = P(M_{\text{LoRA}}, G1) - P(M, G1)$ (likely negative).
- Let $\Delta P_{G2} = P(M_{\text{LoRA}}, G2) - P(M, G2)$ (positive).

Rewriting the inequality:

$$(1 - p) \cdot \text{FPR} \cdot \Delta P_{G1} + p \cdot \text{TPR} \cdot \Delta P_{G2} \geq 0.$$

Solving for $\frac{\text{TPR}}{\text{FPR}}$:

1. Isolate the positive term:

$$p \cdot \text{TPR} \cdot \Delta P_{G2} \geq -(1 - p) \cdot \text{FPR} \cdot \Delta P_{G1}.$$

2. Since $\Delta P_{G1} < 0$, $-\Delta P_{G1} > 0$:

$$p \cdot \text{TPR} \cdot \Delta P_{G2} \geq (1 - p) \cdot \text{FPR} \cdot (-\Delta P_{G1}).$$

3. Divide both sides by $p \cdot \Delta P_{G2}$ (which is positive):

$$\text{TPR} \geq \frac{(1 - p)}{p} \cdot \frac{\text{FPR} \cdot (-\Delta P_{G1})}{\Delta P_{G2}}.$$

4. Divide both sides by FPR (assuming $\text{FPR} > 0$):

$$\frac{\text{TPR}}{\text{FPR}} \geq \frac{(1 - p)}{p} \cdot \frac{-\Delta P_{G1}}{\Delta P_{G2}}.$$

5. Substitute back the definitions of ΔP_{G1} and ΔP_{G2} :

$$\frac{\text{TPR}}{\text{FPR}} \geq \frac{(1 - p)}{p} \cdot \frac{P(M, G1) - P(M_{\text{LoRA}}, G1)}{P(M_{\text{LoRA}}, G2) - P(M, G2)}.$$

Therefore, the ratio of the True Positive Rate to the False Positive Rate must satisfy:

$$\frac{\text{TPR}}{\text{FPR}} \geq \frac{(1 - p)}{p} \cdot \frac{P(M, G1) - P(M_{\text{LoRA}}, G1)}{P(M_{\text{LoRA}}, G2) - P(M, G2)}.$$

This condition ensures that the positive impact on G2 outweighs the negative impact on G1, preventing overall performance degradation.

This completes the proof. \square

B IMPLEMENTATION DETAILS OF FAIRLORA

This section provides the implementation details for FairLoRA, focusing on the group discriminator training, fine-tuning dataset construction, and FairLoRA training configuration. Key components of the implementation are presented in pseudocode to facilitate understanding and reproducibility.

Group Discriminator Training

To effectively identify sensitive attributes, we trained a group discriminator D_ϕ that takes hidden layer representations from a pre-trained model as input and outputs the corresponding sensitive attribute labels. Specifically, we used the penultimate hidden states $h_\theta(x) \in \mathbb{R}^{T \times d}$ as input, where T represents the sequence length and d is the dimensionality of the hidden states.

To aggregate the sequence representation into a global vector, we employed attention pooling, which assigns importance weights to different time steps. This allows the model to focus on the most relevant parts of the sequence when predicting sensitive attributes.

To mitigate bias in predicting sensitive attributes, we employed the worst-group cross-entropy loss:

$$\mathcal{L}_{\text{worst}} = \max_{g \in \mathcal{G}} \mathbb{E}_{(x,s) \sim P_g} [\ell(D_\phi(h_{\text{pool}}(x)), s)],$$

where \mathcal{G} represents the set of all groups, P_g is the data distribution for group g , s is the sensitive attribute label, and $\ell(\cdot)$ denotes the cross-entropy loss function.

The combined pseudocode for the attention pooling mechanism and the group discriminator network is presented below.

Algorithm 1 Group Discriminator with Attention Pooling

Require: Hidden states $h \in \mathbb{R}^{T \times d}$

Ensure: Predicted sensitive attribute label \hat{s}

1: **Attention Pooling:**

2: Initialize learnable parameter vector $w \in \mathbb{R}^d$

3: **for** $t = 1$ **to** T **do**

4: Compute attention score: $a_t \leftarrow w^\top h_t$

▷ Scalar value

5: **end for**

6: Compute attention weights: $\alpha \leftarrow \text{softmax}([a_1, a_2, \dots, a_T])$

7: Compute pooled representation: $h_{\text{pool}} \leftarrow \sum_{t=1}^T \alpha_t h_t$

8: **Group Discriminator Network:**

9: Compute hidden layer activation: $z \leftarrow \text{ReLU}(W_1 h_{\text{pool}} + b_1)$

▷ $W_1 \in \mathbb{R}^{d_1 \times d}$

10: Compute output logits: $o \leftarrow W_2 z + b_2$

▷ $W_2 \in \mathbb{R}^{2 \times d_1}$

11: Compute predicted probabilities: $\hat{p} \leftarrow \text{sigmoid}(o)$

12: Predict sensitive attribute: $\hat{s} \leftarrow \arg \max \hat{p}$

In this algorithm:

Attention Pooling (Lines 2–7): We compute attention scores for each time step using the learnable parameter vector w . The attention weights α are obtained by applying the softmax function to the attention scores. The pooled representation h_{pool} is then calculated as the weighted sum of the hidden states.

