Explainability Methods Can Be Biased: An Empirical Investigation of Gender Disparity in Post-hoc Methods

Anonymous submission

Abstract

While research on applications and evaluations of explanation methods continues to expand, fairness of the explanation methods concerning disparities in their performance across subgroups remains an often overlooked aspect. In this paper, we address this gap by showing that, across three tasks and five language models, widely used post-hoc feature attribution methods exhibit significant gender disparity with respect to their faithfulness, robustness, and complexity. These disparities persist even when the models are pre-trained or fine-tuned on particularly unbiased datasets, indicating that the disparities we observe are not merely consequences of biased training data. Our results highlight the importance of addressing disparities in explanations when developing and applying explainability methods, as these can lead to biased outcomes against certain subgroups, with particularly critical implications in high-stakes contexts. Furthermore, our findings underscore the importance of incorporating the fairness of explanations, alongside overall model fairness and explainability, as a requirement in regulatory frameworks.

1 Introduction

TestPre-trained language models (PLMs) are increasingly used in various natural language processing (NLP) tasks but are often hard-to-understand black boxes, which makes the problems of *explaining PLMs* and *evaluating those explanations* highly valuable. The growing demand to understand how PLMs generate their outputs has led to the increased adoption of Explainable AI methods in NLP. Explainable NLP, in particular, focuses on developing and applying techniques to interpret the inner workings and predictions of NLP models, including PLMs. Model-agnostic *post-hoc* feature importance methods have been particularly favored due to their wide applicability (Jacovi 2023). These methods aim to quantify the importance of each token for a given input and its corresponding model prediction. Such methods can make use of the gradients of the model with respect to its inputs (Sundararajan, Taly, and Yan 2017; Simonyan, Vedaldi, and Zisserman 2013), or use surrogate models (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017).

The growing interest in explainable NLP is evidenced by the increasing number of publications surveying explainability in NLP (Liu, Yin, and Wang 2019; Wallace, Gardner, and Singh 2020; Zhao et al. 2024; Madsen, Reddy, and Chandar 2022; Zini and Awad 2022; Danilevsky et al. 2020). Additionally, as NLP models are frequently applied in high-stakes domains such as medical (Johri et al. 2025) and legal settings (Valvoda and Cotterell 2024) where explainability is essential, a growing number of survey papers now focus on explainability in specific NLP tasks, including fact-checking (Kotonya and Toni 2020), text summarization (Dhaini et al. 2024), and for specific explainability methods in NLP (Mosca et al. 2022). Such surveys highlight the wide application of post-hoc methods in NLP. Furthermore, post-hoc methods are used as main explainers in numerous explainability tools and frameworks proposed in the literature (Arras, Osman, and Samek 2022; Li et al. 2023; Yang et al. 2022; Attanasio et al. 2023; Sarti et al. 2023). These frameworks typically incorporate a range of post-hoc explanation methods while supporting multiple data types and diverse machine learning (ML) model types, including PLMs.

Given the widespread adoption of these methods, evaluating their explanations is increasingly important. Explanation evaluation has become an active research area in recent years (Nauta et al. 2023; Longo et al. 2024), with numerous metrics and properties proposed (DeYoung et al. 2020; Arras, Osman, and Samek 2022; Sithakoul, Meftah, and Feutry 2024) to measure the quality of explanations. One desirable aspect of an explanation method is subgroup fairness: similar quality of the explanation across subgroups such as the different genders. For example,

consider a PLM-based AI system used by clinicians to diagnose patients from textual symptom descriptions and provide post-hoc explanations. The system misdiagnoses both a male and female patient with identical symptoms, where the explanation for the female patient correctly highlights the error, helping the physician identify the mistake. However, the explanation for the male patient falsely emphasizes relevant features in the input text, such as specific symptom-related keywords, misleading the physician into trusting the incorrect diagnosis. This discrepancy could undermine trust and harm patient outcomes.

However, there is a lack of research on evaluating the fairness of explanation methods across demographic groups, particularly in NLP.

Most previous works at the intersection of fairness and

explainability in NLP explore using explainability as a tool to detect bias in language models (T.y.s.s. et al. 2024; Balkir et al. 2022; Mathew et al. 2021; Brandl, Bugliarello, and Chalkidis 2024; Sarti et al. 2023; Vig et al. 2020; Gallegos et al. 2024) or facial recognition models (Huber et al. 2023) while some other recent works examine the influence of explanations on human-AI decision-making (Schoeffer, De-Arteaga, and Kühl 2024; Morrison et al. 2024). Despite the rigorous studies on evaluating fairness and bias in language models, less attention has been given to detecting bias in explanations or, in other words, the fairness of explanations themselves.

In this work, we evaluate disparities in the quality of post-hoc explanations across subgroups. We evaluate explanation quality based on a set of key explanation properties. Specifically, we investigate whether explanation methods produce similar faithfulness, robustness, and complexity across demographic groups , and focus on gender as a protected attribute. [1] We aim to answer the following research question: *Do post-hoc explanation methods perform equivalently across different subgroups, and if not, how can we evaluate gender disparities in explanations?*

Our findings indicate significant gender disparities in the explanations across different language models, even when the models do not exhibit significant bias.

Our main contributions are:

- We evaluate the gender disparity in six post-hoc explanation methods on four BERT-based models and GPT-2 using seven evaluation metrics to measure the quality of explanations with respect to their faithfulness, robustness, and complexity.
- We show that all methods can exhibit significant gender disparities regarding all the evaluation metrics used in the experiments.
- We further demonstrate that gender disparity in explanations persists even when the models are trained solely on an unbiased dataset, leading to the conclusion that the bias we observe is mainly influenced by the explanation methods.
- We finally outline and discuss the implications and considerations for practitioners based on our results.

We present this work as a step toward raising awareness of gender disparities in explanations and their implications, particularly when interpreting language model outcomes in real-world applications. We hope it contributes to ongoing research efforts aimed at improving the fairness and reliability of post-hoc explainability methods.

2 Related Work

A number of studies have highlighted limitations of post-hoc explainability methods (Madsen et al. 2024; Jain and Wallace 2019; Krishna et al. 2024a); however, they fail to consider how these methods perform across different subgroups, thus overlooking issues of fairness of explanations applied to textual datasets.

Wilming et al. (2024) study how bias in BERT can influence explanation correctness. They show how re-training and fine-tuning various components of the BERT architecture can improve explanation accuracy in identifying ground-truth tokens. However, this requires a dataset with ground-truth explanations, which is often not the case in more practically relevant datasets. Our study instead evaluates disparities in explanations using multiple metrics that capture three main properties of explanations, none of which requires a dataset with ground-truth explanations. We also introduce a setup designed to minimize any model-induced bias in the explanations, allowing us to investigate gender disparities independent of potential bias in the language models.

The topic of **disparities in post-hoc explanations** has been addressed in the literature by three studies, all focusing on tabular datasets (Dai et al. 2022; Balagopalan et al. 2022; Mhasawade et al. 2024). Dai et al. (2022) evaluate disparity in the explanation performance with respect to faithfulness (also referred to as fidelity), stability, consistency, and complexity while Balagopalan et al. (2022) focus their evaluation mainly on faithfulness. However, the two works solely experiment on tabular datasets and employ two model classes: linear regression and small neural networks. It remains unclear whether and to what extent explanation methods perform similarly across different subgroups when applied to textual datasets using various PLMs. Other works explore additional factors that may contribute to the level of disparities exhibited by certain post-hoc explainability methods. Balagopalan et al. (2022) and Mhasawade et al. (2024) investigate how specific data properties influence disparities in local explanations, with a particular emphasis on faithfulness. In particular, they examine whether the data representation encodes information about the sensitive attribute, and Mhasawade et al. (2024) further investigates other properties such as limited sample size and covariate shift and evaluates how model characteristics, like model complexity, can result in greater or lesser disparities in the fidelity of LIME (Ribeiro, Singh, and Guestrin 2016) explanations.

