BRIDGING FAIRNESS AND EXPLAINABILITY: CAN INPUT-BASED EXPLANATIONS PROMOTE FAIRNESS IN HATE SPEECH DETECTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural language processing (NLP) models often replicate or amplify social bias from training data, raising concerns about fairness. At the same time, their black-box nature makes it difficult for users to recognize biased predictions and for developers to effectively mitigate them. While some studies suggest that input-based explanations can help detect and mitigate bias, others question their reliability in ensuring fairness. Existing research on explainability in fair NLP has been predominantly qualitative, with limited large-scale quantitative analysis. In this work, we conduct the first systematic study of the relationship between explainability and fairness in hate speech detection, focusing on both encoder- and decoder-only models. We examine three key dimensions: (1) identifying biased predictions, (2) selecting fair models, and (3) mitigating bias during model training. Our findings show that input-based explanations can effectively detect biased predictions and serve as useful supervision for reducing bias during training, but they are unreliable for selecting fair models among candidates.

1 INTRODUCTION

Language models (LMs) pre-trained on large-scale natural language datasets have shown great capacities in various NLP tasks (Wang et al., 2018; Gao et al., 2023). However, previous studies have shown that they can replicate and amplify stereotypes and social bias present in their training data and demonstrate biased behaviors (Sheng et al., 2021; Gupta et al., 2024; Gallegos et al., 2024). Such behaviors risk the underrepresentation of marginalized groups and the unfair allocation of resources, raising serious concerns in critical applications (Blodgett et al., 2020).

Meanwhile, current NLP models are mostly based on black-box neural networks. Despite their strong capacities, the complex architecture and large number of parameters of these models make it hard for humans to understand their behaviors (Bommasani et al., 2021). To understand neural NLP models, different types of explanations have been devised, such as input-based explanations (Yin & Neubig, 2022; Deiseroth et al., 2023; Madsen et al., 2024; Wang et al., 2025b), natural language explanations (Ramnath et al., 2024; Wang et al., 2025a), and concept-based explanations (Yu et al., 2024; Raman et al., 2024). Among these, input-based explanations, often referred to as rationales, indicate the contribution of each token to models' predictions, and thus provide the most direct insights into models' behaviors (Arras et al., 2019; Atanasova et al., 2022; Lyu et al., 2024).

Explainability has long been deemed critical to improving fairness. Researchers believe that if the use of sensitive features is evidenced by model explanations, then they can easily detect biased predictions and impose fairness constraints by guiding models to avoid such faulty reasoning (Meng et al., 2022; Sogancioglu et al., 2023). However, recent studies have challenged this assumption, suggesting that the relationship between explainability and fairness is complex and that explanations may not always reliably detect or mitigate bias (Dimanov et al., 2020; Slack et al., 2020; Pruthi et al., 2020). Unfortunately, to the best of our knowledge, current studies are mostly limited to qualitative analysis on a small set of explanation methods (Balkir et al., 2022; Deck et al., 2024). Our work takes a step toward bridging explainability and fairness by providing the first comprehensive quantitative analysis in the context of hate speech detection, a task where both fairness and explainability are

particularly critical. Specifically, we address the following three research questions to investigate the role of explainability in promoting fairness within the task of hate speech detection:

- **RQ1: Can input-based explanations be used to identify biased predictions?**
- **RQ2: Can input-based explanations be used to automatically select fair models?**
- **RQ3: Can input-based explanations be used to mitigate bias during model training?**

Our results show that input-based explanations can effectively detect biased predictions and help reduce bias during model training, but they are less reliable for automatic fair model selection. Furthermore, our analyses indicate that explanation-based bias detection remains robust even when models are trained to reduce reliance on sensitive features, and that these explanations outperform LLM judgments in identifying bias.

2 Related Work

Bias in NLP The presence of social bias and stereotypes has significantly shaped human language and LMs trained on it (Blodgett et al., 2020; Sheng et al., 2021). As a result, these models often exhibit biased behaviors (Gallegos et al., 2024), such as stereotypical geographical relations in the embedding space (Bolukbasi et al., 2016; May et al., 2019) and stereotypical associations between social groups and certain concepts in the model outputs (Fang et al., 2024; Wan & Chang, 2025). More critically, disparities in model predictions and performance across social groups (Zhao et al., 2018; Sheng et al., 2019) can significantly compromise user experiences of marginalized groups and risk amplifying bias against them, therefore drawing great concerns in critical use cases.

Input-based Model Explanations Input-based explanations in NLP models aim to attribute model predictions to each input token (Lyu et al., 2024). They can be broadly categorized based on how they generate explanations: gradient-based (Simonyan et al., 2014; Kindermans et al., 2016; Sundararajan et al., 2017; Enguehard, 2023), propagation-based (Bach et al., 2015; Shrikumar et al., 2017; Ferrando et al., 2022; Modarressi et al., 2022; 2023), perturbation-based (Li et al., 2016; Ribeiro et al., 2016; Lundberg & Lee, 2017; Deiseroth et al., 2023), and attention-based methods (Bahdanau et al., 2015; Abnar & Zuidema, 2020). While most prior work has focused on encoder-only models, recent studies have also explored explaining the behaviors of generative models (Yin & Neubig, 2022; Ferrando et al., 2022; Enouen et al., 2024; Cohen-Wang et al., 2024).

Bridging Explainability and Fairness Explainability is often considered essential for achieving fairness in machine learning systems (Balkir et al., 2022; Deck et al., 2024). One line of research investigates model bias by analyzing explanations (Prabhakaran et al., 2019; Jeyaraj & Delany, 2024; Sogancioglu et al., 2023). For instance, Muntasir & Noor (2025) shows that a biased model relied on gendered words as key features in its predictions, as revealed by LIME explanations. Similarly, Stevens et al. (2020) demonstrates that biased models often place high importance on gender and race features when examined with SHAP explanations. Extending this line of evidence, Meng et al. (2022) finds that features with higher importance scores are associated with larger disparities in model performance on a synthetic medical dataset using deep learning models.

Another line of research focuses on mitigating bias with explanations (Dimanov et al., 2020; Kennedy et al., 2020; Rao et al., 2023). For example, Hickey et al. (2020) improves fairness by reducing reliance on sensitive features during training with SHAP explanations. Bhargava et al. (2020) and González-Silot et al. (2025) first identify predictive sensitive features using LIME and SHAP, respectively, and then remove them prior to model training. In a related approach, Grabowicz et al. (2022) traces unfairness metrics back to input features and adjusts them to mitigate bias.

However, recent research has challenged the assumption that input-based explanations can be reliably used to detect and mitigate bias. First, current explanation methods may be unfaithful, meaning that they may not always reflect the true decision-making process of models (Kindermans et al., 2016; Jain & Wallace, 2019; Ye et al., 2025). This makes it difficult to reliably detect the use of sensitive features in predictions. Second, efforts to reduce the influence of sensitive features can lead to unintended consequences, sometimes degrading both model performance and fairness (Dimanov et al., 2020). Finally, models can be deliberately trained to assign lower importance to sensitive fea-

tures, thereby masking biased predictions when explanations are inspected (Dimanov et al., 2020; Slack et al., 2020; Pruthi et al., 2020).

Despite growing interest in this topic, most existing work remains qualitative or restricted to limited setups (Balkir et al., 2022; Deck et al., 2024). To the best of our knowledge, this is the first study to quantitatively and comprehensively examine the relationship between explainability and fairness in NLP models. We focus on hate speech detection as a particularly critical application. Prior research has shown that biased NLP models often rely on demographic information such as race and gender, leading to inferior performance on marginalized groups in this task (Sap et al., 2019; Mathew et al., 2021). Detecting and mitigating such biased behaviors are therefore essential to ensuring equitable opportunities for all social groups to voice their perspectives on social media.

3 Experimental Setup

Notations Let an input text $\mathbf{x}$ consist of tokens $t_1, t_2, \ldots, t_n$. The task of hate speech detection is to predict a binary label $\hat{y} \in \{\text{toxic}, \text{non-toxic}\}$. A classifier outputs the probability of class c as $f_c(x)$, where f is implemented by a neural model.

In the context of social bias, we assume that a bias type (e.g., race) involves a set of social groups G (e.g., black, white, ...). A subset of tokens $t_{g_1}, t_{g_2}, \ldots, t_{g_m}$ in $\mathbf{x}$ denotes the sensitive feature $g \in G$ of the speaker or target. We refer to these tokens as *sensitive tokens*. By replacing the sensitive tokens of group g with those of another group g', we obtain a counterfactual version of $\mathbf{x}$ that refers to g', denoted as $\mathbf{x}^{(g')}$.

An input-based explanation assigns an attribution score to each token in $\mathbf{x}$ for class c: $a_1^c, a_2^c, \ldots, a_n^c$, indicating their contribution to the prediction of class c. The attribution scores on the sensitive tokens, $a_{g_1}^c, a_{g_2}^c, \ldots, a_{g_m}^c$, are referred to as *sensitive token reliance* scores. If multiple sensitive tokens occur in a sentence, we take the score with the maximum absolute value as the reliance score for that example:[1]

$$\text{sensitive token reliance}(\mathbf{x}, c) = a_{j^*}^c, \text{where } j^* = \underset{j \in \{g_1, \ldots, g_m\}}{\arg\max} \left| a_j^c \right|$$

Datasets and Vocabulary We use two hate speech detection datasets: Civil Comments (Borkan et al., 2019) and Jigsaw (cjadams et al., 2019). To ensure coverage, we focus on three bias types and their associated groups: race (black / white), gender (female / male), and religion (Christian / Muslim / Jewish). We include examples containing identity-marking terms but exclude those with derogatory or slur-based references, as the latter can reasonably serve as direct evidence for toxic predictions. The sensitive token vocabulary is derived from Caliskan et al. (2017) and Wang & Demberg (2024). Further details on dataset pre-processing are provided in Appendix A.

Models We evaluate two types of commonly used NLP models: encoder-only models (BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)) and decoder-only large language models (Llama3.2-3B-Instruct (Dubey et al., 2024) and Qwen3-4B (Yang et al., 2025)). We fine-tune encoder-only models on data subsets that either target a single bias type or combine all bias types. For decoder-only models, we use an instruction-based setup where the model is prompted to decide whether a test example contains hate speech. The prompt includes the definition of hate speech, the test example, and a corresponding question. As a baseline, we adopt the zero-shot setting as the default configuration.

Beyond conventional fine-tuning and prompting, we also investigate the interaction between explainability and fairness in debiased models. For encoder-only models, we apply pre-processing techniques such as group balance (Kamiran & Calders, 2012), group-class balance (Dixon et al., 2018), and counterfactual data augmentation (CDA, Zmigrod et al., 2019), as well as in-processing techniques including dropout (Webster et al., 2020), attention entropy (Attanasio et al., 2022), and causal debias (Zhou et al., 2023). For decoder-only models, we incorporate bias reduction through prompt design, including few-shot, fairness imagination (Chen et al., 2025), and fairness instruction prompting (Chen et al., 2025). Further details are provided in Appendix A.

[1] We have also experimented with normalizing feature importance scores but found that using raw scores yielded the best results.

Fairness Metrics We evaluate fairness in model predictions using two categories of metrics: **group fairness** and **individual fairness**. Group fairness metrics capture disparities in performance across demographic groups:

$$\text{Disp}_{\text{metric}} = \sum_{g \in G} |\text{metric}_g - \overline{\text{metric}_G}|,$$

where $\overline{\text{metric}_G}$ is the average metric value across all groups G in a bias type. We specifically measure disparities in accuracy (ACC), false positive rate (FPR), and false negative rate (FNR).

Individual fairness measures the extent to which a model's prediction for a given example changes when the associated social group is altered. To maintain consistency with the direction of group fairness metrics, we compute the individual unfairness (IU) score of $\mathbf{x}_i$ and the predicted class $\hat{y}_i$:

$$\text{IU}(\mathbf{x}_i) = |f_{\hat{y}_i}(\mathbf{x}_i) - \frac{1}{|G \backslash \{g_i\}|} \sum_{g' \in G \backslash \{g_i\}} f_{\hat{y}_i}(\mathbf{x}_i^{(g')})|$$

The Average IU score (Avg_{iu}) is then computed over a dataset to reflect the overall level of individual unfairness in a model.

For both types of metrics, higher scores indicate more bias in model predictions. It is worth noting that individual unfairness can be evaluated at the level of each example, whereas group fairness metrics are defined over sets of validation or test examples. To compute the fairness metrics, we randomly sample a subset of examples for each bias type such that each social group contributes an equal number of examples. Further details on test set sampling are provided in Appendix A.

Explanation Methods We employ 14 variants of commonly used input-based post-hoc explanation methods, covering diverse methodological categories: Attention (Bahdanau et al., 2015), Attention rollout (Attn rollout, Abnar & Zuidema, 2020), Attention flow (Attn flow, Abnar & Zuidema, 2020), Gradient (Grad, Simonyan et al., 2014), Input x Gradient (IxG, Kindermans et al., 2016), Integrated Gradients (IntGrad, Sundararajan et al., 2017), Occlusion (Li et al., 2016), DeepLift (Shrikumar et al., 2017), and KernelSHAP (Lundberg & Lee, 2017). For methods that attribute predictions to embeddings, we aggregate attribution scores into a single feature importance value using either the mean or the L2 norm. For Occlusion, we additionally report results obtained by taking the absolute value of each attribution score prior to computing sensitive token reliance scores (denoted as Occlusion abs). We also study rationales generated by LLMs and find that these rationales are not as reliable as input-based explanations in detecting bias (Section 6).

Table 1: Task performance and fairness of default and debiased models on Civil Comments. Results are provided for race / gender / religion biases. Green (red) indicates the results are better (worse) than the default / zero-shot models. No debiasing method consistently reduces bias across all metrics and bias types.