Group Discriminator Network (Lines 8–12): The pooled representation h_{pool} is fed into a fully connected layer with ReLU activation to obtain the hidden activation z . A second linear layer computes the logits o , which are transformed into probabilities \hat{p} using the sigmoid function. The predicted sensitive attribute label \hat{s} is determined by taking the class with the highest probability.

By combining the attention pooling mechanism with the group discriminator network in a single algorithm, we provide a clear and concise representation of how the discriminator processes the input hidden states to predict sensitive attributes.

Using all available data in these experiments ensures that the discriminators achieve high accuracy, thereby improving the model’s capacity to debias effectively without compromising performance. For scenarios with limited sensitive attribute labels, results are presented separately in Table 3.

Partition of dataset

For CelebA and MultiNLI, we used the official splits provided in the respective documentation, following the standard training and test set divisions. For HateXplain, since the official split is not provided, we followed the approach of Lu et al. (2024), where 50% of the samples were used as the test set.

FairLoRA Fine-tuning Dataset Construction

The fine-tuning dataset was constructed to ensure class balance through the following steps:

- **Data with Sensitive Attribute Labels:** We selected samples with a sensitive attribute label of $s = 1$ and performed undersampling to balance the classes.
- **Data without Sensitive Attribute Labels:** A trained discriminator D_ϕ was used to assign pseudo-labels for sensitive attributes. Samples predicted as $s = 1$ were selected, and undersampling was applied to balance the class distribution.

FairLoRA Training Configuration

We employed the AdamW optimizer for training, which effectively handles weight decay and improves generalization. The learning rate was set to 1×10^{-5} to ensure stable convergence during fine-tuning. Training was conducted for 2 epochs, as this was sufficient for the model to converge without overfitting. To maintain consistency, τ was fixed at 0.5 across all experiments. Additionally, we used five different random seeds (5, 15, 25, 35, 45) for each set of experiments to ensure robustness. A validation set can also be utilized to guide hyperparameter tuning if needed.

Pseudocode Implementation

During training, all LoRA adjustments are retained to allow the model to fully learn from the FairLoRA fine-tuning dataset. During inference, the discriminator’s output selectively activates the LoRA adjustments for samples predicted as belonging to sensitive groups. This design ensures that model adjustments are targeted to reduce bias where needed, while maintaining both efficiency and overall performance.

FairLoRA can be extended to accommodate multiple sensitive attributes by introducing additional discriminators and LoRA modules.

Algorithm 2 FairLoRA Forward Pass with Multiple Sensitive Attributes

Require: Input features x , discriminator outputs $\text{dis}_1, \text{dis}_2, \dots, \text{dis}_k$, training mode flag `training`

- 1: Compute base output: $y_{\text{base}} \leftarrow \text{LinearLayer}(x)$
- 2: **for** $i = 1$ to k **do**
- 3: Compute LoRA adjustment: $y_{\text{lora}_i} \leftarrow \text{LoRALayer}_i(x)$
- 4: Determine if LoRA _{i} should be applied: $\text{apply_lora}_i \leftarrow \text{dis}_i > \tau$
- 5: **if not** `training` **then**
- 6: $y_{\text{lora}_i}[\neg \text{apply_lora}_i] \leftarrow 0$
- 7: **end if**
- 8: **end for**
- 9: **return** $y \leftarrow y_{\text{base}} + \sum_{i=1}^k y_{\text{lora}_i}$

This approach enhances the fairness of the model without requiring full access to all sensitive attribute labels, ensuring fairer treatment of underrepresented groups while preserving overall performance.

C COMPREHENSIVE COMPARISON OF EXPERIMENTAL DATA

The evaluation metrics employed in the presented tables are critical for assessing both the performance and fairness of the models:

- **Accuracy (ACC):** The overall proportion of correctly predicted instances among all samples.
- **Balanced Accuracy (BA):** Accounts for class imbalance by computing the average recall obtained on each class. It is calculated as:

$$BA = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

- **Worst Group Accuracy (WGA):** The lowest accuracy observed among all evaluated groups (e.g., different genders, races), highlighting the model’s performance on the most disadvantaged group.
- **Demographic Parity (DP):** Measures the difference in positive prediction rates across different groups. A lower DP indicates more equitable positive prediction distributions among groups.
- **Equal Opportunity (EOp):** Assesses the disparity in true positive rates (TPR) between groups. A smaller EOp suggests that the model provides similar chances of correct positive predictions across groups.
- **Equalized Odds Difference (EOD):** Considers both TPR and false positive rate (FPR) differences between groups. Lower EOD values indicate more balanced predictive performance across groups in terms of both positive and negative classes.
- **Average Error Rate (AER):** The mean error rate across different groups. A lower AER signifies an overall reduction in model errors.