Textual datasets present unique challenges compared to tabular datasets. Among the general text-specific challenges in applying explainability methods (Zini and Awad 2022) compared to tabular datasets, isolating sensitive or protected attributes, such as gender, is particularly more complex in text. This complexity arises from the unstructured nature of text, the implicit representation of gender-related information (as opposed to being explicitly encoded in a single column in tabular datasets), and the context dependency, where gender-related information often depends on the surrounding text. In earlier studies, model selection was limited to either linear regression or a 3- to 4-layer neural network. This work, as detailed later in this paper, explores a diverse set of five transformer-based language models of varying sizes and complexity and two distinct architectures: encoder-only and decoder-only models. Such lan-

[1] Gender, race, age, among others, are referred to as protected attributes under the US anti-discrimination law (Xiang and Raji 2019)

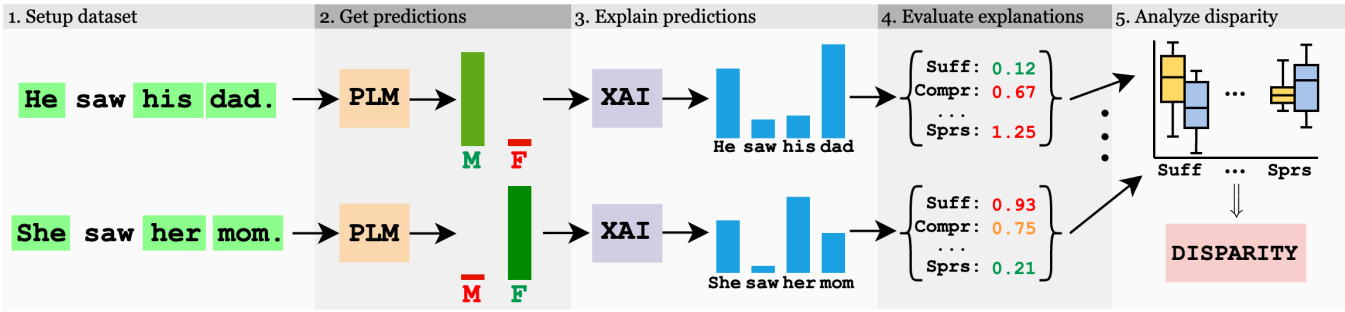


Figure 1: **Overview of our experimental pipeline**, exemplified with the GECO dataset (Wilming et al. 2024). We begin by obtaining predictions for male/female sentence pairs. We then use feature attribution methods to explain the predictions and evaluate the explanations using various metrics. We finally analyze the distributions of evaluation scores per each metric for male and female sentences and observe if the evaluations differ significantly between the two genders, indicating gender bias and disparity in explanations.

guage models demonstrate high performance in text classification tasks, making them an ideal choice for real-world applications. This underscores the importance of investigating disparities in the post-hoc explanations across subgroups when explanation methods are applied to explain the outcomes of these language models.

To the best of our knowledge, this paper presents the first study evaluating gender disparities in post-hoc explanations with respect to multiple quality metrics on various language models on textual datasets.

3 Disparity in Post-hoc Explanations

3.1 Local Post-hoc Explanation Methods

In our evaluation, we focus on six local feature attribution methods: Gradient (Saliency) (Simonyan, Vedaldi, and Zisserman 2013), Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017), SHAP (Lundberg and Lee 2017), LIME (Ribeiro, Singh, and Guestrin 2016), and extensions of Gradient and IG in which the input features are multiplied by the importance scores, named Gradient $\times$ Input (GxI) and IG $\times$ Input (IGxI). The applicability of these post-hoc explanation methods has also made them useful for various downstream tasks (Zini and Awad 2022; Danilevsky et al. 2020), and also recently for obtaining rationales from smaller models to be used in prompting large language models (LLMs) to improve their performance(Krishna et al. 2024b; Bhan et al. 2024).

3.2 Evaluating Explanations

Prior research on evaluating explanations has introduced various properties and desiderata that can be used to assess the quality of explanation methods and the explanations themselves (Robnik-Šikonja and Bohanec 2018; Nauta et al. 2023).

Several studies have built upon these foundational properties to develop metrics for evaluating different aspects of explanation quality. These metrics include both quantitative measures, such as fidelity, stability, consistency, and plausibility (Zhou et al. 2021; Yeh et al. 2019; DeYoung et al. 2020; Li et al. 2023), as well as qualitative approaches, which involve human-based evaluations of the generated explanations (Lakkaraju and Bastani 2020; Jesus et al. 2021) to assess how humans perceive these explanations.

In this paper, we quantitatively evaluate explanation quality based on three main desired properties: faithfulness, robustness, and complexity. Table 2 (Appendix B) presents the metrics we consider to measure the aforementioned properties to evaluate the quality of explanations.

Faithfulness refers to the degree to which an explanation accurately reflects and aligns with the internal workings and decision-making process of a model (Jacovi and Goldberg 2020). High faithfulness in explanations is desirable because it ensures that the explanation truly represents the model's functioning in making a prediction. We evaluate faithfulness using four metrics: comprehensiveness, sufficiency, soft comprehensiveness, and soft sufficiency. While sufficiency and comprehensiveness are commonly used, recent studies suggest they can lead to inaccurate faithfulness measurements due to the complete token removal operation they use (Zhao et al. 2022; Chrysostomou and Aletras 2022); therefore, we also use soft comprehensiveness and soft sufficiency, which have proven more accurate in measuring faithfulness by masking parts of the tokens' embeddings proportional to their importance scores rather than completely removing a fixed number of tokens (Zhao and Aletras 2023). Considering prior literature highlighting disagreement among metrics used to measure faithfulness (Jain and Wallace 2019; Krishna et al. 2024a), we employ multiple metrics that differ in evaluating faithfulness. The focus of our study is not to compare these metrics but rather to investigate disparities in explanations with respect to these metrics.

Robustness refers to the degree to which an explainability method responds to small perturbations and changes of the inputs, consistently producing reliable and stable explanations (Sithakoul, Meftah, and Feutry 2024; Alvarez-Melis and Jaakkola 2018). In particular, we try to compute the worst-case perturbation that results in the most significant change in the explanations within a region around the original input.

Complexity refers to the degree to which users can easily understand and interpret an explanation. Sparse explanations, compared to dense ones, are generally more favorable as they are less complex and easier to comprehend (Sithakoul, Meftah, and Feutry 2024). We evaluate explanation complexity using two measures: sparsity and Gini index. We define and provide the **formulation and the implementation details of these metrics** in Appendix B.

3.3 Implications of Significant Disparity

Based on the identified properties, we discuss the implications of the disparity in these properties. Although we focus this paper on gender, the same considerations can apply to other protected attributes.

Disparities in explanation faithfulness can result in explanations that do not accurately reflect the model's decision-making process across all groups, potentially leading to less accurate explanations for one group (e.g., female inputs) compared to another (e.g., male inputs). As shown in the example in the introduction, this could undermine stakeholders' trust, leading them to rely on incorrect model outputs. Significant *disparity in complexity* implies that the explanations for the model's decisions are more complex and, therefore, more challenging to understand for one group compared to another. *Robustness disparity* implies that explanations for one group exhibit higher sensitivity to slight perturbations, making them more vulnerable to noisy, erroneous data or adversarial attacks.