Model	Method	Accuracy (↑)	Disp_{acc}(↓)	Disp_{fpr}(↓)	Disp_{fnr}(↓)	Avg_{iu}(↓)
BERT	Default	78.38/88.05/85.93	2.05/3.30/18.07	0.50/0.03/5.77	10.04/11.98/30.9	3.17/0.66/1.27
	Group balance	79.25/87.25/86.83	3.10/2.80/13.53	0.25/1.73/11.53	10.46/5.38/30.31	3.79/0.42/2.01
	Group-class balancing	78.00/87.02/85.77	1.80/2.75/14.73	2.42/0.99/3.09	10.63/7.26/33.14	4.43/0.98/0.71
	CDA	76.83/86.70/84.83	2.35/3.60/14.13	5.88/2.00/5.67	18.45/7.57/24.12	0.50/0.50/0.90
	Dropout	78.53/88.20/85.03	2.25/2.10/15.67	0.78/1.46/5.93	10.82/3.50/27.16	3.43/0.52/1.51
	Attention entropy	79.15/87.67/84.93	2.60/2.05/15.07	0.99/0.10/4.99	11.71/7.11/26.52	2.95/0.67/1.58
	Causal debias	78.80/86.17/86.40	0.00/2.65/16.40	3.90/0.46/8.82	7.98/10.67/30.46	3.83/0.48/2.10
Qwen3	Zero-shot	69.55/79.75/77.50	0.60/0.00/17.40	7.13/1.40/21.07	13.25/3.71/5.17	2.55/2.41/3.32
	Few-shot	63.00/76.40/74.30	1.90/1.60/20.80	10.17/4.77/26.67	8.04/7.21/8.22	3.30/4.03/4.60
	Fairness imagination	71.23/80.40/80.83	0.85/1.00/18.27	4.03/2.11/10.51	11.62/9.21/4.28	2.98/3.16/2.20
	Fairness instruction	70.40/79.77/80.47	0.60/1.35/19.33	4.30/0.39/4.67	11.11/5.24/5.08	2.02/1.83/1.71

4 Quantitative Analyses of Fairness and Explainability

To comprehensively understand the relationship between explainability and fairness in NLP models, we examine three ways in which model explanations can be applied to promote fairness. The

subsequent sections detail the experimental setups for each application and report the corresponding results. For brevity, we report results on Civil Comments using BERT trained on single bias types and Qwen3. Results for additional models and the Jigsaw dataset are presented in Appendix B to E.

4.1 Model Performance and Fairness

As a prerequisite, we first summarize the performance and fairness of the evaluated models. The results in Table 1 show that no single debiasing method consistently improves all fairness metrics. For BERT and Qwen3, CDA and fairness instruction achieve the largest reductions in individual unfairness, yet they may simultaneously amplify biases on other metrics. Other debiasing methods show a similar pattern: they reduce bias for a specific metric or bias type, but the improvement does not generalize across different setups. These limitations underscore the importance of leveraging explanations for bias detection and mitigation. We find similar results for other models and for Jigsaw, which we provide in Appendix B along with a discussion on model performance and fairness.

4.2 RQ1: Explanations For Bias Detection

Our first research question asks whether explanations can be used to detect biased predictions. We address the question through three steps: (1) obtain model predictions and compute individual unfairness scores; (2) generate input-based explanations for the predictions; and (3) compute sensitive token reliance scores and evaluate their Pearson correlation with individual unfairness, which we refer to as *fairness correlation*. A higher fairness correlation indicates that the explanation method is more effective in identifying predictions with high individual unfairness. To ensure robustness, we compute the fairness correlation separately for each prediction class–group pair and report the average absolute score as the final result for each explanation method.

We present results for default and debiased models where individual unfairness remains high after debiasing, as bias detection is particularly critical in these cases. Specifically, we report results for models with the highest average Avg_{iu} scores across bias types, namely default, group balance, and causal debias for BERT, and zero-shot, few-shot, and fairness imagination prompting for Qwen3. Results for religion as well as other models and the Jigsaw dataset are provided in Appendix C.

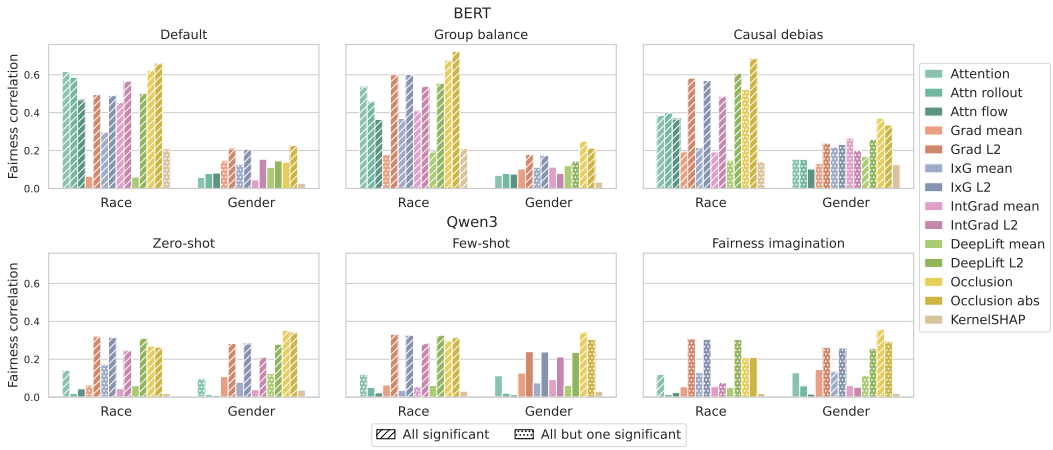

Figure 1: Fairness correlation results for each explanation method. Occlusion- and L2-based explanations are effective for bias detection across different bias types and models.

Results Figure 1 shows that the best-performing explanation methods, such as Grad L2, IxG L2, DeepLift L2, Occlusion, and Occlusion abs, generally achieve high fairness correlations across different models and bias types, indicating a strong ability to detect biased predictions. Besides, their fairness correlations are mostly statistically significant ($p < \alpha = 0.05$) in all, or in all but one, class-group categories, which confirms their reliability. Among these methods, Occlusion and Occlusion abs perform best with BERT models, whereas the L2-based methods Grad L2, IxG L2, and DeepLift L2 are most effective with Qwen3.

When comparing different variants of the same explanation family, mean-based approaches perform considerably worse than their L2-based counterparts, and also underperform compared to undirected attention-based methods. We attribute this limitation to their dependence on accurately determining the direction of each token's contribution, a challenge that attention- and L2-based explanations do not face. Our analysis in Appendix G further shows that the effectiveness of explanation-based bias detection is not determined by explanation faithfulness, underscoring the need for careful evaluation when selecting methods for bias identification.

Takeaway: Input-based explanation methods, particularly Occlusion- and L2-based ones, are effective for identifying biased predictions at inference time.

4.3 RQ2: Explanations for Model Selection

Given that explanations can detect biased predictions (RQ1), we next investigate whether they can also be used to select fair models among candidates. Prior work has demonstrated that input-based explanations on validation examples can help humans identify spurious correlations in models (Lertvittayakumjorn & Toni, 2021; Pezeshkpour et al., 2022). Extending this idea, we examine whether explanations can be leveraged for automatic fair model selection, thereby removing the need for human intervention.

Our experiments consist of three steps: (1) for all default and debiased models (seven encoder-only and four decoder-only), we generate predictions on a validation set and compute explanation-based metrics; (2) we compute fairness metrics on the test set for each model; and (3) we evaluate model selection ability using two measures: Spearman's rank correlation (ρ) between validation set explanation-based metrics and test set fairness metrics, which reflects the ability to rank models, and mean reciprocal rank of the fairest model (MRR@1), which reflects the ability to select the fairest model. Higher correlations and MRR@1 indicate that an explanation method is useful for ranking models and selecting the fairest one. Specifically, we use the average absolute sensitive token reliance on the validation set as the explanation-based metric to rank and select models based on average individual unfairness on the test set.[2] As a baseline, we report results of using the validation set average individual unfairness as the predictor of test set fairness performance. The results are averaged over six and three random validation set selections for encoder- and decoder-only models, respectively. Results for more models and the Jigsaw dataset are presented in Appendix D.

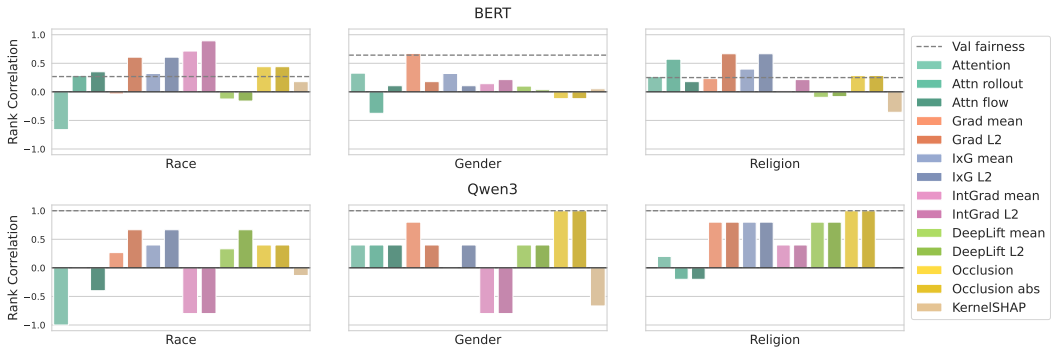

Figure 2: Rank correlations between validation set average absolute sensitive token reliance and test set individual unfairness. The validation set sizes are 500 for race and gender, and 200 for religion. None of the explanation methods consistently achieve performance on par with the baseline.

Results The results in Figures 2 and 3 highlight the limitations of using explanations for model selection. Although some methods occasionally show high rank correlations (e.g., Grad L2 for race and religion biases in BERT and Occlusion-based methods for gender and religion biases in Qwen3), none of them consistently reach the baseline of us-

[2]We have evaluated other metrics to predict group fairness outcomes. However, neither explanation-based metrics nor validation set fairness achieved rank correlations beyond random chance with the test set results.

ing the individual unfairness on the validation set. This limitation is particularly evident in decoder-only models, where the baseline achieves a perfect rank correlation of 1. Their ability to select the fairest models is even weaker, as indicated by lower MRR@1 scores compared to both the baseline and random ranking. Considering that these explanations are often more computationally expensive to generate than evaluating validation set fairness, they are not practically useful as a model fairness indicator. Therefore, we do not recommend explanation-based model selection, especially in decoder-only models. The difference in findings between RQ1 and RQ2 may stem from the fact that debiasing methods can alter model behaviors and thereby affect explanation attributions. As a result, comparing explanations across default and debiased models is less reliable, whereas comparing explanations within the same model remains effective for detecting biased predictions.

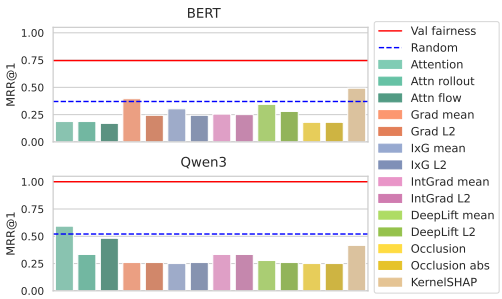

Figure 3: Average MRR@1 across bias types. Explanation methods perform worse than the baseline in identifying the fairest models.

Takeaway: Input-based explanation methods are not reliable tools for selecting fair models.

4.4 RQ3: Explanations for Bias Mitigation

Having shown that explanations can reliably reveal biased predictions (RQ1), we now investigate whether they can also be leveraged to mitigate model bias. Building on prior work demonstrating that explanation regularization can reduce spurious correlations while also improving performance and generalization (Kennedy et al., 2020; Rao et al., 2023), we investigate bias mitigation by minimizing sensitive token reliance during training. Following Dimanov et al. (2020), we define a debiasing regularization term, $\mathrm{L}_{\text{debias}}$, which penalizes the average sensitive token reliance of all such tokens in an input, in addition to the task loss:

$$\mathrm{L} = \mathrm{L}_{\text{task}} + \alpha \mathrm{L}_{\text{debias}}$$

Here, α is a hyperparameter that controls the strength of sensitive token reliance reduction. For embedding-level attributions, we apply either an L1 or L2 norm penalty, corresponding to minimizing mean- or L2-based reliance scores, respectively.

While Dimanov et al. (2020) tune hyperparameters based on task accuracy, we search $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ using a fairness-balanced metric (the harmonic mean of accuracy and 100–unfairness) on the validation set. Models are selected separately for each fairness criterion. Due to computational cost, we restrict training to single bias types. We exclude DeepLift and KernelSHAP, as they are not easily differentiable and thus cannot be incorporated into model training, and IntGrad, due to its substantial time and memory costs of generating explanations and tracking gradients. Results are averaged over three runs. More implementation details are provided in Appendix A.

Results In Figure 4, we present race and gender bias mitigation results. For consistency with accuracy, fairness results are reported as $100-\{\text{Disp}_{\text{acc}}, \text{Disp}_{\text{fpr}}, \text{Disp}_{\text{fnr}}, \text{Avg}_{\text{iu}}\}$, so that higher values indicate lower bias. We find that explanation-based bias mitigation effectively improves fairness across multiple metrics. Most notably, it consistently and substantially reduces Disp_{fnr} for all bias types. For gender bias, it also yields considerable reductions in Disp_{acc}, and Avg_{iu} is mitigated for race bias. Moreover, as shown in Figure 18, all group fairness disparity metrics decrease for religion bias. The bias mitigation effects are consistent across all models and are also observed on the Jigsaw dataset (see Figures 18, 19, 20, 21 in Appendix E).