C.1 COMPARATIVE ANALYSIS OF DEBIASING FOR SINGLE SENSITIVE ATTRIBUTE

The analysis of Table 4 involves evaluating the performance and fairness metrics of different models on the CelebA dataset.

Table 4: Performance and Fairness Metrics of Models on the CelebA Dataset

Model	ACC↑(%)	BA↑(%)	WGA↑(%)	DP↓(%)	EOp↓(%)	EOD↓(%)	AER↑(%)
ERM	95.8 ± 0.1	95.7 ± 0.0	77.9 ± 2.6	37.1 ± 0.6	17.5 ± 2.9	10.0 ± 1.7	69.7 ± 3.9
+ FL Min.	95.8 ± 0.2	95.8 ± 0.1	82.0 ± 2.2	37.3 ± 0.5	14.2 ± 2.4	8.5 ± 1.4	68.7 ± 2.9
+ FL Maj.	95.9 ± 0.1	95.6 ± 0.1	77.2 ± 2.8	36.9 ± 0.6	17.8 ± 3.0	10.0 ± 1.6	67.8 ± 3.0
+ FL All	95.9 ± 0.1	95.8 ± 0.1	81.3 ± 1.5	37.1 ± 0.3	14.6 ± 1.7	8.6 ± 1.0	70.3 ± 4.1
GroupDRO	94.4 ± 0.5	94.4 ± 0.4	87.4 ± 1.4	35.1 ± 0.5	7.5 ± 1.2	4.8 ± 0.6	81.8 ± 6.7
+ FL Min.	94.4 ± 0.5	94.6 ± 0.4	88.8 ± 1.5	35.4 ± 0.4	6.8 ± 1.3	4.7 ± 0.5	83.3 ± 6.7
+ FL Maj.	94.7 ± 0.4	94.6 ± 0.4	84.4 ± 1.1	35.4 ± 0.3	9.7 ± 0.9	5.9 ± 0.4	72.1 ± 3.6
+ FL All	94.7 ± 0.3	94.7 ± 0.3	85.9 ± 1.6	35.6 ± 0.2	9.0 ± 1.6	5.7 ± 0.8	75.1 ± 6.1
DFR	94.3 ± 1.4	94.8 ± 1.0	86.0 ± 2.0	37.5 ± 0.6	11.1 ± 1.6	7.7 ± 0.8	75.1 ± 4.4
+ FL Min.	94.5 ± 1.2	95.0 ± 0.9	87.8 ± 1.9	37.4 ± 0.8	9.6 ± 1.3	6.9 ± 0.8	78.7 ± 8.4
+ FL Maj.	95.6 ± 0.1	95.7 ± 0.0	83.3 ± 2.1	37.2 ± 0.5	13.1 ± 2.3	8.1 ± 1.3	72.3 ± 5.5
+ FL All	95.4 ± 0.1	95.7 ± 0.1	86.0 ± 1.1	37.3 ± 0.3	11.0 ± 1.2	7.1 ± 0.6	74.5 ± 6.1
Lu et al. (2024)	95.4 ± 0.4	95.6 ± 0.4	81.4 ± 4.8	36.8 ± 0.5	14.1 ± 4.1	8.3 ± 2.0	68.7 ± 5.3
+ FL Min.	95.5 ± 0.4	95.7 ± 0.3	86.8 ± 2.2	36.7 ± 0.5	9.8 ± 1.6	6.2 ± 0.7	75.9 ± 8.2
+ FL Maj.	95.9 ± 0.3	95.7 ± 0.3	80.4 ± 4.3	36.7 ± 0.6	14.8 ± 3.5	8.6 ± 1.7	67.3 ± 4.5
+ FL All	95.6 ± 0.3	95.8 ± 0.2	86.6 ± 2.1	36.7 ± 0.7	10.0 ± 1.4	6.3 ± 0.6	75.7 ± 9.0

* Bold values indicate the best performance in each category.