4 Experimental Setup

4.1 Datasets

An ideal dataset to test our hypothesis would contain male and female inputs where the only difference between a pair of male/female inputs is the gender in those inputs, and the difference in gender should have an influence on the task, i.e., the model should not be able to learn to ignore the genders. While ensuring that two inputs differ only in gender is easier to do with tabular data where gender is a categorical attribute, it is harder in textual data in which gender can be apparent in a number of ways: as the subject, or as an object, either explicitly through pronouns or more implicitly through nouns such as *sister/brother* or *actor/actress*. Ensuring that the models cannot ignore the genders is also non-trivial since it is hard to measure what features of a sentence actually reliably influence the inputs. In fact, knowing that would in a way be equivalent to having ground-truth feature importance explanations, as we would know that the words signaling gender in a sentence strongly influence the prediction. We experiment with three datasets, taking different approaches with respect to these two considerations. Table 1 displays example inputs from each of our datasets.

The first dataset is **GECO** (Wilming et al. 2024), consisting of pairs of sentences that only differ in their words signaling gender; e.g., replacing *him* with *her* and *sister* with *brother*. The task is to classify the gender in a sentence, either of the entire sentence or only the subject of a sentence. Thus GECO strictly enforces that pairs of sentences are identical except gender, and that those genders strongly influence the predictions, as they *are* the predictions themselves.

Next, inspired by the CrowS-Pairs dataset (Nangia et al. 2020), we construct the synthetic **Stereotypes** dataset by prompting Claude 3.5 Sonnet (Anthropic 2024) (see Appendix D for the details). It consists of sentence pairs differing only in their gendered words as in GECO, but the task is to classify if a sentence expresses a valid stereotype or not. A valid stereotype is one that is (even if factually inaccurate) associated with one gender more than the other, and the invalid sentences associate the same stereotype with the other gender. This way, we again have pairs of inputs identical except gender, but now the gender in a sentence is not the label directly although it strongly affects it. We validate our dataset first by manually verifying a subset of the inputs, and then by observing that models fine-tuned on this dataset can achieve high accuracy, indicating that the task is meaningful and can be solved with the information in the sentences.

Finally, we convert the tabular **COMPAS** (Angwin et al. 2016) dataset for recidivism prediction to text in order to obtain a dataset without pairs identical up to gender, and one in which the gender attribute has a weaker, although not negligible, influence on the task. Following earlier work (Fang et al. 2024), we convert each row to a comma-separated string such as `3 priors, score factor 1, under 45, under 25, African American, male, misdemeanor`. We do not process the COMPAS dataset to have input pairs that only differ in their gender. That would require assigning labels to previously unseen data points, which we avoid doing to not modify the original relationships between the existing features.

4.2 Models

We use five open-source language models that are accessible through the HuggingFace Hub for our experiments, with more information as well as hyperlinks in Table 3 in the appendix. The first two are a base BERT (Devlin et al. 2019) model and a distilled TinyBERT (Qian et al. 2022). The third is the GPT-2 model released by OpenAI (Radford et al. 2019), and the fourth is the RoBERTa-large model (Liu et al. 2019) which is the largest model we experiment with, with around 355M parameters. Finally, we experiment with a version of BERT released by Meta and named FairBERTa (Jiao et al. 2020). We chose to include FairBERTa as it is fine-tuned on a dataset in which inputs containing gender information are perturbed to non-binary words (e.g., he/she $\rightarrow$ they), that is argued to lead to a model which exhibits less disparity between genders.

4.3 Explanation Methods & Evaluation

For the implementation of our explanation methods (Section 3.1), we use the ferret library (Attanasio et al. 2023) which provides off-the-shelf support for models available through the Hugging Face transformers library. We also use ferret and an extension of it for its implementation of comprehensiveness, sufficiency, and sensitivity metrics. We provide our own implementations based on earlier work for

Table 1: Example inputs from the datasets used in our experiments. Bold words indicate those that are changed between male and female sentences of each pair.

Dataset	Examples	Labels	Task	Size
GECO ALL (Wilming et al. 2024)	**She** is cynically false about **her** childhood. **He** is cynically false about **his** childhood.	Female Male	Classify gender	Female: 1,610 Male: 1,610
GECO SUBJ (Wilming et al. 2024)	**She** takes her to a hospital. **He** takes her to a hospital.	Female Male	Classify gender	Female: 1,610 Male: 1,610
Stereotypes	As a **woman** CFO, **she** cut budgets ruthlessly. As a **man** CFO, **he** cut budgets ruthlessly.	Yes No	Detect stereotype	Female: 1,675 Male: 1,675
COMPAS (Angwin et al. 2016)	1 priors, score factor 0, under 45, under 25, Hispanic, male 0 priors, score factor 0, under 45, under 25, other race, female	Yes No	Predict recidivism	Female: 1,175 Male: 4,997

the sparsity, Gini index, and soft sufficiency/comprehensiveness.

Lower values are preferred for sufficiency (AOPC), soft sufficiency, and sparsity, while higher values are preferred for the other metrics.

4.4 Quantifying Disparity

We obtain feature-importance explanations for each input in our test set and evaluate the explanations with our metrics, resulting in a list of evaluation scores for the male and female subsets of the dataset, per explanation method and metric. We can then compare these lists per metric to quantify if there is a statistically significant difference or not, and if so how strong it is (i.e. the effect size).

To measure if the disparity is statistically significant, we follow the previous work (Dai et al. 2022) and use the Mann-Whitney U test that is applicable to subgroups with different sizes to test the null hypothesis that for any pair of values chosen from the subgroups, they are equally likely to be greater than each other. This corresponds to the methods performing equivalently between the two subgroups. We conclude there is a statistically significant difference if $p \leq 0.05$ and quantify the effect size with the *Cohen's d* metric we define in Appendix E.

5 Results and Analysis

Our main results are shown as follows: Tables 9 and 11 (Appendix H) display the counts of runs (out of five) resulting in statistically significant ($p \leq .05$) disparity, highlighting the cases with considerable effect size ($|d| \geq 0.2$, following the literature (Sawilowsky 2009)) with bold. Moreover, blue cells indicate that the male sentences have higher scores for that metric, while red cells indicate female sentences' scores are higher, with the strength of the color varying with respect to the count in the cell. To evaluate if the disparity we observe is a consequence of the models being pre-trained on biased data, we also report results after training BERT and GPT-2 from scratch on GECO in Section 5.4. We explain our training setup in more detail in Appendix G, and show the average effect sizes of disparities for each configuration in Tables 5, 6, 7, 8 in Appendix H, as well as further boxplots displaying the distributions of scores in Figures 3 in Appendix H. We also present a bias analysis in Appendix F using GECO and show that the models' predictive performance does not exhibit significant disparity.

In total, of the 5,040 combinations of dataset, model, explanation, metric, seeds in our experiments, 3,647 (72.4%) exhibit statistically significant ($p \leq 0.05$) disparity and 2,761 (54.8%) do so with a considerable effect size ($|d| \geq 0.2$).

5.1 Disparity per Explanation Method

To analyze disparity per method, we aggregate results in Tables 9,11, and 10 (Appendix H) (or Tables 5, 6,7 and 8 in the Appendix) row-wise considering all experimental combinations with metrics, models and datasets. Results show that IGxI (60%), SHAP (59%) and LIME (57%) exhibit the highest values for significant disparity with considerable effect size

followed by Grad (51.3%), GxI (51.1%), and IG (49%) where these values are notably high and reflect significant disparity and bias in the performance of these explanations.

Excluding IGxI, gradient-based methods (Grad, GxI, IG) show relatively less significant explanation disparity than perturbation-based methods (SHAP and LIME). Overall, results demonstrate that all six methods we considered exhibit significant disparity with considerable effect size on more than 49% of the combinations. These results are also reflected in the differences and gaps between the evaluation scores distributions in Figure 2 (and Figure 3 in Appendix H).

5.2 Disparity Across Metrics

Faithfulness Disparity On both GECO datasets, more than 95% of the runs with *soft sufficiency* and *soft comprehensiveness* results in significant disparity, with 85% also exhibiting considerable effect size, while the *comprehensiveness* and *sufficiency* less frequently result in significant disparity. Most noticeably on GECO-SUBJ, all runs with the soft metrics result in significant disparity. Furthermore, for all faithfulness metrics, the direction of their disparities for each model is often consistent between the explanation methods, indicating that the model plays a larger role in determining this direction than the explanation method.