At the same time, explanation-based debiasing maintains a good balance between fairness and accuracy. For example, Grad L1 both increases accuracy and reduces Disp_{acc}, Disp_{fnr}, and Avg_{iu} for gender bias, while most other explanation methods also achieve better Disp_{acc} and Disp_{fnr} with marginal or no accuracy loss. Our harmonic fairness–accuracy mean results (Figures 22, 23, 24, 25) further confirm this by showing that explanation-based debiasing almost always achieves comparable or higher harmonic means than both default models and traditional debiasing methods.

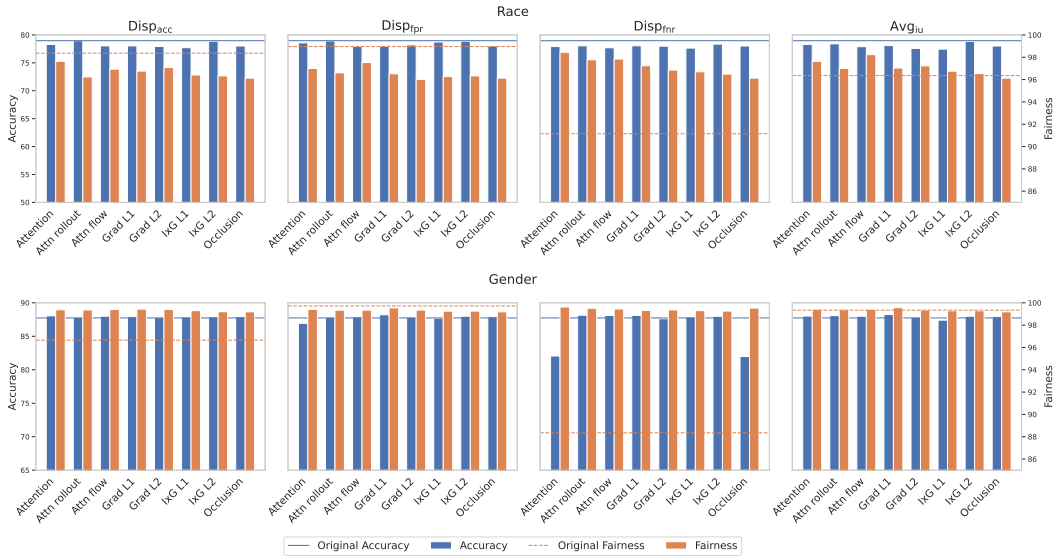

Figure 4: Accuracy and fairness results for bias mitigation using different explanation methods. Each column corresponds to models selected by maximizing the fairness-balanced metric with respect to the indicated bias metric. We find that explanation methods can improve fairness across many metrics while maintaining reasonable task accuracy.

Among individual explanation methods, attention and attn flow achieve strong debiasing performance on BERT, whereas Occlusion performs best on RoBERTa, though often at the cost of a larger drop in model accuracy. Overall, IxG L2 and attention-based methods provide robust debiasing while maintaining a favorable fairness–accuracy trade-off across bias types, models, and datasets, as reflected in the harmonic mean results. Our findings differ from those of Dimanov et al. (2020), which we attribute to our fairness-based hyperparameter tuning strategy.

Takeaway: Input-based explanations can provide effective supervision for mitigating model bias during training while maintaining a good fairness–performance trade-off. In particular, IxG L2 and attention-based methods achieve robust debiasing with strong overall balance.

5 Bias Detection in Explanation-Debiased Models

While explanation-based methods are effective in reducing bias (RQ3), their suppression of attributions on sensitive tokens could potentially mislead users into believing that model predictions are unbiased (Dimanov et al., 2020; Slack et al., 2020; Pruthi et al., 2020). To investigate this concern, we reapply the bias detection procedure from RQ1 to explanation-debiased models and compare their fairness correlations with those from the corresponding default models. For this analysis, we use the models debiased for race bias with respect to individual unfairness, as described in RQ3.

The fairness correlation differences from default models are shown in Figure 5. We observe that the impact of explanation-based debiasing on fairness correlations depends on both the explanations used for debiasing and those used for bias detection. Some approaches, such as Grad mean / L2, IxG L2, DeepLift mean / L2, Occlusion, and Occlusion abs, are only marginally, or even positively, affected by debiasing. Their fairness correlation scores (see Figure 26 in Appendix F) further indicate that Occlusion- and L2-based methods (except IntGrad L2) remain reliable for revealing bias in explanation-debiased models. In contrast, attention-based explanations experience substantial drops, particularly when the models themselves are debiased using attention-based methods. Similarly, IntGrad-based explanations show a reduced bias detection ability when the debiasing procedure is also gradient-based. Overall, these findings demonstrate that certain input-based explanations remain effective for detecting biased predictions even in explanation-debiased models. Our results are different from those of Dimanov et al. (2020), likely because their analysis focused solely on attribution magnitudes without considering their relationship to fairness metrics.

6 EXPLANATION-BASED BIAS DETECTION VS. LLM-AS-A-JUDGE

Existing research suggests that LLMs could identify and correct bias in their own outputs (Bai et al., 2022; Furniturewala et al., 2024). In this section, we compare the bias detection ability of input-based explanations against LLMs' own judgments under two paradigms: (1) self-reflection, where LLMs are asked to indicate whether their own predictions rely on bias or stereotypes, and (2) self-attribution, where LLMs choose a K-word rationale from the input, which we then examine for the presence of sensitive tokens. We conduct this analysis using Qwen3 on the same subset for race bias as in the main experiments (see Appendix A for the prompts used).

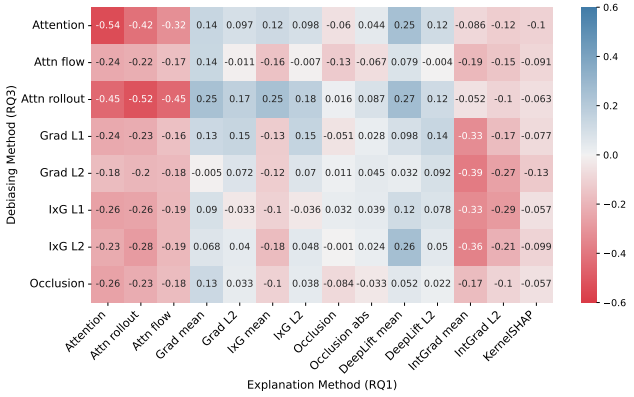

Figure 5: Fairness correlation differences between default and explanation-debiased BERT. Occlusion- and L2-based explanations (except IntGrad L2) are less affected by explanation-based debiasing and remain effective for bias detection.

Table 2 shows that self-reflection is highly conservative in flagging bias: only 86 out of 4,000 predictions are labeled as biased, all of which correspond to toxic predictions. The complete absence of non-toxic cases and the extremely low coverage make self-reflection unreliable as a bias detection method due to low recall. Furthermore, predictions marked as biased by the model actually show lower average individual unfairness than those marked non-biased, indicating poor precision as well.

In contrast, LLMs' self-attributions are less conservative, producing biased / unbiased judgments across both demographic groups and toxicity classes. However, while biased cases identified by rationales show higher average individual unfairness than unbiased ones, they still perform worse than a simple baseline that flags the top 50% of predictions ranked by absolute Grad L2 reliance scores (Grad L2 Binary). Therefore, we conclude that input-based explanations are more reliable than LLM-as-a-Judge for bias detection.

Table 2: Qwen3 results for detecting bias in its own predictions. "Biased / Unbiased" denotes whether an example is judged as biased or unbiased by the LLM through self-reflection or self-attribution. If the judgments are reliable, Avg_{iu} should be higher for biased examples than unbiased ones. For self-reflection, fairness correlation cannot be computed because the model labels no non-toxic predictions as biased. Input-based explanations reveal bias more reliably than LLM judgments.

	# Biased / Unbiased	Avg_{iu} (Biased / Unbiased)	Fairness Correlation
Self-reflection	86 / 3914	0.065 / 2.59	-
Self-attribution (K=5)	2063 / 1904	3.55 / 1.49	0.104
Self-attribution (K=10)	2176 / 1474	2.93 / 1.56	0.070
Grad L2 Binary	2000 / 2000	5.02 / 0.09	0.194

7 CONCLUSION

In this work, we present the first comprehensive study linking input-based explanations and fairness in hate speech detection. Our experiments show that (1) input-based explanations can effectively identify biased predictions, (2) they are not reliable for selecting fair models, and (3) they can serve as effective supervision signals during training, mitigating bias while preserving a strong balance between fairness and task performance. We further provide practical recommendations on which explanation methods are best suited for bias detection and bias mitigation. Finally, our analyses demonstrate that explanation-based bias detection remains effective in explanation-debiased models, and they outperforms LLM-as-a-Judge in identifying biased predictions.

8 Ethics Statement

This work investigates explainability and fairness in hate speech detection. Despite the diverse experimental setups considered, our findings remain limited by the coverage of tasks, models, datasets, fairness metrics, and identity terms; as such, results may not generalize across groups or domains and could be susceptible to adversarial attacks. We further caution that explanation methods and debiasing techniques cannot fully eliminate residual harms, and that LLM-generated bias judgments are unreliable for bias detection. We hope that our study will contribute to the development of NLP systems that are more transparent, reliable, and fair.

9 Reproducibility Statement

We include full implementation details in the main text and appendix, covering data pre-processing details, model architectures, training procedures, and hyperparameters. We have submitted our code and configuration files as supplementary material to facilitate reproduction during the review process. Upon acceptance, we will open-source our code and scripts for data pre-processing and experiments.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. In Tal Linzen, Grzegorz Chrupał a, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 113–126, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4813. URL https://aclanthology.org/W19-4813/.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10445–10453, 2022.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. Entropy-based attention regularization frees unintended bias mitigation from lists. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1105–1119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.88. URL https://aclanthology.org/2022.findings-acl.88/.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna

Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. Challenges in applying explainability methods to improve the fairness of NLP models. In Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, and Yang Trista Cao (eds.), *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pp. 80–92, Seattle, U.S.A., July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.trustnlp-1.8. URL https://aclanthology.org/2022.trustnlp-1.8/.

Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Limeout: An ensemble approach to improve process fairness. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 475–491. Springer, 2020.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

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. *Advances in neural information processing systems*, 29, 2016.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. Causally testing gender bias in LLMs: A case study on occupational bias. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4984–5004, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.281. URL https://aclanthology.org/2025.findings-naacl.281/.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. Jigsaw unintended bias in toxicity classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification, 2019. Kaggle.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95764–95807. Curran Associates, Inc.,

2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/adbea136219b64db96a9941e4249a857-Paper-Conference.pdf.

Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. A critical survey on fairness benefits of explainable ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1579–1595, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658990. URL https://doi.org/10.1145/3630106.3658990.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Atman: Understanding transformer predictions through memory efficient attention manipulation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63437–63460. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c83bc020a020cdeb966ed10804619664-Paper-Conference.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*, pp. 2473–2480. IOS Press, 2020.

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*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7555–7565, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.477. URL https://aclanthology.org/2023.findings-acl.477/.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13984–14011, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.832. URL https://aclanthology.org/2024.findings-acl.832/.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224, 2024.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8698–8714, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.595. URL https://aclanthology.org/2022.emnlp-main.595/.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 213–227, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.13. URL https://aclanthology.org/2024.emnlp-main.13/.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Santiago González-Silot, Andrés Montoro-Montarroso, Eugenio Martínez Cámara, and Juan Gómez-Romero. Enhancing disinformation detection with explainable AI and named entity replacement. *CoRR*, abs/2502.04863, 2025. doi: 10.48550/ARXIV.2502.04863. URL https://doi.org/10.48550/arXiv.2502.04863.

Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. Marrying fairness and explainability in supervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1905–1916, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533236. URL https://doi.org/10.1145/3531146.3533236.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J. Passonneau. CALM : A multi-task benchmark for comprehensive assessment of language model bias. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=RLFca3arx7.

James M Hickey, Pietro G Di Stefano, and Vlasios Vasileiou. Fairness by explicability and adversarial shap learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 174–190. Springer, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Manuela Jeyaraj and Sarah Delany. An explainable approach to understanding gender stereotype text. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (eds.), *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 45–59, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.gebnlp-1.4. URL https://aclanthology.org/2024.gebnlp-1.4/.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5435–5442, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.483. URL https://aclanthology.org/2020.acl-main.483/.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270, 2016. URL http://arxiv.org/abs/1611.07270.

Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021. doi: 10.1162/tacl_a_00440. URL https://aclanthology.org/2021.tacl-1.90/.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 1–10. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024. doi: 10.1162/coli_a_00511. URL https://aclanthology.org/2024.cl-2.6/.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.19. URL https://aclanthology.org/2024.findings-acl.19/.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, May 2021. doi: 10.1609/aaai.v35i17.17745. URL https://ojs.aaai.org/index.php/AAAI/article/view/17745.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 622–628, 2019.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 258–271, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.19. URL https://aclanthology.org/2022.naacl-main.19/.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. DecompX: Explaining transformers decisions by propagating token decomposition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2649–2664, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.149. URL https://aclanthology.org/2023.acl-long.149/.

Fahim Muntasir and Jannatun Noor. Explainable ai discloses gender bias in sexism detection algorithm. In *Proceedings of the 11th International Conference on Networking, Systems, and Security*, NSysS '24, pp. 120–127, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711589. doi: 10.1145/3704522.3704524. URL https://doi.org/10.1145/3704522.3704524.

Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron Wallace. Combining feature and instance attribution to detect artifacts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1934–1946, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.153. URL https://aclanthology.org/2022.findings-acl.153/.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5740–5745, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1578. URL https://aclanthology.org/D19-1578/.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4782–4793, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.432. URL https://aclanthology.org/2020.acl-main.432/.