The **ERM model** achieves high overall accuracy (ACC) and balanced accuracy (BA), with scores of approximately 95.8% and 95.7%, respectively. However, the model presents fairness concerns as indicated by the worst-group accuracy (WGA), which is relatively low at 77.9%. This suggests suboptimal performance for the least advantaged group. Incorporating the **FL Min.** strategy increases the WGA to 82.0%, demonstrating improved performance on the worst-performing group. Additionally, there is a reduction in the Equal Opportunity (EOp) metric from 17.5% to 14.2% and in Equalized Odds Difference (EOD) from 10.0% to 8.5%, indicating a significant decrease in group disparities and an overall enhancement in fairness.

The **GroupDRO model** initially performs well with a high WGA of 87.4%, reflecting strong baseline performance for the worst-performing group. When **FL Min.** is applied, the WGA further increases to 88.8%, enhancing the model’s robustness across groups. Moreover, there are decreases in EOp from 7.5% to 6.8% and in EOD from 4.8% to 4.7%, implying a reduction in group disparities and improved fairness metrics.

The **DFR model** attains a WGA of 86.0%, suggesting favorable fairness performance at the baseline level. With the application of **FL Min.**, the WGA improves to 87.8%, indicating better performance on the worst-performing group. Concurrently, the EOp decreases from 11.1% to 9.6%, and the EOD reduces from 7.7% to 6.9%, which enhances fairness by mitigating disparities between different groups.

The **Lu et al. (2024) model** starts with a WGA of 81.4%, highlighting room for improvement in addressing the worst-performing group. Upon incorporating **FL Min.**, the WGA significantly increases to 86.8%, indicating substantial improvement for disadvantaged groups. Additionally, notable reductions are observed in EOp from 14.1% to 9.8%, and in EOD from 8.3% to 6.2%, demonstrating enhanced fairness by reducing inter-group disparities.

Table 5: Performance comparison across different attributes of CelebA dataset.

Method	Heavy Makeup			Wearing Lipstick		
	ACC↑(%)	WGA↑(%)	EOD↓(%)	ACC↑(%)	WGA↑(%)	EOD↓(%)
ERM	95.8 ± 0.1	45.4 ± 3.2	27.9 ± 1.9	95.8 ± 0.1	57.4 ± 3.5	29.3 ± 2.4
+ FL Min.	95.8 ± 0.1	54.5 ± 3.1	24.4 ± 1.7	95.8 ± 0.2	63.0 ± 2.7	25.1 ± 2.0
GroupDRO	94.4 ± 0.5	65.4 ± 2.7	25.8 ± 1.6	94.4 ± 0.5	70.2 ± 2.5	25.9 ± 1.9
+ FL Min.	94.4 ± 0.4	70.1 ± 2.5	22.7 ± 1.5	94.5 ± 0.4	74.3 ± 2.4	22.5 ± 1.9
DFR	94.3 ± 1.4	58.0 ± 2.2	27.0 ± 1.8	94.3 ± 1.4	68.1 ± 1.9	26.7 ± 1.8
+ FL Min.	94.5 ± 1.5	63.8 ± 1.9	24.1 ± 2.0	94.4 ± 1.4	73.2 ± 2.0	22.3 ± 1.7
Lu et al.	95.4 ± 0.4	61.4 ± 2.5	28.0 ± 2.2	95.4 ± 0.4	67.8 ± 2.1	27.5 ± 1.7
+ FL Min.	95.6 ± 0.5	69.8 ± 2.9	23.2 ± 2.5	95.4 ± 0.4	74.1 ± 2.3	23.1 ± 1.5

We also conducted experiments using other sensitive attributes, such as “Heavy Makeup” and “Wearing Lipstick”. The results, presented in Table 5, are consistent with those in Table 4, demonstrating the robustness of our proposed method.

The analysis of Table 6, which presents the performance and fairness metrics of models on the MultiNLI dataset, follows a similar structure to that of Table 1. The general observations about model performance and the impact of incorporating fairness learning strategies (such as FL Min., FL Maj., and FL All) are consistent with the results discussed for the CelebA dataset.

In summary, incorporating the **FL Min.** strategy across all models for the MultiNLI dataset leads to similar improvements as observed with the CelebA dataset. The WGA increases, and the fairness disparities (as indicated by DP, EOp, and EOD) are reduced. These results emphasize that focusing on disadvantaged groups during model training enhances both the performance for those groups and overall fairness.