In particular, the soft metrics show a noticeable decrease in the number of runs with significant disparity on the COM-

PAS and Stereotypes datasets, to less than 40% on COMPAS and 20% on Stereotypes. Nevertheless, the regular comprehensiveness and sufficiency metrics show a smaller decrease with 63% of runs on COMPAS and 71.5% on Stereotypes showing significant disparity. These results indicate that the soft removal operations can help reduce disparities in the faithfulness of explanations as long as the sensitive attribute is not the label directly.

We also observe in particular on the more practically relevant COMPAS dataset that unbiased pre-training as in FairBERTa and using larger models such as RoBERTa might help reduce the occurrence of disparities. Nevertheless, the high amount of faithfulness disparities visible across models and explanation methods for the comprehensiveness and sufficiency metrics highlights that feature attribution methods can lead to unfair performance between sensitive attributes such as gender.

Complexity Disparity On the GECO datasets, the complexity metrics Gini index and *sparsity*, exhibit disparity less frequently than the faithfulness metrics, at 70% and 49.5% respectively, with 53% and 42% of the total runs also resulting in disparity with considerable effect size. Results for the COMPAS and Stereotypes datasets also follow similar percentages, with 65% (38% with considerable effect size) and 68% (54% with considerable effect size) for the Gini Index on COMPAS and Stereotypes, and likewise 57% (27%) and 40% (30%) for sparsity.

Similar to the results with faithfulness metrics, using larger models such as RoBERTa decreases the disparity in the complexity of explanations, as it is most strongly visible when very few of the Grad, GxI, IG, and IGxI explanations show significant disparity in complexity on GECO. Nevertheless, despite this behavior, LIME and SHAP almost always result in disparity with RoBERTa on GECO, highlighting that the amount of disparity is not only a consequence of the model, but it varies with the explanation methods as well.

Sensitivity Disparity Considering sensitivity, 85% of runs on GECO results in disparities with considerable effect size, which is the highest among all evaluation metrics. Although not the highest among all metrics, 42% of run on Stereotypes and 45% on COMPAS also result in disparities with considerable effect size, indicating that such disparities persists across different datasets and explanation methods.

Following the trend from the previous metrics, the larger RoBERTa model again results in the least amount of disparity in sensitivity on COMPAS and Stereotypes, highlighting again that larger models may be less, although not completely, prone to gender disparities in their explanations. However, this is not visible on the GECO datasets, where all runs result in statistically significant disparities in sensitivity.

5.3 Disparities Across Datasets

Overall, our results confirm the hypothesis that the GECO datasets would show the highest amount of disparity, since the sensitive attribute had a stronger influence on the task by way of being the label itself. As we decrease the impact of gender on the task, first with the Stereotypes and then with the COMPAS datasets, we observe less disparity but still a significant one. This indicates that the amount of disparity depends not only on the model or the explanation methods, but on the dataset as well. However, most crucially, this is not because the datasets are particularly biased but because the dataset determines the influence the sensitive attribute has on the predictions.

5.4 Disparity when the Models are Trained from Scratch

To eliminate the possibility that the disparity we observe is just a consequence of the data the models were pre-trained on, we now apply our pipeline to models trained only on the two variants of the GECO dataset. More specifically, using BERT and GPT-2, we initialize the models randomly and then train them either on GECO-ALL or GECO-SUBJ for 50 epochs.

Table 10 displays the number of runs out of five resulting in statistically significant ($p \leq .05$) disparity, and highlights those that has a considerable effect size ($d \geq 0.2$) in bold. Similar to the results in Table in 9, more than 80% of runs for both models show significant gender disparity. Moreover, the direction of the disparity per metric, as indicated by the colors of cells in the table, also follows a similar pattern. For instance, in both sets of results, male sentences have explanations with higher soft sufficiency and sparsity scores, and female sentences have higher soft comprehensiveness scores. Thus, we conclude that even if trained only on an unbiased dataset such as GECO[2], the explanation methods frequently result in disparate treatments of the two genders. These results confirm that while datasets and models can influence the disparities in explanations, aligning with (Mhasawade et al. 2024),

they are not the sole cause and explanation methods themselves can contribute to these disparities.

5.5 Implications and Considerations for Researchers and Practitioners

Although disparity results can vary between metrics, all explanation methods under study consistently exhibit significant explanation disparities with considerable effect sizes across all included metrics. These results underscore the need for stakeholders (e.g., practitioners, developers, researchers) to consider general and metric-specific explanation disparities when using explanations for PLMs outputs to make informed decisions, depending on their use case.

[2] We refer to GECO as "unbiased" as it does not distinguish between the two genders. Concretely, if the ground truth explanation words were masked, it would be impossible to determine which sentences were male sentences and which were female sentences. This is because with the masked inputs, the two sentences in each male-female pair appear identical, and there is no way to distinguish between the two genders since the dataset is perfectly balanced as well. This observation implies that there is no property of the dataset besides the gender words that affect the labels in any way. This is unlike a potentially biased dataset such as COMPAS where even if the gender of each data point was masked, the remaining features' statistics could be used to infer the masked genders to an extent.

Practitioners use explainability methods to interpret a model's decision-making process, debug the model, or improve its performance. As previously discussed, the convenience and ease of use provided by explainability frameworks make them popular among developers seeking explanations for real-world applications. However, directly applying post-hoc methods or relying on frameworks that support them can mislead developers when evaluating the model, particularly in gender-related tasks, leading to biased decisions and critical outcomes for both the system under development and any subsequent projects that utilize such frameworks. Practitioners should therefore recognize that these methods can exhibit significant gender disparities. We recommend they carefully consider such disparities based on the explanation properties most relevant to their use case. As discussed earlier, each explanation property has different implications. Accordingly, practitioners must assess which properties are most critical in their specific contexts.

For example, certain disparities in explanations can be more critical for some stakeholder groups than for others. For instance, explanations with complexity disparity might not be critical for developers. However, it can be very relevant for laypeople who often need simple and easily understandable explanations. On the other side, significant disparity in faithfulness represents a major concern for all stakeholder groups as it implies explanations that inaccurately reflect the model's decision-making process between subgroups, which could result in critical consequences, similar to the example presented in section 1. We urge practitioners to thoroughly audit the properties of explanations for each subgroup.

Researchers can benefit from our open-source pipeline to develop new explanation methods and evaluation metrics and identify the reasons behind the disparity we observe. Practitioners can also use this pipeline to run tests to identify potential failure modes of the methods they are using. For example, detecting that explanation quality varies significantly by gender in a task where it should be gender-neutral raises a potential red flag, especially in gender-sensitive contexts. Thus, we call upon researchers and developers to account for gender disparities in post-hoc explanations when introducing new libraries and frameworks that employ these methods.

For **developers** of AI systems, particularly those integrating PLMs, employing explainability methods that exhibit gender disparities in systems can lead to non-compliance with transparency requirements outlined in regulations such as the EU AI Act (Council of European Union 2024). This risk is particularly pronounced in high-risk settings, where biased explanations can undermine the system's fairness, disadvantage certain subgroups, and impose significant liability on developers and deployers of such systems. As post-hoc methods are widely used across various AI applications, **end-users**, such as doctors, who receive explanations generated by these methods, should be aware that they may exhibit gender disparities, particularly in gender-sensitive tasks or use cases. For **regulators and policymakers**, our findings emphasize the importance of explicitly integrating explanation fairness as a requirement in both existing and future regulations alongside model and data fairness.