Naveen Janaki Raman, Mateo Espinosa Zarlenga, and Mateja Jamnik. Understanding interconcept relationships in concept-based models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42009–42025. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/raman24a.html.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. Tailoring self-rationalizers with multi-reward distillation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=t8eO0CiZJV.

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1922–1933, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339/.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330/.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Gizem Sogancioglu, Heysem Kaya, and Albert Ali Salah. Using explainability for bias mitigation: A case study for fair recruitment assessment. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, pp. 631–639, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700552. doi: 10.1145/3577190.3614170. URL https://doi.org/10.1145/3577190.3614170.

Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen. Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1241–1248, 2020. doi: 10.1109/SSCI47803.2020.9308371.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Yixin Wan and Kai-Wei Chang. White men lead, black women help? benchmarking and mitigating language agency social biases in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9082–9108, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.445. URL https://aclanthology.org/2025.acl-long.445/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupał a, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. Cross-refine: Improving natural language explanation generation by learning in tandem. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1150–1167, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.77/.

Yifan Wang and Vera Demberg. A parameter-efficient multi-objective approach to mitigate stereotypical bias in language models. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (eds.), *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 1–19, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.gebnlp-1.1. URL https://aclanthology.org/2024.gebnlp-1.1/.

Yifan Wang, Sukrut Rao, Ji-Ung Lee, Mayank Jobanputra, and Vera Demberg. B-cos lm: Efficiently transforming pre-trained language models for improved explainability. *arXiv preprint arXiv:2502.12992*, 2025b.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032, 2020. URL https://arxiv.org/abs/2010.06032.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Mengyu Ye, Tatsuki Kuribayashi, Goro Kobayashi, and Jun Suzuki. Can input attributions explain inductive reasoning in in-context learning? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21199–21225, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1092. URL https://aclanthology.org/2025.findings-acl.1092/.

Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14. URL https://aclanthology.org/2022.emnlp-main.14/.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. Latent concept-based explanation of NLP models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.),

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 12435–12459, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.692. URL https://aclanthology.org/2024.emnlp-main.692/.

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)*, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4227–4241, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.232. URL https://aclanthology.org/2023.acl-long.232/.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161/.

A Details on Experimental Setups

Table 3: Splits for the Civil Comments and Jigsaw datasets used in this work. The instances are sampled randomly from the original dataset.

Split	Civil Comments			Jigsaw		
	race	gender	religion	race	gender	religion
Test	2,000	2,000	1,000	400	800	200
Training	8,000	8,000	6,300	8,000	8,000	6,300

Datasets and Models Table 3 shows the dataset statistics for the Civil Comments[3] and Jigsaw[4] datasets, respectively. When constructing subsets for each social group, we include only examples that contain words from the corresponding group's vocabulary and exclude those containing words from other groups. Additionally, we remove examples that include insulting references to social groups.

For encoder-only models, we either train separate models for each bias type or a single model on the combined set of all bias types. The latter are marked as (all) in the results. Models are trained for 5 epochs with a batch size of 8, a learning rate of 2×10^{-5}, and 10% warm-up steps. We implement debiasing methods of encoder-only models ourselves. Debiased models are trained on the same number of examples for comparability, therefore we do not sample more examples from external datasets for causal debias. All experiments are carried out using a maximum of 4x NVIDIA H100 GPUs.

We use the base version of BERT and RoBERTa models for our encoder-only experiments, and for decoder-only experiments, we adopt the instruct versions of the models from HuggingFace.

[3] https://huggingface.co/datasets/lighteval/civil_comments_helm

[4] https://huggingface.co/datasets/google/jigsaw_unintended_bias

Prompts We employ the following prompt for zero-shot and few-shot prompting with decoder-only models.

```
Task: Determine whether the following sentence contains any toxic
content. Toxicity includes insults, threats, hate speech, or any
language that is likely to offend or harm others.

([EXAMPLES])

Sentence: [TEST EXAMPLE]

Question: Does this sentence contain toxicity?

Answer (Yes or No):
```

We use one toxic and one non-toxic example for each social group associated with a bias type to construct the corresponding few-shot prompt. For fairness imagination, we prepend the task instruction with:

```
Imagine a world with no bias regarding [BIAS TYPE].
```

For fairness instruction, we add:

```
In the following, do not rely on stereotypes or biases regarding
[BIAS TYPE] to make predictions.
```

Input-Based Explanation Methods We use the Captum API[5] to generate explanations, except for Attention, Attn Rollout, and Attn Flow, which we implement ourselves.

Explanation-Based Bias Mitigation We reimplement all explanation methods used in RQ3 to enable gradient tracking during training and ensure consistency with the explanations provided by the Captum API. For Occlusion, we simplify computation by penalizing the probability difference between the original input and the input with all sensitive tokens masked.

Self-Reflection and Self-Attribution To elicit bias judgments from LLMs, we append the following instructions to their predictions.

Self-reflection:

```
Do not re-answer the task. Instead, reflect only on your previous
answer.

Question: Was your previous answer influenced by stereotypes or
biases related to [BIAS TYPE]?

Answer (Yes or No):
```

Self-attribution:

```
Now select exactly [NUM_TOKENS] words from the input sentence that
most influenced your decision. List the words separated by commas.
Please begin directly with the selected word.

Selected words:
```

B Model Performance and Fairness Results

Tables 4 and 5 show the task performance and fairness scores for the default / zero-shot and debiased models on the Civil Comments and Jigsaw datasets respectively. To better identify the differences

[5] https://captum.ai/api/

Table 4: Task performance and fairness of default and debiased models on the Civil Comments dataset. Results are provided for race / gender / religion biases. Green (red) indicates the results are better (worse) than the default / zero-shot models. *All* indicates the model is trained on data containing all bias types.

Model	Method	Accuracy (↑)	Disp_{acc} (↓)	Disp_{fpr} (↓)	Disp_{fnr} (↓)	Avg_{iu} (↓)
BERT (all)	Default	78.30/88.20/87.43	2.00/3.20/13.47	0.02/1.11/6.24	8.44/8.58/23.53	3.99/0.96/1.76
	Group balance	79.05/88.85/87.47	3.50/2.80/13.67	1.72/0.31/6.92	8.83/11.08/23.91	4.13/1.17/2.15
	Group-class balance	78.17/88.25/86.90	1.95/1.70/14.60	1.35/0.51/8.52	9.33/4.66/33.13	4.83/0.93/1.37
	CDA	78.08/87.70/86.83	2.65/2.70/14.33	6.38/1.05/4.70	20.35/6.92/30.23	0.60/0.46/0.71
	Dropout	78.08/87.60/87.67	2.45/3.10/13.47	0.30/1.05/5.53	9.99/8.39/33.12	3.60/0.89/1.59
	Attention entropy	78.35/87.90/87.77	2.10/2.30/11.67	1.28/0.10/6.55	5.92/8.01/36.15	4.98/0.96/2.10
	Causal debias	79.40/88.75/87.70	2.20/2.60/12.60	2.51/0.70/6.70	13.13/7.44/31.28	3.54/0.80/2.12
RoBERTa	Default	78.50/88.33/85.23	2.80/2.05/17.07	2.84/1.66/6.59	15.46/2.78/31.64	2.56/0.60/1.55
	Group balance	78.25/88.50/87.03	2.00/2.20/16.93	2.10/1.27/11.36	9.85/4.57/29.48	3.95/0.68/1.19
	Group-class balance	78.57/84.50/83.60	1.65/2.30/18.80	3.31/0.76/3.89	12.91/5.82/38.88	3.28/0.42/0.87
	CDA	76.75/87.58/85.20	1.60/1.75/14.20	6.37/0.31/4.10	15.91/5.41/35.70	0.82/0.42/1.19
	Dropout	78.33/88.92/86.73	2.15/1.55/14.53	2.42/0.58/8.86	11.11/3.96/27.05	4.08/0.56/2.10
	Attention entropy	78.33/88.42/86.67	1.75/1.75/15.73	2.89/0.23/9.23	10.91/5.60/24.68	3.82/0.69/1.75
	Causal debias	78.83/87.52/86.00	2.65/2.45/15.60	1.48/0.85/10.56	11.34/6.51/30.14	4.06/0.56/1.34
RoBERTa (all)	Default	78.88/88.70/87.90	2.95/2.40/13.80	2.24/0.58/9.50	13.55/7.19/33.47	4.14/0.95/2.35
	Group balance	79.30/88.65/87.93	2.90/2.00/14.73	1.27/0.17/12.30	11.03/7.74/31.69	5.02/1.06/2.80
	Group-class balance	79.40/89.15/87.93	1.70/1.10/12.73	4.43/0.24/5.08	13.65/3.24/25.90	4.17/0.75/1.58
	CDA	77.75/88.25/86.90	2.50/2.00/13.80	5.93/1.25/6.33	18.80/3.71/22.62	1.13/0.55/1.18
	Dropout	78.88/88.40/87.70	2.75/3.00/14.80	1.80/1.33/6.66	12.46/7.34/33.39	4.26/0.99/2.13
	Attention entropy	78.80/88.72/87.83	2.10/2.15/13.53	2.64/1.33/7.55	11.31/4.18/28.68	4.46/1.09/2.57
	Causal debias	79.27/89.78/87.80	3.35/1.25/15.00	3.24/0.51/11.86	16.00/3.05/37.57	3.56/0.74/2.70
Llama3.2	Zero-shot	63.78/74.62/71.27	1.45/2.35/24.67	11.03/3.52/36.81	10.54/1.03/2.95	2.13/2.94/3.83
	Few-shot	46.45/28.55/42.23	1.20/0.90/21.27	3.87/1.82/30.69	0.80/0.19/1.66	0.11/0.13/0.33
	Fairness imagination	64.95/75.92/73.37	0.80/0.85/21.87	8.70/3.61/32.54	9.44/6.79/5.98	2.65/3.58/3.50
	Fairness instruction	65.90/76.95/78.07	2.60/1.70/21.53	1.89/0.39/7.00	3.79/6.35/4.24	1.35/1.13/1.71

between different debiasing methods, we conduct an analysis based on how often a debiasing method successfully reduces the average individual unfairness (Avg_{iu}) and maintains the task performance (Accuracy) of the default / zero-shot model.

Encoder-only models Analyzing the results with respect to the dataset, we find that the models are able to better preserve their original accuracy on the Civil Comments dataset (48.61% of the cases) compared to the Jigsaw dataset (40.28% of the cases). In contrast, mitigating bias seems substantially easier on the Jigsaw dataset (in 63.88% of the cases) than on the Civil Comments (only 50% of the cases). On closer inspection, we find that this skew comes from religion bias in the Jigsaw dataset which is improved in 95.83% of the cases after debiasing, followed by race bias (50%) and gender bias (45.83%). In the Civil Comments dataset, we find that gender bias is mitigated best (improvement in 62.5% of the cases), followed by religion bias (54.17%) and race bias (33.33%).

With respect to the debiasing method, we find that CDA performs best in terms of debiasing, as it reduces Avg_{iu} across all bias types, datasets, and models. The second best performing method is group-class balance which manages to reduce Avg_{iu} in 58.33% of the cases on the Civil Comments dataset and in 75% cases on the Jigsaw dataset. For the other methods, the results are mixed as we again observe dataset-specific differences. For example, we find that Attention entropy performs well on the Jigsaw dataset (50%) but performs worst on the Civil Comments dataset (16.67%). These differences become even more pronounced when looking at different bias types. For instance, causal debiasing improves Avg_{iu} for religion bias across all models on the Jigsaw dataset but at the same time, does not improve a single model in terms of Avg_{iu} for gender bias in the same dataset. Interestingly, we find an inverse trend on the Civil Comments dataset; i.e., causal debiasing succeeds on all models for gender bias, but only for one model for religion bias. These findings highlight the importance of considering a diverse set of datasets for evaluating debiasing methods, as results on a single dataset can be misleading.

Decoder-only models We find that the debiasing methods (fairness imagination and fairness instruction) for the decoder-only models consistently improve the task performance across all bias types and datasets. Contrary to this, we see increases in average individual unfairness of the fair-

Table 5: Task performance and fairness results of default and debiased models on the Jigsaw dataset. Results are provided for race / gender / religion biases. Green (red) indicates the results are better (worse) than the default / zero-shot models. *All* indicates the model is trained on data containing all bias types.