Table 6: Performance and Fairness Metrics of Models on the MultiNLI Dataset

Model	ACC↑(%)	BA↑(%)	WGA↑(%)	DP↓(%)	EOp↓(%)	EOD↓(%)	AER↑(%)
ERM	82.6 ± 0.3	82.6 ± 0.3	67.3 ± 2.6	47.6 ± 1.2	14.6 ± 1.1	12.5 ± 1.5	57.1 ± 4.0
+ FL Min.	82.7 ± 0.4	82.7 ± 0.4	71.0 ± 1.5	45.5 ± 0.7	12.2 ± 1.0	10.8 ± 1.4	60.2 ± 3.8
+ FL Maj.	82.8 ± 0.2	82.8 ± 0.2	66.8 ± 2.7	47.7 ± 1.4	14.7 ± 1.2	12.7 ± 1.5	55.6 ± 4.1
+ FL All	82.8 ± 0.2	82.8 ± 0.2	70.5 ± 2.2	45.8 ± 1.1	12.5 ± 1.1	11.0 ± 1.0	59.0 ± 4.0
GroupDRO	80.8 ± 0.6	80.8 ± 0.3	77.2 ± 1.2	40.7 ± 0.4	8.8 ± 0.7	5.9 ± 0.9	74.8 ± 6.5
+ FL Min.	80.7 ± 0.8	80.7 ± 0.8	78.3 ± 1.4	39.6 ± 0.7	7.5 ± 0.6	5.5 ± 0.8	77.2 ± 7.1
+ FL Maj.	81.2 ± 0.5	81.2 ± 0.5	75.0 ± 2.9	42.5 ± 0.6	9.1 ± 0.7	6.0 ± 1.2	72.5 ± 5.6
+ FL All	81.2 ± 0.4	81.2 ± 0.4	76.8 ± 1.0	41.6 ± 0.9	8.3 ± 0.8	5.7 ± 0.9	74.9 ± 6.7
DFR	81.9 ± 0.4	81.9 ± 0.4	74.1 ± 1.0	43.1 ± 0.5	9.1 ± 0.7	6.7 ± 0.8	65.1 ± 5.2
+ FL Min.	81.9 ± 0.3	81.9 ± 0.3	76.0 ± 1.0	42.0 ± 0.4	8.0 ± 0.6	6.3 ± 0.7	67.3 ± 5.4
+ FL Maj.	82.1 ± 0.7	82.1 ± 0.7	73.0 ± 2.1	43.9 ± 0.8	9.0 ± 1.0	6.8 ± 0.9	63.4 ± 7.1
+ FL All	82.1 ± 0.5	82.1 ± 0.5	74.7 ± 1.5	42.9 ± 0.5	8.5 ± 0.7	6.6 ± 0.7	66.0 ± 6.0
Lu et al. (2024)	82.0 ± 0.2	82.0 ± 0.2	72.8 ± 0.7	44.7 ± 0.9	10.1 ± 0.6	8.3 ± 0.6	64.7 ± 5.1
+ FL Min.	82.0 ± 0.2	82.0 ± 0.2	75.0 ± 0.6	42.6 ± 0.8	9.0 ± 0.5	7.5 ± 0.6	66.9 ± 5.2
+ FL Maj.	82.5 ± 0.4	82.5 ± 0.4	71.8 ± 1.5	44.8 ± 1.2	10.7 ± 1.2	8.4 ± 1.0	62.7 ± 6.7
+ FL All	82.6 ± 0.1	82.6 ± 0.1	74.7 ± 0.6	43.1 ± 0.9	9.1 ± 0.6	7.7 ± 0.6	66.3 ± 5.0

C.2 CALCULATION OF CORRELATION COEFFICIENTS

To verify that mitigating a new bias does not interfere with previously achieved fairness improvements, we calculated the Pearson correlation coefficients between performance changes across debiasing stages. Specifically, we examined the changes in metrics unrelated to gender bias after mitigating gender bias, relative to the original ERM model. The following metrics were used for each model: **DP (R)** (racial fairness), **EOp (R)**, **EOD (R)** and **ACC** (accuracy).

1. Extract Metrics and Compute Changes

The metrics were extracted from Table 7. For each metric M , we calculated the change ΔM at each debiasing stage relative to the ERM baseline.

For DistilBERT-base, the changes are:

- **Changes at FLoRa Afr. stage:** $\Delta DP (R)_{Afr.} = 33.7 - 38.2 = -4.5$, $\Delta EOp (R)_{Afr.} = 14.2 - 14.9 = -0.7$, $\Delta EOD (R)_{Afr.} = 24.4 - 26.5 = -2.1$, $\Delta ACC_{Afr.} = 79.6 - 79.5 = +0.1$.
- **Changes at FLoRa Fe. stage:** $\Delta DP (R)_{Fe.} = 32.8 - 38.2 = -5.4$, $\Delta EOp (R)_{Fe.} = 13.1 - 14.9 = -1.8$, $\Delta EOD (R)_{Fe.} = 23.0 - 26.5 = -3.5$, $\Delta ACC_{Fe.} = 79.7 - 79.5 = +0.2$.

2. Form Vectors of Changes

We form vectors of the changes for the two debiasing stages: $\mathbf{X} = [-4.5, -0.7, -2.1, +0.1]$ (FLoRa Afr.), $\mathbf{Y} = [-5.4, -1.8, -3.5, +0.2]$ (FLoRa Fe.).