6 Conclusion

In this paper, we presented the first study investigating disparities in the quality of post-hoc feature attribution methods for language models across subgroups, focusing specifically on gender as a protected attribute. We showed that every investigated explanation method presents a significant degree of bias across various metrics, even when the models are trained from scratch on an unbiased dataset, with the most pronounced disparities emerging in faithfulness and sensitivity. These results underscore the importance of going beyond model-level fairness and scrutinizing the fairness of explanations themselves.

Despite the limitations discussed in Section A, this work can serve as an essential foundation for researchers and practitioners seeking to evaluate existing and novel methods of interpreting language models. By unveiling potential gaps in how explanation quality varies for different demographic groups and metrics, we highlight the broader need for fairness-promoting algorithms that address explanation-level bias. Building on recent efforts to mitigate model bias (Brandl, Bugliarello, and Chalkidis 2024), we advocate for fairness-focused strategies aimed at reducing disparities in explanation performance.

Looking forward, there are multiple avenues for **future work**:

- Extending methods and metrics: Incorporate new approaches (Leemann et al. 2025; Deiseroth et al. 2023) or additional implementations of existing metrics.
- Broadening data coverage: Generate synthetic datasets or augment existing textual datasets to capture protected attributes beyond gender, ensuring compatibility with our disparity measurement pipeline. Additionally, we plan to expand our dataset collection by integrating further datasets addressing gender bias in NLP tasks, such as WinoBias (Zhao et al. 2018), while still being aware of the shortcomings of such datasets (Blodgett et al. 2021).
- Combining quantitative with human-based evaluation: Complement the standard metrics with human-grounded assessments to evaluate explanations (Nauta et al. 2023; Longo et al. 2024) to capture nuanced disparities early.

Our findings point to the need for deeper theoretical and empirical investigation into the causes of explanation bias and the contribution of each cause, whether stemming from the explanation methods themselves, the fine-tuning process, or dataset design. A better understanding of these underlying mechanisms will be pivotal for developing robust mitigation strategies and ensuring that explanation fairness is upheld alongside transparency and predictive fairness.

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the Robustness of Interpretability Methods. arXiv:1806.08049.

Angwin, J.; Larson, J.; Kirchner, L.; and Mattu, S. 2016. Machine bias.

Anthropic, A. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Arras, L.; Osman, A.; and Samek, W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81: 14–40.

Attanasio, G.; Pastor, E.; Di Bonaventura, C.; and Nozza, D. 2023. ferret: a Framework for Benchmarking Explainers on Transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Balagopalan, A.; Zhang, H.; Hamidieh, K.; Hartvigsen, T.; Rudzicz, F.; and Ghassemi, M. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1194–1206. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Balkir, E.; Kiritchenko, S.; Nejadgholi, I.; and Fraser, K. 2022. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In Verma, A.; Pruksachatkun, Y.; Chang, K.-W.; Galstyan, A.; Dhamala, J.; and Cao, Y. T., eds., *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 80–92. Seattle, U.S.A.: Association for Computational Linguistics.

Bhan, M.; Vittaut, J.-N.; Chesneau, N.; and Lesot, M.-J. 2024. Self-AMPLIFY: Improving Small Language Models with Self Post Hoc Explanations. *arXiv preprint arXiv:2402.12038*.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., 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)*, 1004–1015. Online: Association for Computational Linguistics.

Brandl, S.; Bugliarello, E.; and Chalkidis, I. 2024. On the Interplay between Fairness and Explainability. In Ovalle, A.; Chang, K.-W.; Cao, Y. T.; Mehrabi, N.; Zhao, J.; Galstyan, A.; Dhamala, J.; Kumar, A.; and Gupta, R., eds., *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, 94–108. Mexico City, Mexico: Association for Computational Linguistics.

Chrysostomou, G.; and Aletras, N. 2022. An Empirical Study on Explanations in Out-of-Domain Settings. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6920–6938. Dublin, Ireland: Association for Computational Linguistics.

Council of European Union. 2024. Council regulation (EU) no 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Dai, J.; Upadhyay, S.; Aivodji, U.; Bach, S. H.; and Lakkaraju, H. 2022. Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 203–214. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In Wong, K.-F.; Knight, K.; and Wu, H., eds., *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–459. Suzhou, China: Association for Computational Linguistics.

Deiseroth, B.; Deb, M.; Weinbach, S.; Brack, M.; Schramowski, P.; and Kersting, K. 2023. AtMan: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation. arXiv:2301.08110.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

Dhaini, M.; Erdogan, E.; Bakshi, S.; and Kasneci, G. 2024. Explainability Meets Text Summarization: A Survey. In Mahamood, S.; Minh, N. L.; and Ippolito, D., eds., *Proceedings of the 17th International Natural Language Generation Conference*, 631–645. Tokyo, Japan: Association for Computational Linguistics.

Fang, X.; Xu, W.; Anting Tan, F.; Zhang, J.; Hu, Z.; Qi, Y.; Nickleach, S.; Socolinsky, D.; Sengamedu, S.; and Faloutsos, C. 2024. Large language models on tabular data–a survey. *arXiv e-prints*, arXiv–2402.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 3681–3688.

Hsia, J.; Pruthi, D.; Singh, A.; and Lipton, Z. 2024. Goodhart's Law Applies to NLP's Explanation Benchmarks. In

Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1322–1335. St. Julian's, Malta: Association for Computational Linguistics.

Huber, M.; Fang, M.; Boutros, F.; and Damer, N. 2023. Are Explainability Tools Gender Biased? A Case Study on Face Presentation Attack Detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, 945–949.

Jacovi, A. 2023. Trends in Explainable AI (XAI) Literature. arXiv:2301.05433.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In Burstein, J.; Doran, C.; and Solorio, T., 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)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Jentzsch, S.; and Turan, C. 2022. Gender Bias in BERT-Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 184–199.

Jesus, S.; Belém, C.; Balayan, V.; Bento, J. a.; Saleiro, P.; Bizarro, P.; and Gama, J. a. 2021. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 805–815. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4163–4174. Online: Association for Computational Linguistics.

Johri, S.; Jeong, J.; Tran, B. A.; Schlessinger, D. I.; Wongvibulsin, S.; Barnes, L. A.; Zhou, H.-Y.; Cai, Z. R.; Van Allen, E. M.; Kim, D.; et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 1–10.

Kotonya, N.; and Toni, F. 2020. Explainable Automated Fact-Checking: A Survey. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 5430–5443. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Krishna, S.; Han, T.; Gu, A.; Wu, S.; Jabbari, S.; and Lakkaraju, H. 2024a. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *Transactions on Machine Learning Research*.

Krishna, S.; Ma, J.; Slack, D.; Ghandeharioun, A.; Singh, S.; and Lakkaraju, H. 2024b. Post hoc explanations of language models can improve language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Lakkaraju, H.; and Bastani, O. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 79–85. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.

Leemann, T.; Fastowski, A.; Pfeiffer, F.; and Kasneci, G. 2025. Attention Mechanisms Don't Learn Additive Models: Rethinking Feature Importance for Transformers. arXiv:2405.13536.

Li, X.; Du, M.; Chen, J.; Chai, Y.; Lakkaraju, H.; and Xiong, H. 2023. M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 1630–1643. Curran Associates, Inc.

Liu, H.; Yin, Q.; and Wang, W. Y. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5570–5581. Florence, Italy: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Ser, J. D.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; Jiang, R.; Khosravi, H.; Lecue, F.; Malgieri, G.; Páez, A.; Samek, W.; Schneider, J.; Speith, T.; and Stumpf, S. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106: 102301.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lyu, Q.; Apidianaki, M.; and Callison-Burch, C. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 50(2): 657–723.

Madry, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Madsen, A.; Lakkaraju, H.; Reddy, S.; and Chandar, S. 2024. Interpretability Needs a New Paradigm. arXiv:2405.05386.