Model	Method	Accuracy (↑)	Disp_{acc} (↓)	Disp_{fpr} (↓)	Disp_{fnr} (↓)	Avg_{iu} (↓)
BERT	Default	85.50/93.00/90.50	0.50/2.25/6.00	0.64/2.34/5.22	0.70/3.28/21.54	2.02/0.36/1.33
	Group balance	84.88/92.75/89.67	2.75/1.00/10.67	1.28/0.82/3.90	7.77/4.56/38.29	1.90/0.36/0.67
	Group-class balance	84.38/92.81/90.83	0.25/0.62/6.33	1.58/0.15/1.98	8.03/9.64/43.57	0.97/0.65/0.34
	CDA	85.25/91.81/90.50	4.00/3.63/10.00	4.12/3.44/5.10	2.97/7.38/37.39	0.39/0.28/0.45
	Dropout	85.62/92.69/89.83	1.25/3.37/9.67	0.31/3.03/5.46	6.51/8.41/27.37	2.75/0.36/1.00
	Attention entropy	85.00/92.06/89.83	0.00/3.12/9.33	0.62/3.03/4.29	1.72/6.00/28.06	2.93/0.50/0.98
	Causal debias	85.50/93.38/89.83	4.00/0.75/7.33	1.28/0.28/3.55	13.73/12.77/17.12	3.16/0.43/1.10
BERT (all)	Default	85.62/93.19/90.33	1.25/1.12/9.33	1.59/1.51/4.65	12.69/0.36/21.76	1.30/0.33/1.18
	Group balance	83.38/93.19/90.17	1.75/1.12/9.67	1.56/1.10/4.66	3.10/3.23/26.79	2.81/0.40/0.76
	Group-class balance	84.88/92.94/90.00	1.25/0.87/10.00	1.27/0.41/2.09	0.37/7.49/58.07	1.29/0.28/0.47
	CDA	85.62/92.19/90.00	3.25/1.88/7.00	2.86/1.78/4.02	4.17/4.41/38.24	0.69/0.29/0.46
	Dropout	86.50/93.44/91.00	3.00/1.38/7.00	1.26/1.10/5.60	10.24/6.10/13.16	1.91/0.33/1.27
	Attention entropy	85.25/93.75/91.50	0.50/2.75/8.00	0.65/2.62/5.19	0.57/5.54/34.85	2.57/0.41/1.07
	Causal debias	84.50/93.44/90.50	1.00/1.38/9.00	2.22/1.38/4.27	4.35/3.38/24.10	1.40/0.40/1.00
RoBERTa	Default	84.50/93.00/90.33	1.00/3.75/10.33	2.87/3.44/1.82	6.54/8.31/47.47	2.55/0.30/0.89
	Group balance	85.50/92.31/89.83	2.50/0.62/11.33	0.94/0.27/1.55	9.11/6.41/38.00	2.44/0.26/0.46
	Group-class balance	85.00/92.50/90.67	1.00/1.50/5.33	1.59/0.26/2.01	11.53/14.87/24.59	1.55/0.53/0.62
	CDA	85.12/93.19/89.33	0.75/1.88/8.67	4.12/1.10/3.90	12.64/11.13/25.89	0.36/0.23/0.40
	Dropout	83.88/93.69/90.17	1.75/0.88/7.67	1.29/0.82/2.97	3.10/3.28/26.86	2.71/0.23/0.87
	Attention entropy	85.00/93.50/90.33	0.50/1.75/6.67	2.23/2.06/1.01	6.55/0.62/22.78	2.39/0.24/0.81
	Causal debias	86.25/92.19/89.50	2.00/3.37/10.00	2.23/2.33/1.84	0.60/14.77/43.47	2.09/0.39/0.66
RoBERTa (all)	Default	85.50/93.75/91.50	0.50/1.75/7.00	0.01/1.51/5.56	3.06/5.74/31.14	2.52/0.35/1.55
	Group balance	85.38/93.62/91.67	1.75/3.25/9.33	0.01/2.47/4.12	9.01/11.90/40.29	2.76/0.30/0.96
	Group-class balance	86.38/92.56/90.17	2.25/1.88/10.67	0.62/1.37/2.58	9.05/8.62/64.35	4.75/0.23/0.34
	CDA	85.25/92.56/90.67	1.00/0.62/7.67	1.59/0.13/1.80	11.53/7.49/31.28	0.52/0.23/0.74
	Dropout	86.00/93.00/90.17	2.50/1.75/4.67	1.27/1.51/4.19	17.51/6.21/28.72	1.02/0.33/0.79
	Attention entropy	86.75/93.50/91.50	0.50/2.50/7.00	0.96/2.06/3.16	6.54/8.05/24.59	3.40/0.38/1.19
	Causal debias	85.38/93.25/91.00	0.25/3.50/10.00	0.01/2.62/5.41	1.88/13.69/34.14	2.55/0.40/0.80
Qwen3	Zero-shot	66.75/77.25/77.33	3.50/3.75/16.33	4.21/3.78/17.40	0.80/4.05/5.89	3.05/2.31/3.67
	Few-shot	56.12/61.44/75.67	7.75/1.63/9.67	9.95/2.11/11.48	0.23/1.69/2.44	3.93/4.96/4.13
	Fairness imagination	73.75/82.88/86.33	3.00/1.00/10.33	5.12/0.89/5.79	5.39/0.82/24.13	3.14/2.97/2.51
	Fairness instruction	78.00/89.50/89.33	3.00/0.50/9.33	4.14/0.26/3.13	2.04/5.23/26.74	1.95/1.43/1.61
Llama3.2	Zero-shot	54.00/70.50/65.17	8.50/1.00/25.67	10.91/1.53/31.87	0.20/4.56/8.33	2.39/3.00/4.28
	Few-shot	27.62/23.69/83.00	4.75/4.37/6.00	6.92/4.90/5.74	2.42/3.74/28.13	0.04/0.03/0.02
	Fairness imagination	57.75/73.56/66.83	5.00/1.62/26.33	6.47/1.63/30.03	0.26/1.74/17.48	2.86/3.73/3.92
	Fairness instruction	77.00/89.00/87.17	2.00/0.75/10.67	2.87/0.84/3.97	2.19/3.33/36.36	1.39/0.97/1.87

ness imagination approach for race and gender bias across both datasets. Only for religion, fairness imagination leads to a decrease of the individual unfairness. For fairness instruction, we observe a consistent improvement across all three bias types and both datasets, showing the clear superiority of the approach. The consistency of the results is especially surprising when considering that both decoder-only models are instruction-tuned, and that Chen et al. (2025) identify a bias amplification effect from instruction tuning. We conclude that fairness instruction is a good baseline to evaluate other debiasing methods for decoder-only models.

C Bias Detection Results

Fairness correlation We present the full fairness correlation results of encoder- and decoder-only models with different debiasing methods on Civil Comments and Jigsaw in Figures6, 7, 8, 9. Consistent with findings presented in the main text, Occlusion- and L2-based explanation methods achieve strong fairness correlations across different setups.

Comparing different debiasing methods, we find that low correlation scores primarily occur when individual unfairness is less pronounced, such as in CDA models. In these cases, the models themselves produce fewer biased predictions, making the detection of bias through explanations less critical. The lower correlations therefore do not substantially undermine the role of explanations in bias identification.

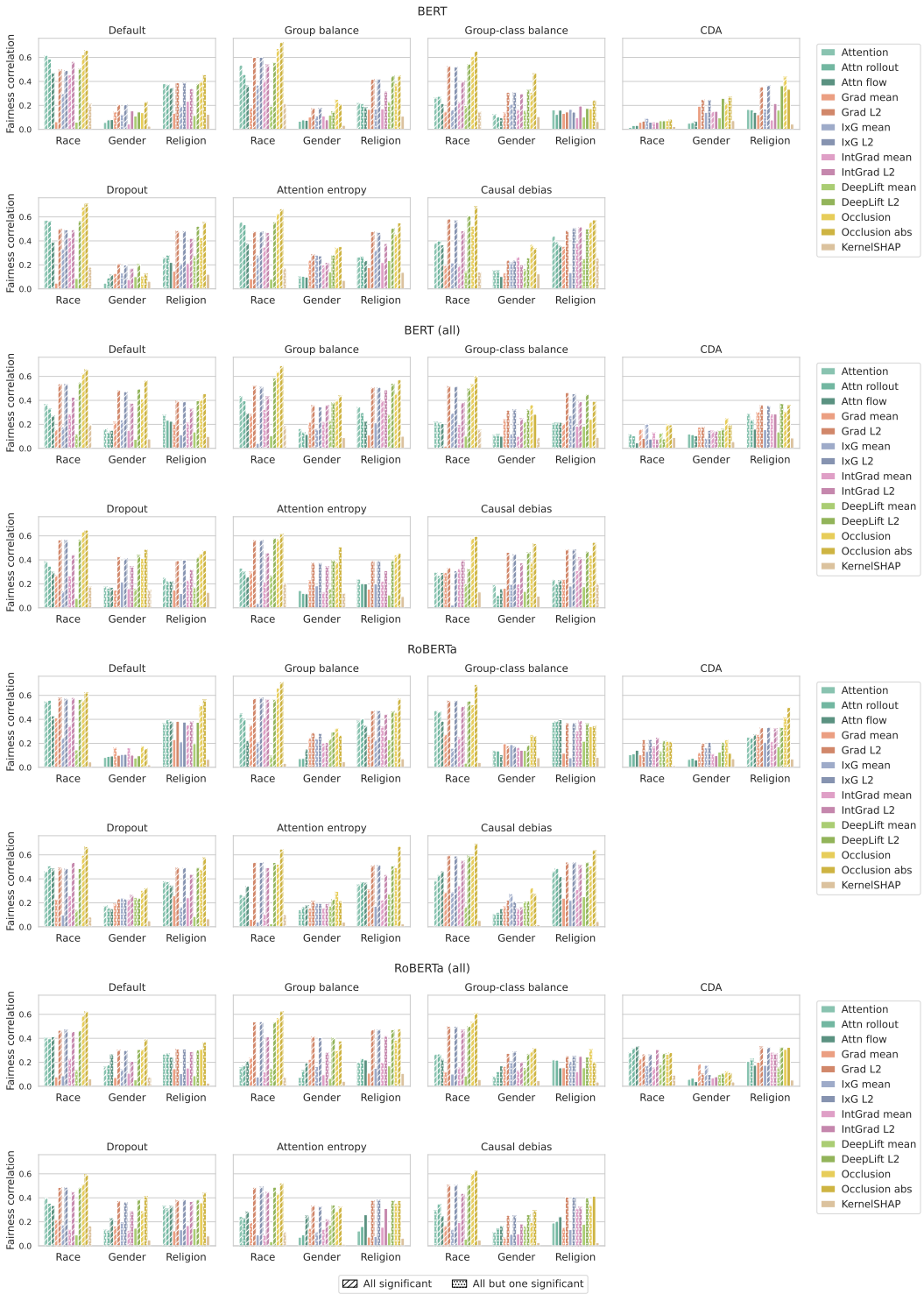

Figure 6: Fairness correlation results on Civil Comments for each explanation method across encoder-only models and bias types. Higher values indicate that the method is more effective and reliable in detecting biased predictions at inference time. *All* indicates the model is trained on data containing all bias types.

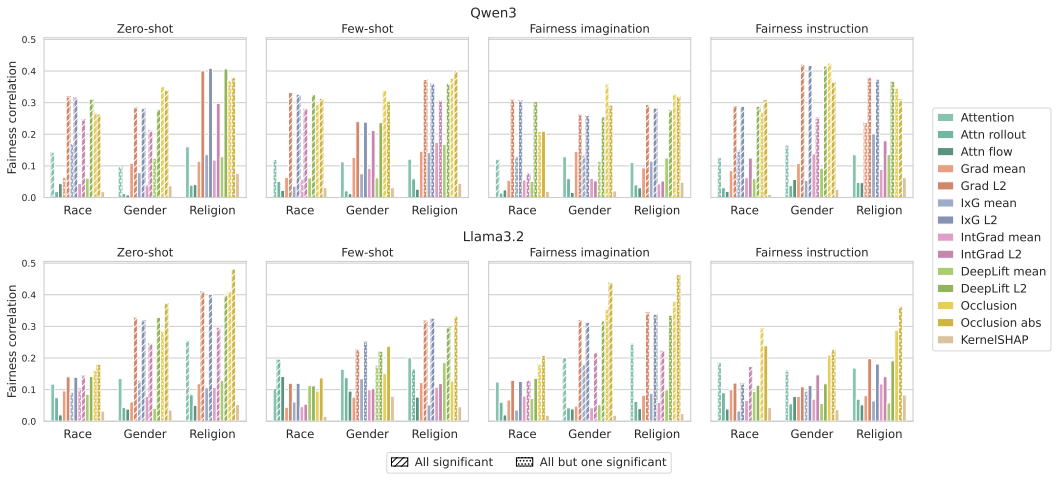

Figure 7: Fairness correlation results on Civil Comments for each explanation method across decoder-only models and bias types. Higher values indicate that the method is more effective and reliable in detecting biased predictions at inference time.

D Model Selection Results

Explanation-Based Metrics We evaluate several explanation-based metrics for selecting fair models with respect to different fairness criteria:

- **Average absolute sensitive token reliancee:** used to predict average individual unfairness, under the assumption that higher reliance on sensitive tokens implies greater sensitivity to group substitutions.
- **Group differences in average absolute sensitive token reliance:** used to predict disparities in accuracy, assuming that stronger reliance on sensitive features increases the risk of incorrect predictions.
- **Group differences in average absolute sensitive token reliance for positive/negative predictions:** used to predict disparities in false positive and false negative rates, respectively.

Among these, only average absolute sensitive token reliance exhibits rank correlations above random chance with its target fairness metric (individual unfairness). The correlations for other metrics remain at chance level. Figures 10, 11, 12, 13 demonstrate that no explanation methods can consistently match baseline rank correlation results. Figures 14, 15, 16, 17 further reveal that explanation methods are not able to select the fairest models. These findings underline the unreliability of explanation-based model selection.

E Bias Mitigation Results

The complete bias mitigation results are presented in Figures 18, 19, 20, 21. The findings are in line with conclusions from the main paper, that explanation-based debiasing can effectively reduce model biases across different fairness metrics, bias types, models, and datasets. In addition, the accuracy-fairness harmonic mean results shown in Figures 22, 23, 24, 25 demonstrate that explanation-based debiasing achieves comparable or superior balance between fairness and task performance than default models and traditional debiasing approaches.