3. Compute Correlation Coefficient

The Pearson correlation coefficient r between the vectors \mathbf{X} and \mathbf{Y} was calculated. For DistilBERT-base, the resulting correlation coefficient is:

$$r = 0.97$$

4. Results for BERT-base Model

Similarly, for the BERT-base model, we calculated:

- **Changes at FLoRa Afr. and FLoRa Fe. stages:** $\mathbf{X} = [-13.1, -4.6, -8.9, -0.2]$ (FLoRa Afr.), $\mathbf{Y} = [-14.7, -5.8, -10.3, -0.1]$ (FLoRa Fe.).
- **Correlation Coefficient:** $r = 0.99$.

Table 7: Performance and fairness comparison during progressive debiasing of sensitive attributes for DistilBERT-base and BERT-base.

Metric	DistilBERT-base			BERT-base		
	ERM	FLoRa Afr.	FLoRa Fe.	ERM	FLoRa Afr.	FLoRa Fe.
Other TPR	75.2 \pm 2.0	75.1 \pm 1.9	77.0 \pm 1.7	77.1 \pm 1.7	77.0 \pm 1.8	78.2 \pm 1.6
Afr. TPR	90.1 \pm 1.7	90.1 \pm 1.7	90.1 \pm 1.5	90.1 \pm 1.1	85.4 \pm 2.0	85.4 \pm 2.2
Other FPR	18.9 \pm 3.1	18.7 \pm 2.7	19.6 \pm 2.5	20.5 \pm 4.0	19.3 \pm 3.7	20.9 \pm 4.1
Afr. FPR	57.1 \pm 2.7	52.4 \pm 2.9	52.4 \pm 2.0	47.6 \pm 4.4	33.3 \pm 3.6	33.3 \pm 2.9
DP (R) \downarrow	38.2 \pm 1.4	33.7 \pm 1.4	32.8 \pm 1.1	27.1 \pm 0.9	14.0 \pm 1.0	12.4 \pm 0.7
EOP (R) \downarrow	14.9 \pm 1.1	14.2 \pm 1.0	13.1 \pm 1.0	13.0 \pm 0.8	8.4 \pm 1.1	7.2 \pm 1.0
EOD (R)\downarrow	26.5 \pm 0.7	<u>24.4 \pm 0.6</u>	23.0 \pm 0.6	20.1 \pm 0.4	<u>11.2 \pm 0.6</u>	9.8 \pm 0.5
Male TPR	80.5 \pm 2.1	80.5 \pm 2.1	80.6 \pm 2.0	82.5 \pm 2.1	81.0 \pm 1.8	81.0 \pm 1.7
Fe. TPR	67.5 \pm 2.9	67.5 \pm 3.1	78.6 \pm 2.8	64.3 \pm 2.0	64.3 \pm 2.0	74.2 \pm 1.8
Male FPR	19.7 \pm 2.1	19.3 \pm 1.7	20.1 \pm 2.1	21.0 \pm 3.8	20.1 \pm 3.6	21.0 \pm 3.7
Fe. FPR	28.6 \pm 3.0	28.6 \pm 3.1	33.0 \pm 2.9	28.6 \pm 2.6	28.6 \pm 2.6	28.6 \pm 2.1
DP (G) \downarrow	7.4 \pm 1.3	7.6 \pm 1.1	12.9 \pm 2.2	7.6 \pm 1.5	8.5 \pm 1.4	7.6 \pm 1.0
EOP (G) \downarrow	13.0 \pm 0.5	13.0 \pm 0.5	2.0 \pm 2.1	18.2 \pm 0.8	16.7 \pm 0.4	8.8 \pm 1.4
EOD (G)\downarrow	11.3 \pm 1.1	<u>11.2 \pm 0.7</u>	7.4 \pm 0.6	12.9 \pm 1.1	<u>12.6 \pm 0.9</u>	8.2 \pm 0.4
ACC\uparrow	79.5 \pm 0.2	<u>79.6 \pm 0.2</u>	79.7 \pm 0.3	79.8 \pm 0.3	79.6 \pm 0.5	<u>79.7 \pm 0.4</u>

* Bold values indicate the best performance in each category, while underlined values represent the second-best results. “R” refers to Race, and “G” refers to Gender.

5. Summary

The high correlation coefficients (0.97 for DistilBERT and 0.99 for BERT) indicate a strong positive relationship between the changes in metrics across debiasing stages, demonstrating that mitigating a new bias does not adversely affect previously achieved improvements, effectively preventing catastrophic forgetting.