Madsen, A.; Reddy, S.; and Chandar, S. 2022. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.*, 55(8).

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 14867–14875.

Mhasawade, V.; Rahman, S.; Haskell-Craig, Z.; and Chunara, R. 2024. Understanding Disparities in Post Hoc Machine Learning Explanation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2374–2388. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Morrison, K.; Spitzer, P.; Turri, V.; Feng, M.; Kühl, N.; and Perer, A. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Mosca, E.; Szigeti, F.; Tragianni, S.; Gallagher, D.; and Groh, G. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 4593–4603. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s).

Qian, R.; Ross, C.; Fernandes, J.; Smith, E. M.; Kiela, D.; and Williams, A. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9496–9521.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " 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*, 1135–1144.

Robnik-Šikonja, M.; and Bohanec, M. 2018. *Perturbation-Based Explanations of Prediction Models*, 159–175. Cham: Springer International Publishing. ISBN 978-3-319-90403-0.

Sarti, G.; Feldhus, N.; Sickert, L.; and van der Wal, O. 2023. Inseq: An Interpretability Toolkit for Sequence Generation Models. In Bollegala, D.; Huang, R.; and Ritter, A., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 421–435. Toronto, Canada: Association for Computational Linguistics.

Sawilowsky, S. S. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8: 597–599.

Schoeffer, J.; De-Arteaga, M.; and Kühl, N. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sithakoul, S.; Meftah, S.; and Feutry, C. 2024. BEExAI: Benchmark to Evaluate Explainable AI. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 445–468. Cham: Springer Nature Switzerland. ISBN 978-3-031-63787-2.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

T.y.s.s., S.; Baumgartner, N.; Stürmer, M.; Grabmair, M.; and Niklaus, J. 2024. Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16500–16513. Torino, Italia: ELRA and ICCL.

Valvoda, J.; and Cotterell, R. 2024. Towards Explainability in Legal Outcome Prediction Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7269–7289. Mexico City, Mexico: Association for Computational Linguistics.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12388–12401. Curran Associates, Inc.

Wallace, E.; Gardner, M.; and Singh, S. 2020. Interpreting Predictions of NLP Models. In Villavicencio, A.; and Van Durme, B., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 20–23. Online: Association for Computational Linguistics.

Wilming, R.; Dox, A.; Schulz, H.; Oliveira, M.; Clark, B.; and Haufe, S. 2024. GECOBench: A Gender-Controlled Text Dataset and Benchmark for Quantifying Biases in Explanations. *arXiv preprint arXiv:2406.11547*.

Xiang, A.; and Raji, I. D. 2019. On the Legal Compatibility of Fairness Definitions. arXiv:1912.00761.

Yang, W.; Le, H.; Laud, T.; Savarese, S.; and Hoi, S. C. H. 2022. OmniXAI: A Library for Explainable AI. arXiv:2206.01612.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (In)fidelity and Sensitivity of Explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Zhao, Z.; and Aletras, N. 2023. Incorporating Attribution Importance for Improving Faithfulness Metrics. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4732–4745. Toronto, Canada: Association for Computational Linguistics.

Zhao, Z.; Chrysostomou, G.; Bontcheva, K.; and Aletras, N. 2022. On the Impact of Temporal Concept Drift on Model Explanations. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4039–4054. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhou, J.; Gandomi, A. H.; Chen, F.; and Holzinger, A. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5).

Zini, J. E.; and Awad, M. 2022. On the Explainability of Natural Language Processing Deep Models. *ACM Comput. Surv.*, 55(5).

A Limitations

The main limitation of our evaluation is that it is limited to gender disparity and binary classification tasks. More insights into the disparities amplified by the explanation methods can be gained by analyzing different sensitive attributes such as race, as well as other tasks, such as text generation. Our evaluations are also limited to transformer-based models, although such models currently see the highest use.

We also acknowledge that the design process for the synthetic Stereotypes dataset could benefit from recommendations in the literature to avoid potential pitfalls associated with evaluation corpus design (Blodgett et al. 2020, 2021). Finally, evaluating explanations with respect to desirable properties such as faithfulness and robustness is an active research area itself, with new evaluation methods frequently being proposed to address the shortcomings of existing methods (Jacovi and Goldberg 2020; Lyu, Apidianaki, and Callison-Burch 2024; Hsia et al. 2024). Thus, our analysis is also limited by the current state of the evaluation literature and would benefit from future developments.

B Definition and Formulation of Evaluation Metrics

B.1 Comprehensiveness

Measures how relevant the tokens assigned high-importance are for classification. Let f_j be the output probability for the correct class j. Top-k tokens r are removed and the difference $f_j(x) - f_j(x \setminus r)$ is the comprehensiveness value. In ferret, comprehensiveness is measured using the area over perturbation curve (AOPC) that is computed by varying k (the number of tokens to remove) by varying the threshold such that the tokens with an importance score above the threshold are removed, and averaging the resulting comprehensiveness values. The resulting values thus lie in the interval [0,1]. We vary threshold from 0.1 to 1.0 in increments of 0.1. A high value indicates a significant change in the model's output, which implies that the removed tokens were important for classification. Then we conclude that an explanation successfully captures the relevant tokens if it has a high comprehensiveness value.

B.2 Sufficiency

As opposed to comprehensiveness, only the top-k tokens r are input to the model and the sufficiency value is the difference $f_j(x) - f_j(r)$ in the model's output. A small value indicates that only the tokens assigned high importance were enough to obtain the same output, and hence that the explanation was able to capture the most relevant tokens. Then the number k is varied similar to the comprehensiveness metric, except this time removing the tokens with importance scores below the threshold, and the AOPC is computed by averaging the resulting sufficiency values, with the final values between 0 and 1.

B.3 Soft Sufficiency and Comprehensiveness

Removing tokens entirely can lead to out-of-distribution inputs, meaning that the explanations are evaluated on kinds of inputs the model never saw during training and is unlikely to see in real use. To reduce this difference between the actual inputs and those used in evaluation, (Zhao and Aletras 2023) instead propose to mask a fraction of each token's embeddings based on that token's importance score. For a the vector representation $\mathbf{x}$ of a token with importance score s normalized between 0 and 1, the input is perturbed to obtain $\mathbf{x}'$ such that

$$\mathbf{x}' = \mathbf{x} \odot \mathbf{e}, \quad \mathbf{e}_i \sim \text{Ber}(q) \tag{1}$$

with $q = s$ if the elements are to be retained (for sufficiency) and $q = 1 - s$ if they are to be removed (for comprehensiveness). Finally for original and perturbed sentences $\mathbf{X}$ and $\mathbf{X}'$

Table 2: Overview of the considered explanation properties and metrics used to evaluate explanation quality.

Property	Metric	Definition
Faithfulness	Comprehensiveness (DeYoung et al. 2020)	Measures whether the explanation captures all the evidence (i.e., tokens) used by the model to make a prediction by assessing the drop in model probability when relevant tokens are removed
	Sufficiency (DeYoung et al. 2020)	Measures whether the tokens identified by the explanation are sufficient for the model to make a prediction .
	Soft Comprehensiveness & Soft Sufficiency (Zhao and Aletras 2023)	To prevent evaluating explanations on out-of-distribution inputs as a result of removing tokens entirely as in comprehensiveness and sufficiency, for the *soft* versions each token's embedding is masked proportionally to its importance score.
Complexity	Sparsity (Dai et al. 2022)	Counts the number of features with an attributed importance greater than a given threshold.
	Gini-index(Sithakoul, Meftah, and Feutry 2024)	Measures the concentration of explanations on specific features by computing the Gini index of attribution vector. A high value, close to 1, indicates a greater concentration of attribution on fewer tokens, which is more desirable compared to a low value, close to 0, where attribution is more evenly distributed across multiple tokens. .
Robustness	Sensitivity (Yeh et al. 2019)	Measures the extent of change in the explanation when there is a slight alteration in the input. High sensitivity in explanations can be problematic, as it may render the explanation method more susceptible to adversarial attacks (Ghorbani, Abid, and Zou 2019).

with true class y, soft sufficiency and comprehensiveness are defined as

$$\text{Soft-S} = 1 - \max(0, p(y|X) - p(y|\mathbf{X}')) \in [0,1] \quad (2)$$

$$\text{Soft-C} = \max(0, p(y|X) - p(y|\mathbf{X}')) \in [0,1] \quad (3)$$

where p denotes the model output logits.