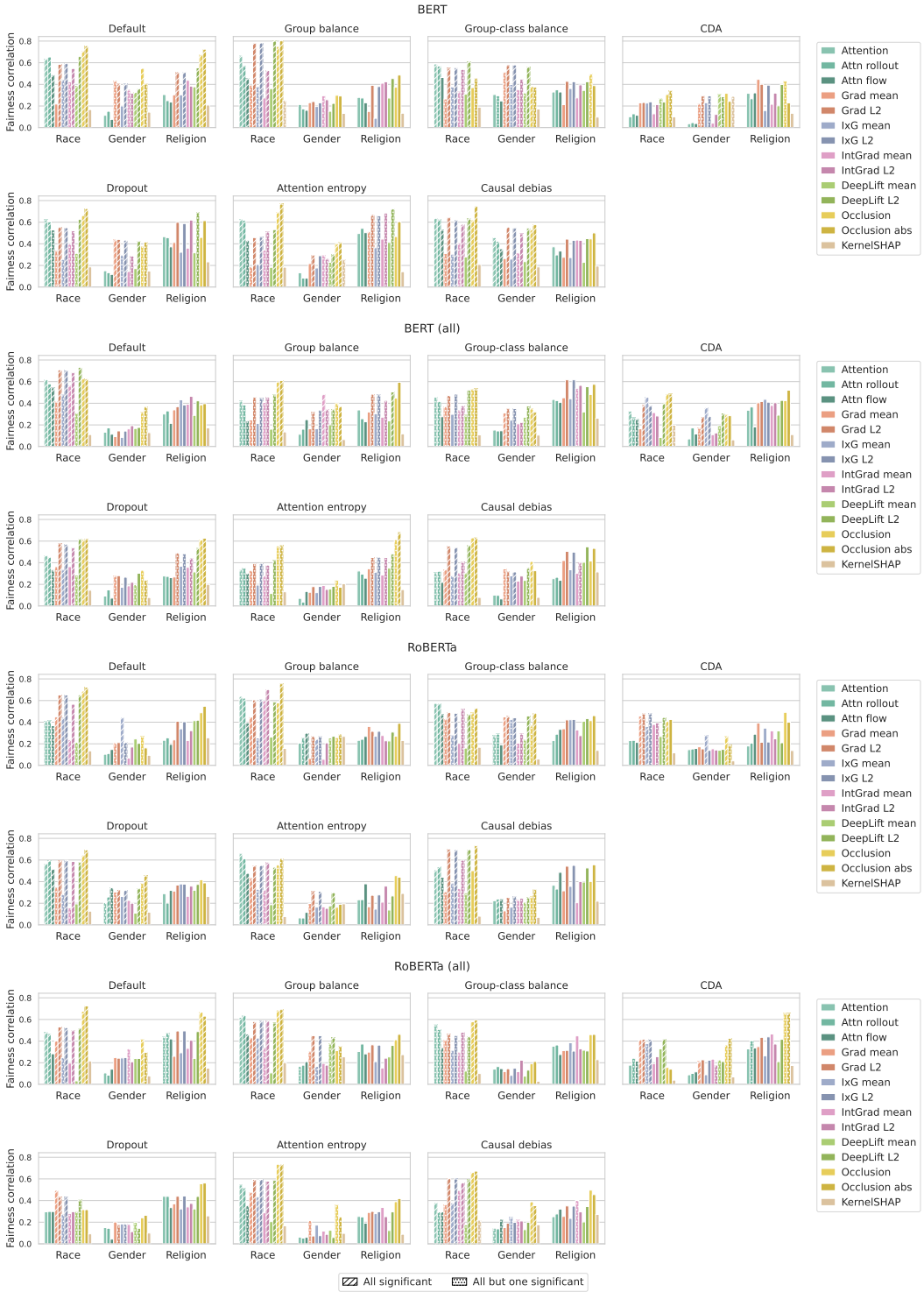

Figure 8: Fairness correlation results on Jigsaw for each explanation method across encoder-only models and bias types. Higher values indicate that the method is more effective and reliable in detecting biased predictions at inference time. *All* indicates the model is trained on data containing all bias types.

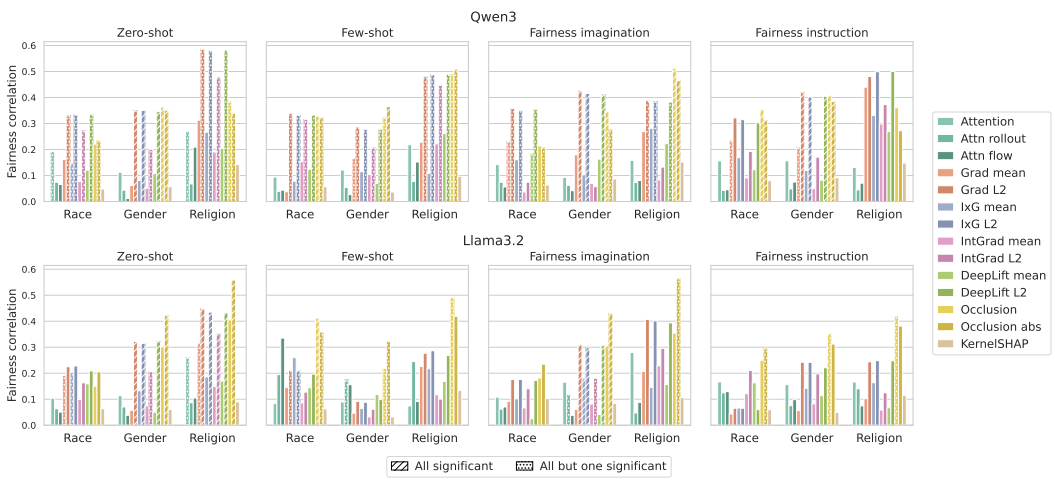

Figure 9: Fairness correlation results on Jigsaw for each explanation method across decoder-only models and bias types. Higher values indicate that the method is more effective and reliable in detecting biased predictions at inference time.

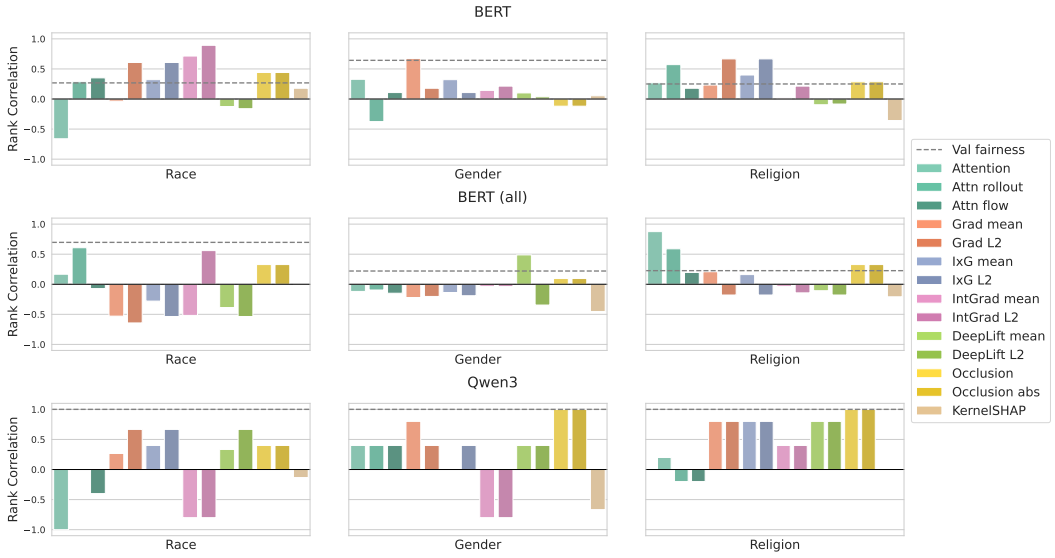

Figure 10: Rank correlations between validation set average absolute sensitive token reliance and individual unfairness on the test set on Civil Comments. The validation set sizes are 500 for race, 500 for gender, and 200 for religion. Higher correlation values indicate greater effectiveness in ranking models. *All* indicates the model is trained on all bias types.

F Fairness Correlations in Explanation-Debiased Models

Figure 26 presents the fairness correlation scores computed on explanation-debiased models. We find that Grad L2, IxG L2, DeepLift L2, and Occlusion-based explanations still show strong bias mitigation ability in the debiased models.

G Faithfulness as an Indicator of Bias Detection Ability

What factors influence the reliability of explanations in detecting bias? In this section, we examine the relationship between explanation faithfulness and their ability to identify bias, reflected by

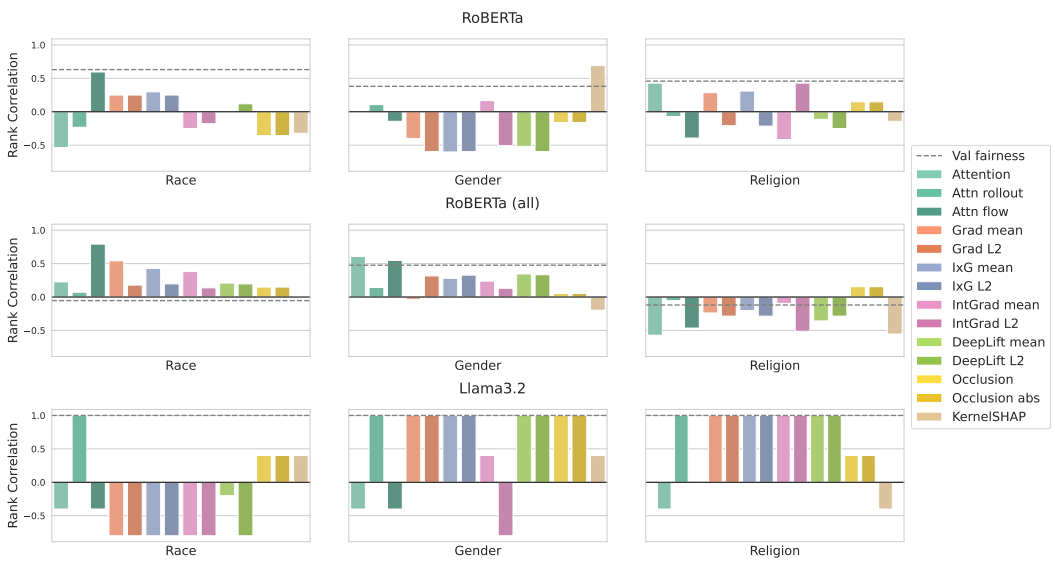

Figure 11: Rank correlations between validation set average absolute sensitive token reliance and individual unfairness on the test set on Civil Comments. The validation set sizes are 500 for race, 500 for gender, and 200 for religion. Higher correlation values indicate greater effectiveness in ranking models. *All* indicates the model is trained on all bias types.

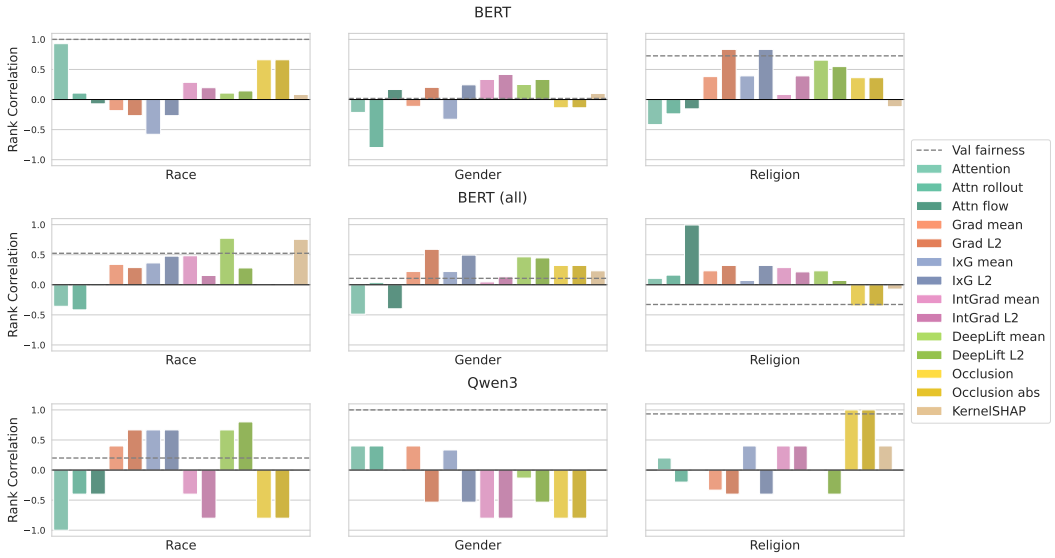

Figure 12: Rank correlations between validation set average absolute sensitive token reliance and individual unfairness on the test set on Jigsaw. The validation set size is 200. Higher correlation values indicate greater effectiveness in ranking models. *All* indicates the model is trained on all bias types.

fairness correlation scores in RQ1. We assess the faithfulness of explanation methods using two perturbation-based metrics: comprehensiveness and sufficiency AOPC (Area Over the Perturbation Curve; DeYoung et al., 2020), computed by masking 5%, 10%, 20%, and 50% of the input tokens. For substitution, we use the [MASK] token in BERT and the [PAD] token in Qwen3. Higher comprehensiveness and lower sufficiency scores indicate more faithful explanations.

Our results on race bias in Civil Comments (Figure 27 and Table 6) reveal no clear link between faithfulness and fairness correlation of explanations. In particular, mean-based explanations may

Figure 13: Rank correlations between validation set average absolute sensitive token reliance and individual unfairness on the test set on Jigsaw. The validation set size is 200. Higher correlation values indicate greater effectiveness in ranking models. *All* indicates the model is trained on all bias types.