C.3 EXPLORING THE IMPACT OF PROCESSING ORDER ON MULTI-SENSITIVE ATTRIBUTES

Table 8: Performance and fairness comparison during progressive debiasing of sensitive attributes for DistilBERT-base and BERT-base.

Metric	DistilBERT-base			BERT-base		
	ERM	FLoRa Fe.	FLoRa Afr.	ERM	FLoRa Fe.	FLoRa Afr.
DP (R) \downarrow	38.2 \pm 1.4	37.8 \pm 1.2	32.9 \pm 1.2	27.1 \pm 0.9	26.7 \pm 0.9	12.1 \pm 1.0
EOP (R) \downarrow	14.9 \pm 1.1	14.7 \pm 1.1	13.2 \pm 1.1	13.0 \pm 0.8	12.4 \pm 1.2	7.0 \pm 1.1
EOD(R)\downarrow	26.5 \pm 0.7	<u>26.0 \pm 0.7</u>	23.1 \pm 0.7	20.1 \pm 0.4	<u>19.7 \pm 0.5</u>	9.6 \pm 0.7
DP (G) \downarrow	7.4 \pm 1.3	13.0 \pm 2.1	12.8 \pm 2.2	7.6 \pm 1.5	8.0 \pm 1.2	8.5 \pm 1.0
EOP (G) \downarrow	13.0 \pm 0.5	5.0 \pm 1.9	3.7 \pm 1.7	18.2 \pm 0.8	8.9 \pm 1.5	8.8 \pm 1.2
EOD(G)\downarrow	11.3 \pm 1.1	<u>7.3 \pm 0.7</u>	7.2 \pm 0.6	12.9 \pm 1.1	<u>8.4 \pm 0.7</u>	8.3 \pm 0.5
ACC\uparrow	79.5 \pm 0.2	79.6 \pm 0.3	79.6 \pm 0.2	<u>79.8 \pm 0.3</u>	<u>79.8 \pm 0.5</u>	79.9 \pm 0.4

* Bold values indicate the best performance in each category, while underlined values represent the second-best results. “R” refers to Race, and “G” refers to Gender.

We conducted additional experiments to investigate the impact of varying the sequence of debiasing (FairLoRA Race first) and addressing multiple biases simultaneously. As shown in Table 8, the results indicate that the order of debiasing has negligible impact on the final outcomes. This finding aligns with our theoretical explanation that FairLoRA exhibits a “forgetting-avoidance” property, whereby corrections for distinct sensitive attributes are encapsulated in independent LoRA modules.

This design ensures that adjustments made for one attribute do not interfere with those made for others.

Moreover, as illustrated in Table 9, the results demonstrate that whether biases are mitigated sequentially or simultaneously, the overall outcomes remain largely consistent. This robustness arises from FairLoRA’s modular architecture, which stores adjustments for each sensitive attribute in separate LoRA modules, allowing independent corrections without cross-attribute interference.

Table 9: Comparison of Progressive Debiasing and Simultaneous Debiasing Approaches.

Metric	DistilBERT-base			BERT-base		
	Afr.Fisrt	Fe.Fisrt	Together	Afr.Fisrt	Fe.Fisrt	Together
DP (R)↓	32.8 ± 1.1	32.9 ± 1.2	33.2 ± 1.2	12.4 ± 0.7	12.1 ± 1.0	12.8 ± 1.1
EOP (R)↓	13.1 ± 1.0	13.2 ± 1.1	13.3 ± 1.1	7.2 ± 1.0	7.0 ± 1.1	7.5 ± 1.2
EOD(R)↓	23.0 ± 0.6	23.1 ± 0.7	23.3 ± 0.8	9.8 ± 0.5	9.6 ± 0.7	10.0 ± 0.7
DP (G)↓	12.9 ± 2.2	12.8 ± 2.2	13.1 ± 2.3	7.6 ± 1.0	8.5 ± 1.0	8.0 ± 1.2
EOP (G)↓	2.0 ± 2.1	3.7 ± 1.7	4.7 ± 2.2	8.8 ± 1.4	8.8 ± 1.2	9.0 ± 1.4
EOD(G)↓	7.4 ± 0.6	7.2 ± 0.6	7.4 ± 0.8	8.2 ± 0.4	8.3 ± 0.5	8.5 ± 0.7
ACC↑	79.7 ± 0.3	79.6 ± 0.2	79.6 ± 0.3	79.7 ± 0.4	79.9 ± 0.4	79.7 ± 0.4

* Afr.First refers to applying FairLoRA to address bias for African Americans first, while Fe.First refers to addressing bias for females first, and Together represents simultaneous bias mitigation for both groups.