B.4 Sparsity

For a given explanation vector $(s_1, ..., s_n)$, we compute the share of scores exceeding a threshold τ (0.1 for our experiments) in absolute value:

$$\text{Sparsity} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[|s_i| \geq \tau\right] \in [0,1] \quad (4)$$

where $\mathbf{1}$ denotes the indicator function. Lower non-zero values are preferred as they indicate only a few tokens were assigned high scores, which makes the explanation easier to understand. Sparsity of zero is not desired since it means all tokens were assigned relatively low scores with respect to the threshold.

B.5 Gini Index

For the explanation vector $\mathbf{s} = (s_1, ..., s_n)$ sorted in an ascending way with respect to the scores' absolute values and $\mathbf{k} = (k_1, ..., k_n)$ denoting the indices of the original elements in the sorted vector, we compute

$$\text{Gini Index} := 1 - 2 \sum_{i=1}^{n} \frac{s_i}{\|\mathbf{s}\|_1} \cdot \frac{n - k_i + 0.5}{n} \in [0,1] \quad (5)$$

with higher values (more sparse) being preferred.

B.6 Sensitivity

Given input x and model f with explainer Φ, we find the input y within a ball of radius r around x such that the change in the explanation is maximized (i.e. the worst-case perturbation). A lower worst-case difference indicates the explanation method is more robust to small perturbations:

$$\text{Sensitivity} = \max_{y; \|x-y\| \leq r} \frac{\|\Phi(f,y) - \Phi(f,x)\|}{\|\Phi(f,x)\|} \in [0, \infty). \quad (6)$$

We use a projected gradient descent (PGD) (Madry 2017) attack in which the input is perturbed in the direction of the gradient maximizing the prediction error, and projected back onto the ball after each gradient step.

C Reproducibility

Our end-to-end pipeline is designed to be easily reproducible. We base our experiments on the publicly available GECO (Wilming et al. 2024) and COMPAS datasets (Angwin et al. 2016), as well as the synthetic Stereotypes dataset we create and make public. We use the publicly available models from Huggingface (see Table 3) running on a single NVIDIA V100 GPU. Including fine-tuning, and generating and evaluating explanations, one model/dataset run takes between 40-60 minutes without the sensitivity metric and 10 hours with sensitivity. We use the open source ferret library (Attanasio et al. 2023) for implementations of the explanation methods and metrics. For the metrics not available in ferret, we either use and include in our codebase other publicly available implementations with appropriate licenses, or provide our own implementations.

Table 3: Information about the models used in our experiments. The names are hyperlinks directing to their respective HuggingFace Hub pages.

Name	Type	Num. Params
TinyBERT	Encoder-only	14,350,874
FairBERTa	Encoder-only	124,647,170
BERT	Encoder-only	167,357,954
GPT-2	Decoder-only	124,442,112
RoBERTa-large	Encoder-only	355,359,744

D Prompting Claude for the Stereotypes Dataset

To create our Stereotypes dataset, we prompt Claude 3.5 Sonnet (Anthropic 2024), the most recent version as of November 2024. Rather than using a single prompt, we start with an initial prompt, and then iterate in a few steps of conversation depending on the quality of the sentences generated. Once the sentences fulfill our requirements, we repetitively ask Claude to generate a number of sentences.

The initial prompt is:

> *I want you to generate a small dataset. It will consist of pairs of sentences. The only difference between the sentences in each pair will be the subject's gender. E.g.:*
>
> - *He is a doctor.*
> - *She is a doctor.*
>
> *The second characteristic of the dataset is that the first sentence in each pair will express a stereotype towards one gender. So the second sentence will be the same, just with the gender flipped, and the stereotype naturally won't hold for that gender. E.g.*
>
> - *She was a bad doctor, no surprises.*
> - *He was a bad doctor, no surprises.*
>
> *Do you understand? Generate one sentence pair so I can see if you get the task.*

After this prompt, we give feedback for two steps until the outputs are of desired quality. Our feedback consists of the instructions

> *You get the point but the examples you generated are not very good. Generate a few more and I will tell you the best. Then we will refine.*

and

> *But the stereotypes are not explicitly obvious in the sentences. I want them to be more clear. Something like "I was surprised to see a woman doctor articulate herself so well."*

E Definition of Cohen's d

To quantify the effect size in our experiments we use the Cohen's d metric defined as

$$d = \frac{\bar{x}_M - \bar{x}_F}{s} \quad \text{with} \quad s = \sqrt{\frac{\sigma_M^2 + \sigma_F^2}{2}} \tag{7}$$

with $\bar{x}_M, \bar{x}_F$ average male/female scores and σ_M^2, σ_F^2 the variances.

F Bias Analysis

To ensure that the bias we observe is independent of the model, we quantify the gender bias in each of our models through a bias analysis after fine-tuning, with the results displayed below. The true positive rate (TPR), true negative rate (TNR), and the average prediction difference (APD) (Jentzsch and Turan 2022) defined as follows:

$$\text{TPR} = \frac{\text{Accurately predicted male}}{\text{Total male}} \tag{8}$$

$$\text{TNR} = \frac{\text{Accurately predicted female}}{\text{Total female}} \tag{9}$$

$$\text{APD} = \mathbb{E}_{x \sim \mathcal{D}} \left| f_{\text{sm}}^{M}(x_M) - f_{\text{sm}}^{F}(x_F) \right| \tag{10}$$

where x_M and x_F are the male and female versions of the same input, and $\mathcal{D}$ denotes the dataset, and f_{sm} the softmax output probabilities of the model for the male (M) and female (F) classes.

Thus a discrepancy between the TPR and TNR values indicate the model is better at identifying one gender than the other, and while a high APD value can indicate a discrepancy between the subgroup accuracies, it might also indicate that the model is more certain (i.e. higher softmax probabilities

when making predictions for one gender compared to the other one.

Table 4 displays the bias analysis results for our four models averaged over 5 runs. The average predictions differences are smaller for the ALL dataset compared to the slightly harder SUBJ dataset. Nevertheless, all models achieve almost-perfect accuracy on ALL and very high accuracy on SUBJ. While accuracies for male sentences is slightly higher for all models as the TPR values are higher than TNR values, the very small differences often within $\pm$ one standard deviation and in in the order of 0.01 leads us to believe there is no significant bias inherent in the models after fine-tuning.

G Experimental Pipeline

To obtain our results, for each of our models and datasets, we split the dataset into an 80/20 train/test split with balanced classes. We download pre-trained model weights (As discussed in section 5.4, for the experiments where we train models from scratch, we initialize the models randomly and then train them either on GECO-ALL or GECO-SUBJ for 50 epochs (1 epoch for RoBERTa to ensure high test accuracy without overfitting)) from Huggingface, and update all weights during training for 50 epochs for the GECO dataset, 5 for COMPAS, and 10 epochs for the Stereotypes dataset. We use the AdamW optimizer (Loshchilov 2017) with initial learning rate 0.001 and a linear learning rate schedule with 500 warm-up steps. Since our tasks are binary classification tasks, we use the binary cross-entropy loss. We observed that after one epoch of fine-tuning, the models perform hardly better than random guessing, so we trained each model for

Table 4: Bias analysis results after fine-tuning each model, averaged over 5 runs (TPR: true positive rate, TNR, true negative rate, APD: average prediction difference).