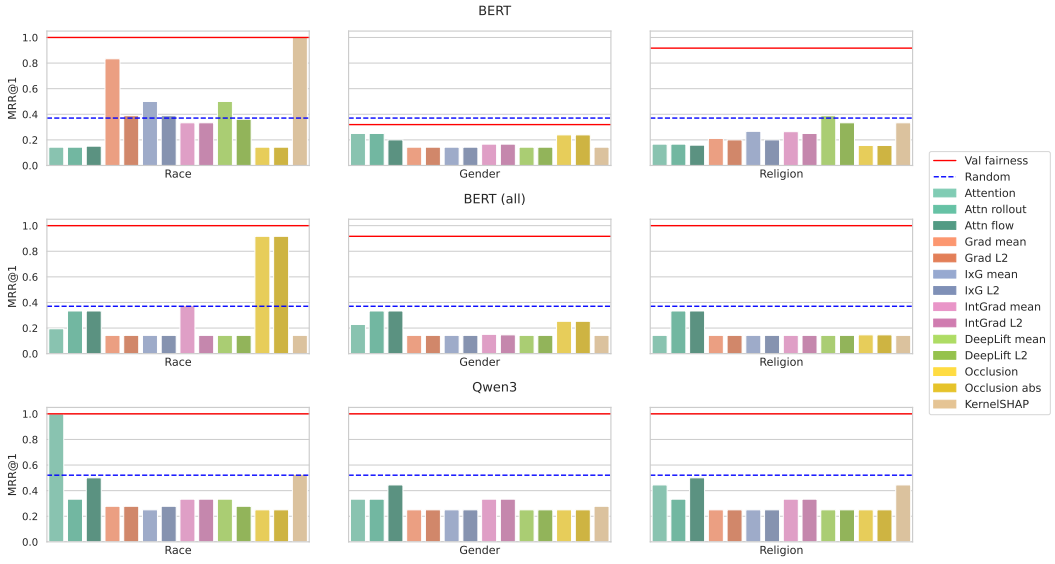

Figure 14: MRR@1 results on Civil Comments. The validation set sizes are 500 for race, 500 for gender, and 200 for religion. Higher MRR@1 scores indicate explanations are more effective in selecting the fairest models. *All* indicates the model is trained on all bias types.

achieve better faithfulness scores than their L2-based counterparts, yet they consistently perform significantly worse in identifying bias. We attribute this discrepancy to two key differences between the faithfulness metrics and our fairness correlation measure. First, faithfulness evaluates attribution scores across all input tokens, whereas our fairness correlation measure only considers sensitive token reliance. Second, perturbation-based faithfulness assesses the impact of masking tokens on model predictions, while our individual unfairness metric compares predictions when one social group is substituted for another. Taken together, these findings suggest that explanation faithful-

Figure 15: MRR@1 results on Civil Comments. The validation set sizes are 500 for race, 500 for gender, and 200 for religion. Higher MRR@1 scores indicate explanations are more effective in selecting the fairest models. *All* indicates the model is trained on all bias types.

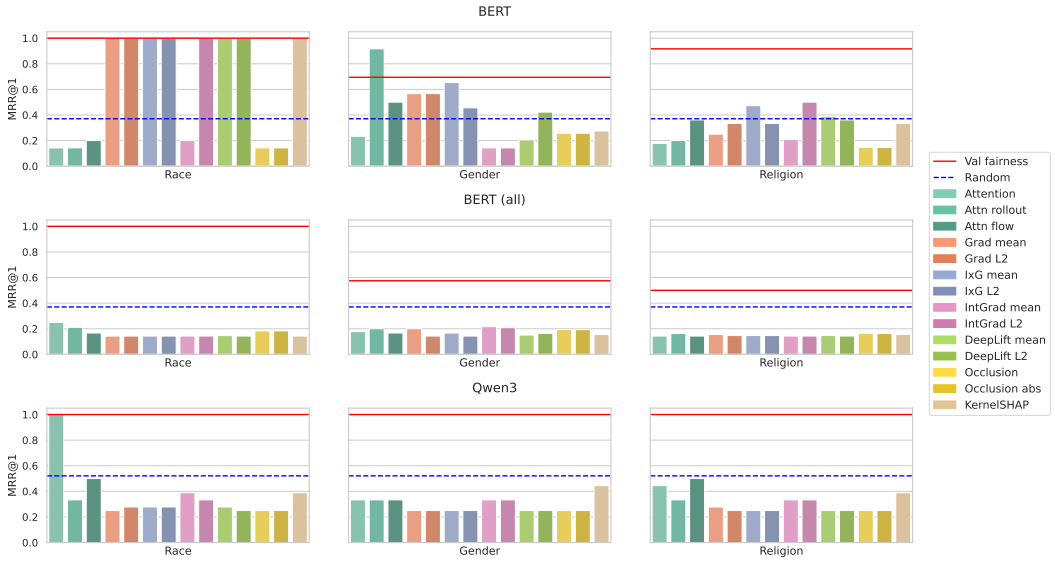

Figure 16: MRR@1 results on Jigsaw. The validation set size is 200. Higher MRR@1 scores indicate explanations are more effective in selecting the fairest models. *All* indicates the model is trained on all bias types.

ness is not a reliable indicator of bias detection ability. We therefore do not recommend selecting explanation methods for fairness on the basis of faithfulness results alone.

H LLM USAGE

Apart from the models evaluated in our experiments and analyses, we used LLMs (ChatGPT) solely to polish the writing in this work.

Figure 17: MRR@1 results on Jigsaw. The validation set size is 200. Higher MRR@1 scores indicate explanations are more effective in selecting the fairest models. *All* indicates the model is trained on all bias types.

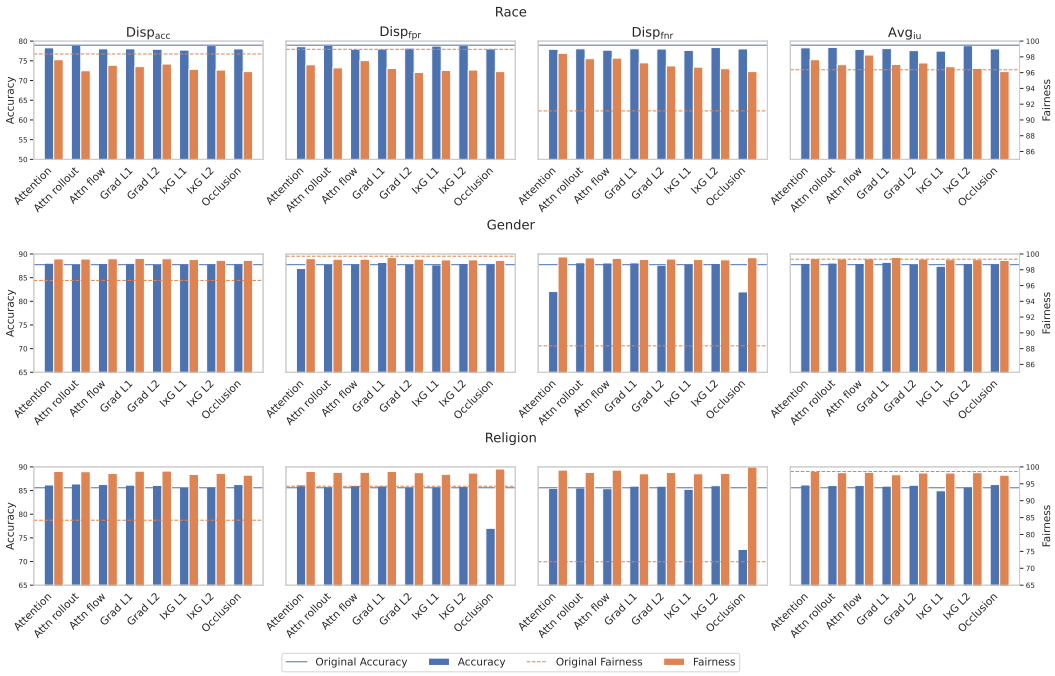

Figure 18: Accuracy and fairness results for bias mitigation in BERT on the Civil Comments dataset, using different explanation methods during training. For consistency with accuracy, fairness results are reported as $100-\{\text{Disp}_{\text{acc}}, \text{Disp}_{\text{fpr}}, \text{Disp}_{\text{fnr}}, \text{Avg}_{\text{iu}}\}$, so that higher values indicate better debiasing performance. Each column corresponds to models selected by maximizing the fairness-balanced metric with respect to the indicated bias metric.

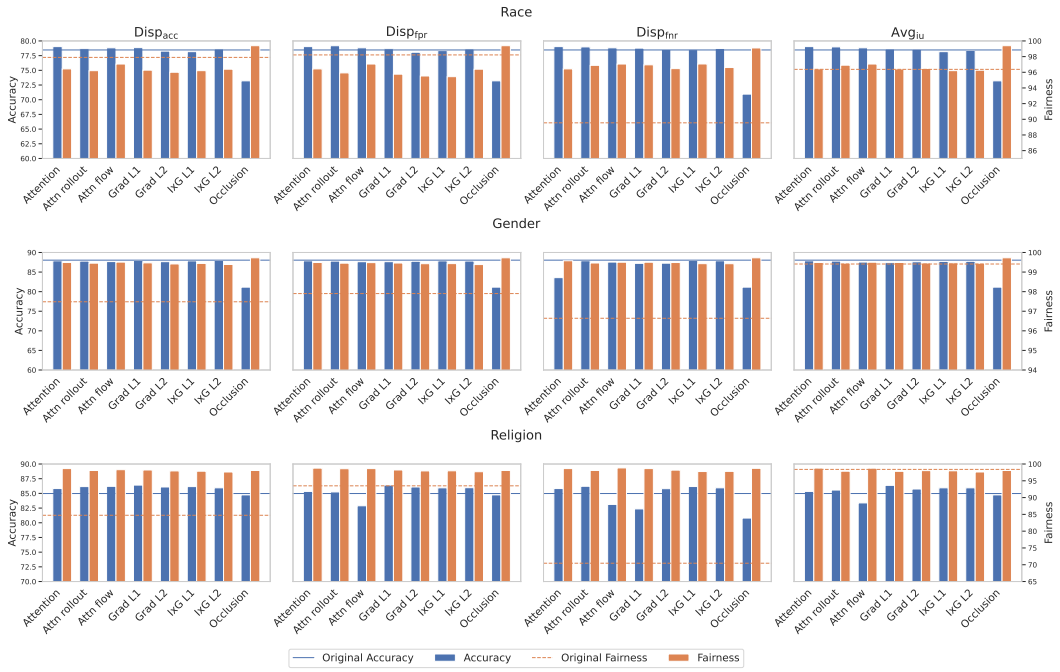

Figure 19: Accuracy and fairness results for bias mitigation in RoBERTa on the Civil Comments dataset, using different explanation methods during training. For consistency with accuracy, fairness results are reported as $100 - \{\text{Disp}_{\text{acc}}, \text{Disp}_{\text{fpr}}, \text{Disp}_{\text{fnr}}, \text{Avg}_{\text{iu}}\}$, so that higher values indicate better debiasing performance. Each column corresponds to models selected by maximizing the fairness-balanced metric with respect to the indicated bias metric.

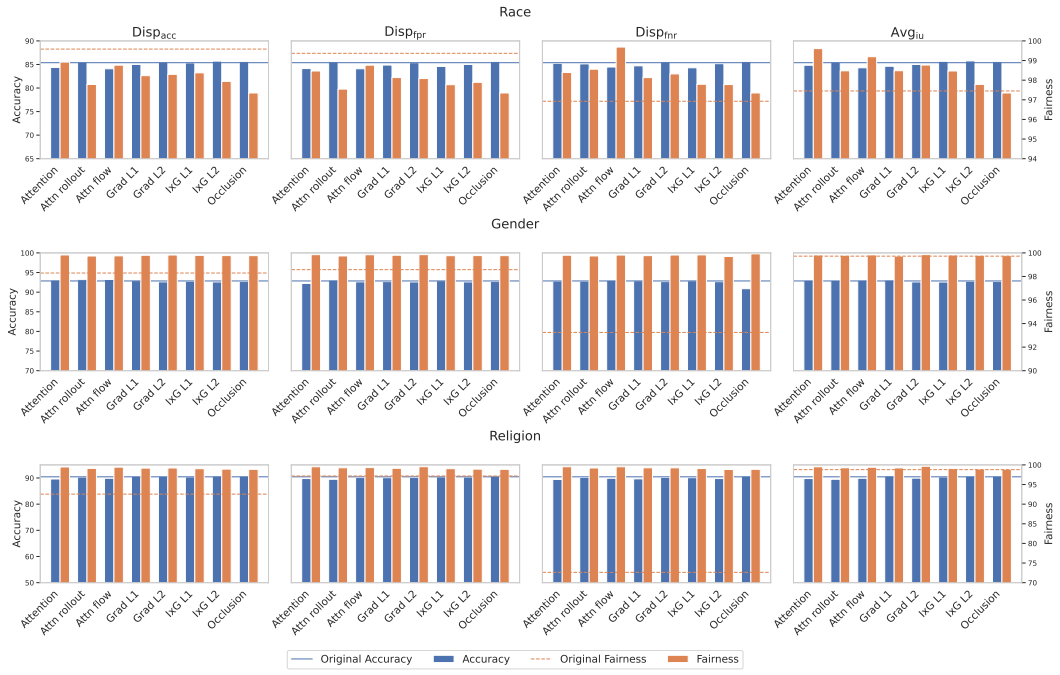

Figure 20: Accuracy and fairness results for bias mitigation in BERT on the Jigsaw, using different explanation methods during training. For consistency with accuracy, fairness results are reported as $100 - \{\text{Disp}_{\text{acc}}, \text{Disp}_{\text{fpr}}, \text{Disp}_{\text{fnr}}, \text{Avg}_{\text{iu}}\}$, so that higher values indicate better debiasing performance. Each column corresponds to models selected by maximizing the fairness-balanced metric with respect to the indicated bias metric.