D IMPACT OF THRESHOLD ON DISCRIMINATOR TPR AND FPR FOR DEMOGRAPHIC GROUPS

African American Group Analysis (Left Pair of Plots in Figure 3) The top-left plot illustrates the variation of True Positive Rate (TPR) and False Positive Rate (FPR) for the “African American” group as a function of the threshold. As the threshold increases, both TPR and FPR decrease. The reduction in TPR suggests that a higher threshold leads to stricter classification, reducing the number of true positives. Meanwhile, the rapid decrease in FPR indicates fewer false positives.

The bottom-left plot shows the TPR/FPR ratio across different thresholds. This ratio peaks at approximately 0.7-0.8, indicating an optimal balance between TPR and FPR. Beyond this peak, the ratio declines, suggesting diminishing benefits from further increasing the threshold due to a disproportionate reduction in TPR compared to the decline in FPR. Therefore, this peak threshold can be used to guide optimal threshold selection, ensuring fairness and maintaining model performance.

Female Group Analysis (Right Pair of Plots in Figure 3) The top-right plot shows the changes in TPR and FPR for the “Female” group, following a similar pattern to the “African American” group. As the threshold increases, both TPR and FPR decrease, with higher thresholds making the model stricter, leading to a reduction in both true positives and false positives.

The bottom-right plot depicts the TPR/FPR ratio, which also peaks around the 0.7-0.8 threshold range, indicating the threshold range that maximizes classification efficiency for the “Female” group. After this peak, the ratio starts to decline, suggesting that further increases in the threshold reduce classification effectiveness. Thus, selecting a threshold near this peak ensures optimal fairness while retaining classification accuracy.

Summary For both the “African American” and “Female” groups in the HateXplain dataset, the TPR/FPR ratio reaches its peak around a threshold of 0.7-0.8, indicating that this range provides the optimal balance between fairness and classification performance. For other datasets, a similar analysis can be conducted to determine the optimal threshold range that ensures FairLoRA effectively mitigates biases while maintaining overall model efficacy.

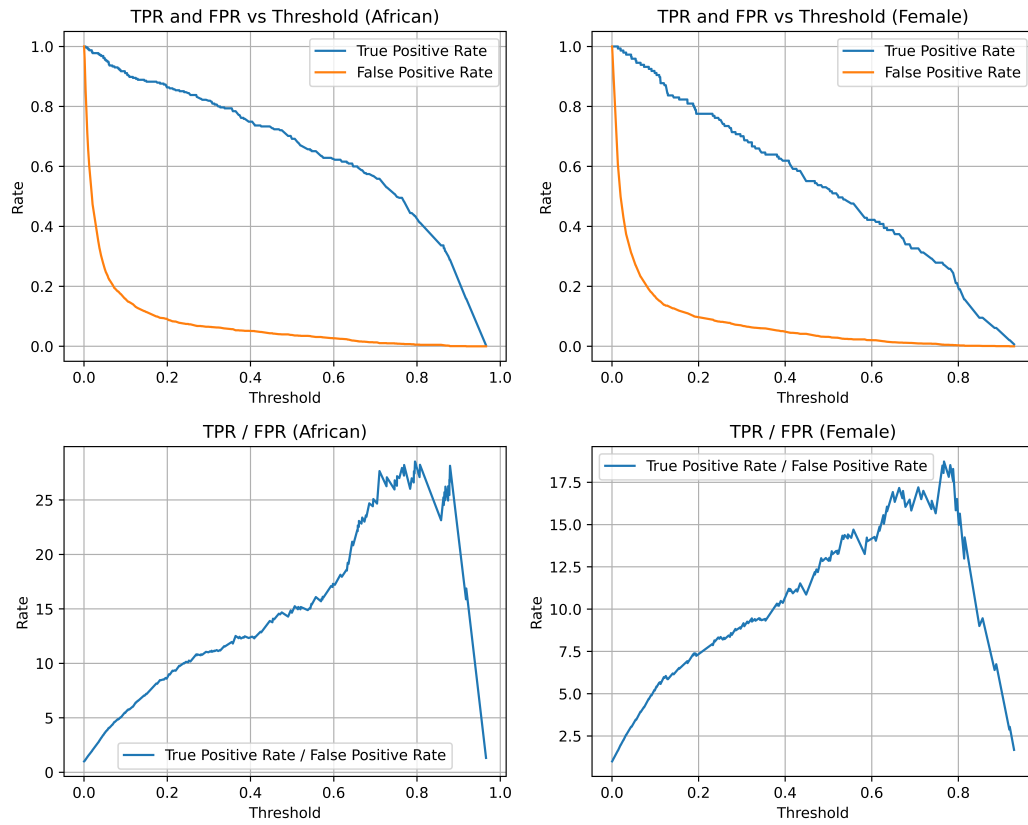


Figure 3: TPR and FPR Analysis with TPR/FPR Ratio for African American and Female Groups across Different Thresholds.