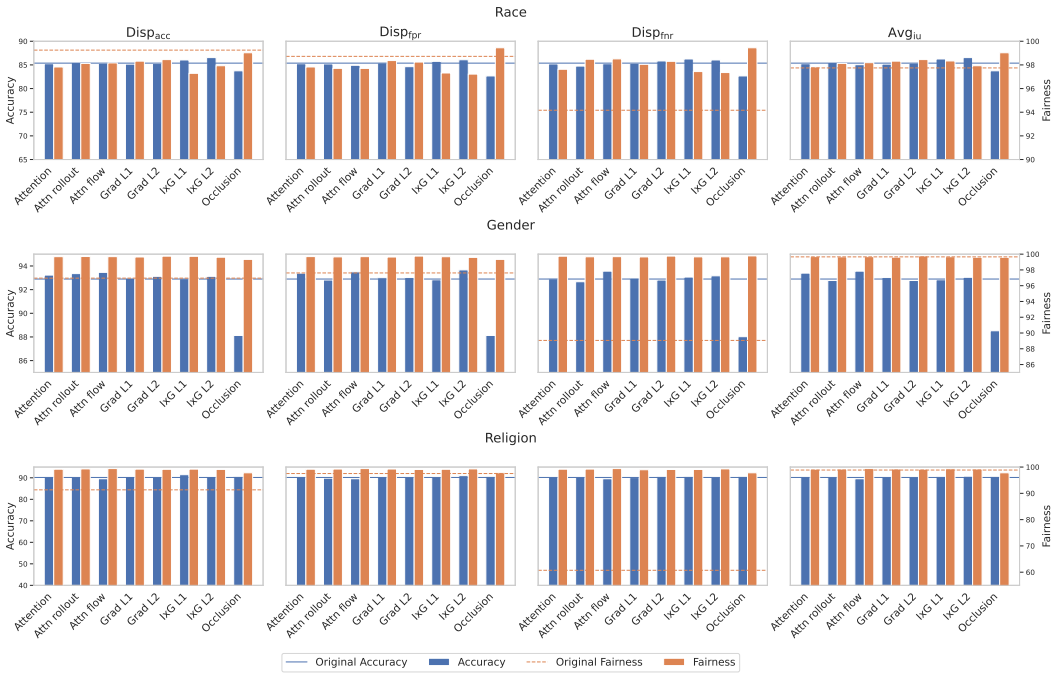

Figure 21: Accuracy and fairness results for bias mitigation in RoBERTa on the Jigsaw dataset, using different explanation methods during training. For consistency with accuracy, fairness results are reported as $100 - \{\text{Disp}_{\text{acc}}, \text{Disp}_{\text{fpr}}, \text{Disp}_{\text{fnr}}, \text{Avg}_{\text{iu}}\}$, so that higher values indicate better debiasing performance. Each column corresponds to models selected by maximizing the fairness-balanced metric with respect to the indicated bias metric.

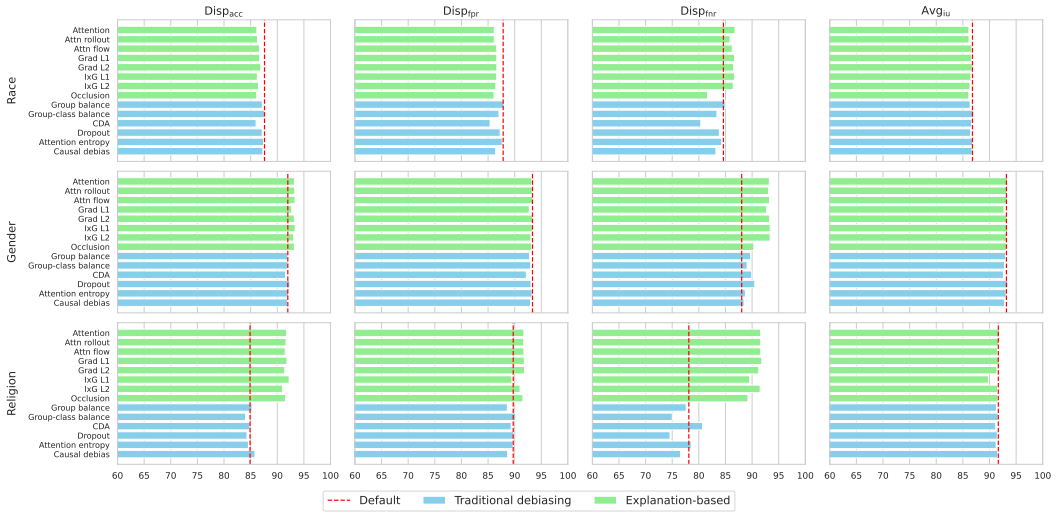

Figure 22: Harmonic mean between accuracy and fairness for established debiasing methods and explanation-based methods for BERT on Civil Comments. A higher score indicates better balance between model performance and fairness.

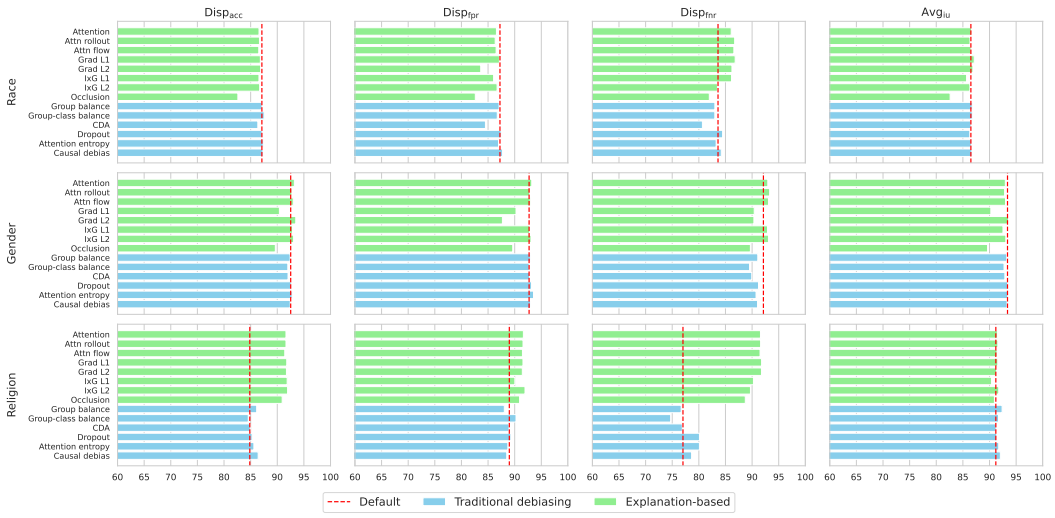

Figure 23: Harmonic mean between accuracy and fairness for established debiasing methods and explanation-based methods for RoBERTa on Civil Comments. A higher score indicates better balance between model performance and fairness.

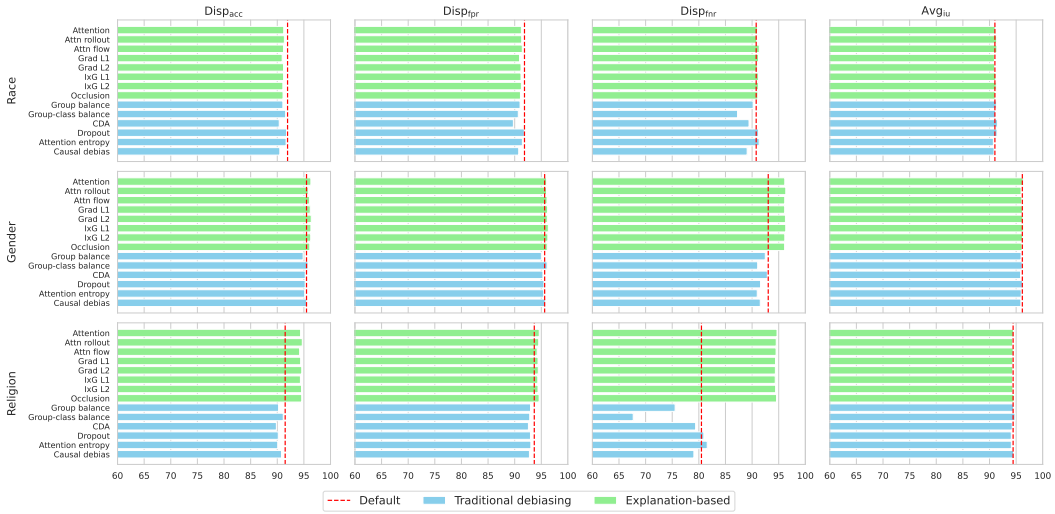

Figure 24: Harmonic mean between accuracy and fairness for established debiasing methods and explanation-based methods for BERT on Jigsaw. A higher score indicates better balance between model performance and fairness.

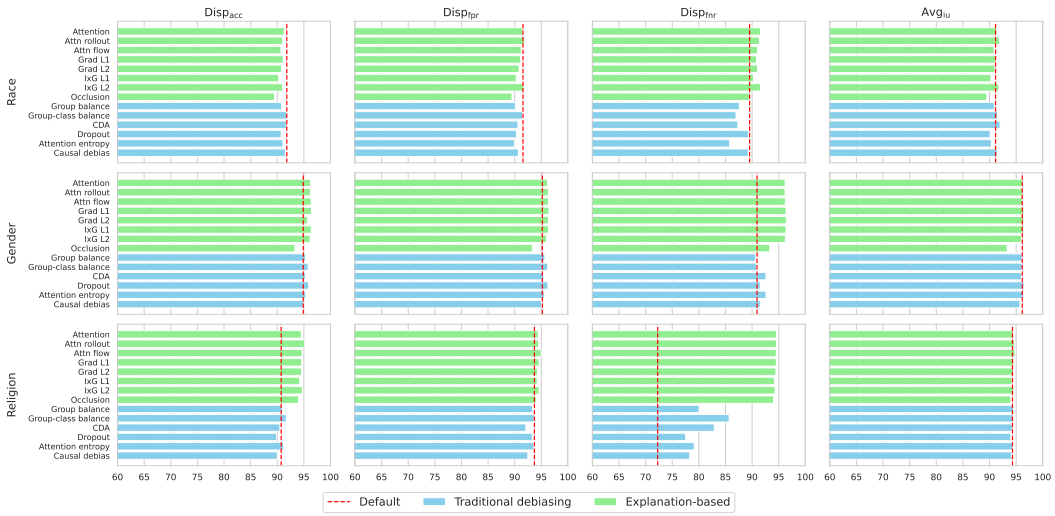

Figure 25: Harmonic mean between accuracy and fairness for established debiasing methods and explanation-based methods for RoBERTa on Jigsaw. A higher score indicates better balance between model performance and fairness.

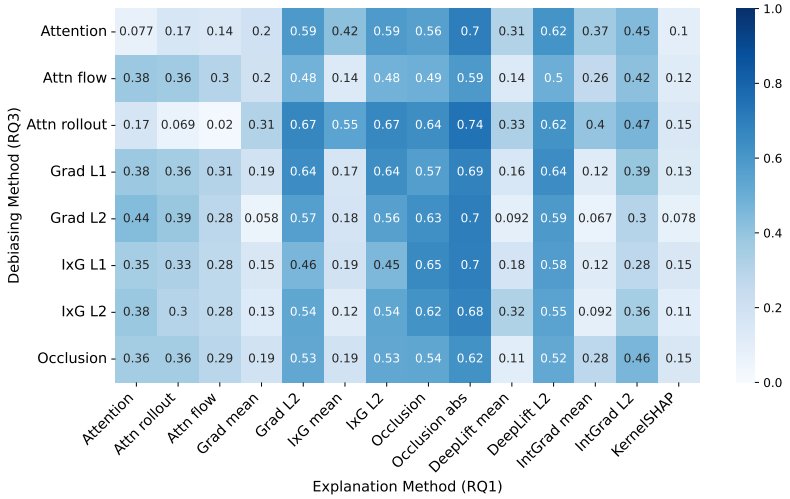

Figure 26: Fairness correlation results on BERT models with race bias mitigated through explanation-based methods on Civil Comments.

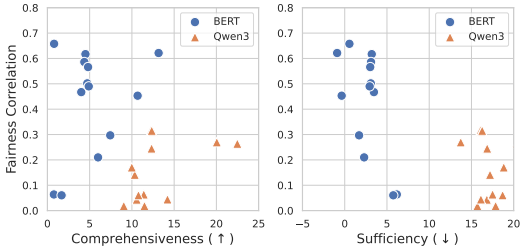

Figure 27: Faithfulness and fairness correlation results of different explanation methods. No clear relationship between explanation faithfulness and their bias detection ability is observed. Each point represents the faithfulness and fairness correlation of one explanation method applied to default / zero-shot models.

Table 6: Faithfulness results of different explanation methods on BERT and Qwen3 models.

Explanation	Comp. (↑)	Suff. (↓)	Fairness Correlation (↑)	Comp. (↑)	Suff. (↓)	Fairness Correlation (↑)
	BERT			Qwen3		
Attention	4.50	3.20	61.70	10.34	17.20	14.17
Attn rollout	4.37	3.11	58.60	9.04	15.70	1.86
Attn flow	4.01	3.46	46.73	10.57	16.82	4.38
Grad L2	4.82	2.99	49.57	12.30	16.09	31.80
Grad mean	0.77	6.16	6.37	11.41	17.50	6.40
DeepLift L2	4.72	3.09	50.18	12.44	16.17	31.08
DeepLift mean	1.68	5.75	6.04	10.78	18.69	6.11
IxG L2	4.89	2.95	48.99	12.35	16.27	31.53
IxG mean	7.44	1.70	29.69	9.99	18.82	17.07
IntGrad L2	4.81	3.02	56.60	12.33	16.86	24.52
IntGrad mean	10.68	-0.36	45.31	14.21	16.12	4.46
Occlusion	13.16	-0.90	62.14	20.05	13.73	26.98
Occlusion abs	0.79	0.56	65.77	22.48	20.36	26.41
KernelSHAP	5.99	2.30	20.99	11.49	17.86	1.86