Dataset	**Model**	**TPR**	**TNR**	**APD**
ALL	BERT	$0.996_{0.001}$	$0.991_{0.001}$	$0.006_{0.000}$
	FairBERTa	$0.997_{0.000}$	$0.996_{0.001}$	$0.002_{0.004}$
	GPT2	$0.997_{0.000}$	$0.994_{0.000}$	$0.004_{0.000}$
	TinyBERT	$0.992_{0.003}$	$0.988_{0.001}$	$0.013_{0.001}$
	RoBERTa	$0.979_{0.020}$	$0.975_{0.036}$	$0.054_{0.031}$
SUBJ	BERT	$0.985_{0.001}$	$0.979_{0.001}$	$0.029_{0.002}$
	FairBERTa	$0.983_{0.003}$	$0.960_{0.008}$	$0.050_{0.005}$
	GPT2	$0.979_{0.003}$	$0.967_{0.003}$	$0.046_{0.005}$
	TinyBERT	$0.984_{0.001}$	$0.971_{0.004}$	$0.039_{0.001}$
	RoBERTa	$0.895_{0.076}$	$0.973_{0.009}$	$0.170_{0.098}$
COMPAS	BERT	$0.626_{0.000}$	$0.720_{0.000}$	$0.230_{0.000}$
	FairBERTa	$0.654_{0.000}$	$0.720_{0.000}$	$0.237_{0.000}$
	GPT2	$0.590_{0.000}$	$0.735_{0.000}$	$0.249_{0.000}$
	TinyBERT	$0.595_{0.035}$	$0.749_{0.021}$	$0.244_{0.017}$
	RoBERTa	$0.000_{0.000}$	$1.000_{0.000}$	$0.115_{0.002}$
Stereotypes	BERT	$0.994_{0.000}$	$0.997_{0.000}$	$0.003_{0.000}$
	FairBERTa	$0.994_{0.000}$	$1.000_{0.000}$	$0.006_{0.000}$
	GPT2	$0.997_{0.000}$	$1.000_{0.000}$	$0.002_{0.000}$
	TinyBERT	$1.000_{0.000}$	$1.000_{0.000}$	$0.000_{0.000}$
	RoBERTa	$1.000_{0.000}$	$0.800_{0.400}$	$0.001_{0.003}$

a larger number of epochs to achieve a high test accuracy without overfitting. We repeat this process 5 times for each model and dataset pair to report aggregate results.

H Additional Results

We display further results from our experiments in the tables and figures below. Tables 5, 6, 7, and 8 display further results on the number of runs resulting in significant disparity as well as the average effect sizes. Figures 3, 4, 5, 6, and 7 display box plots of the distributions of all evaluation scores obtained over all runs, including the ones not resulting in significant disparity for the remainder of our models.

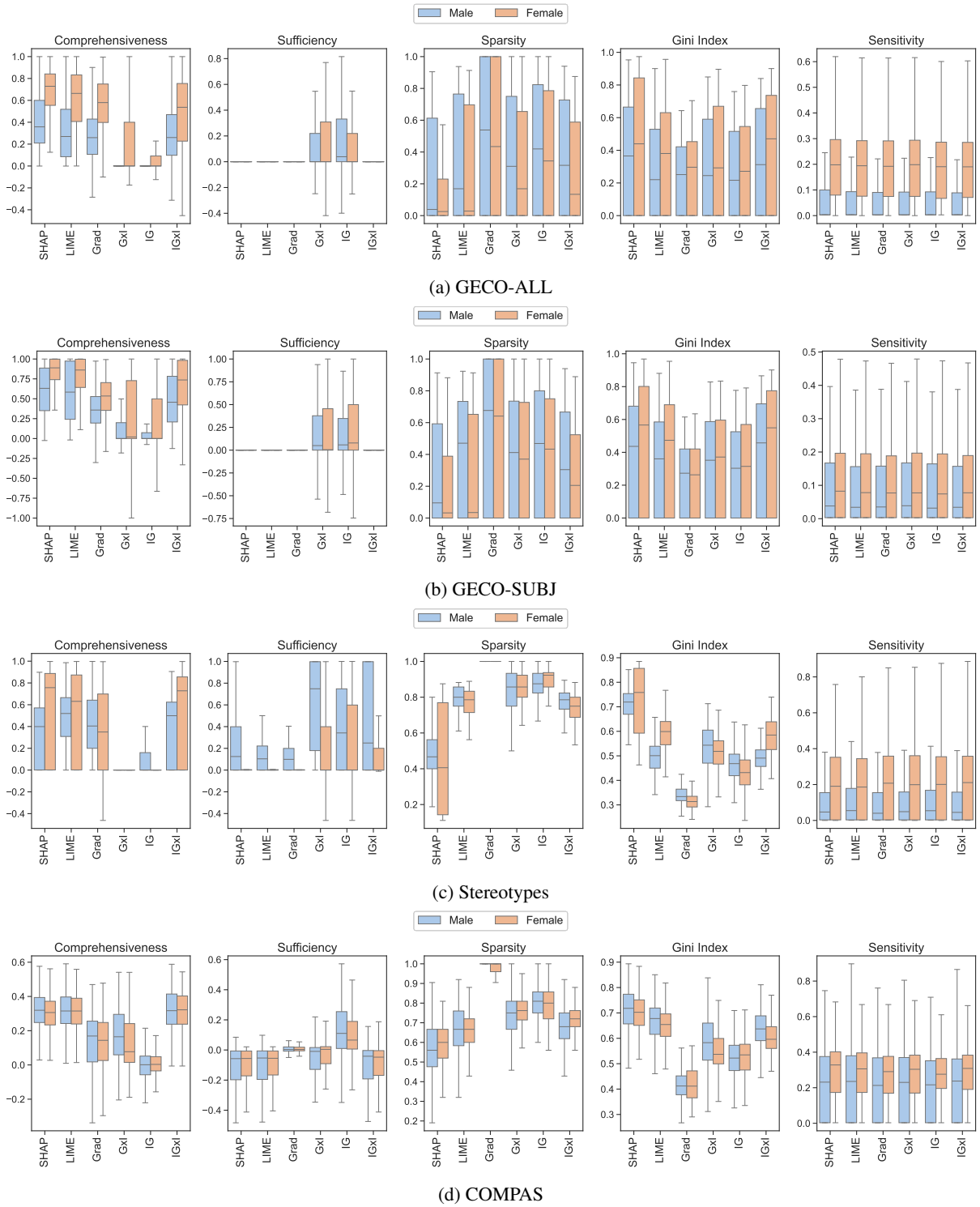


Figure 2: **Box-plots of evaluation scores** obtained over 5 runs for each using TinyBERT on GECO, Stereotypes, and COMPAS, including the runs not resulting in statistically significant disparity.

Table 5: **Occurence of disparity and effect sizes on GECO-ALL**. The numbers in parentheses display how many of the 5 runs resulted in statistically significant disparity, along with the effect sizes (Cohen's d). Bold font indicates considerable effect size ($|d| \geq 0.2$).

Model	**Method**	AOPC Compr. (↑)	AOPC Suff. (↓)	Soft Compr. (↑)	Soft Suff. (↓)	Gini Index (↑)	Sparsity (↓)	Sens. (↓)
BERT	Grad	**(5) -1.68±1.28**	(4) .16±.22	**(5) .60±1.32**	(0) NA	**(5) 1.00±1.92**	(5) -.06±.19	**(5) -.61±1.41**
	GxI	**(4) -.22±.29**	**(4) .27±.05**	**(5) .74±.95**	(0) NA	**(5) 1.06±1.89**	**(4) -.34±.49**	**(5) -.77±.92**
	IG	**(4) -.31±.35**	(5) -.11±.19	**(4) .88±1.15**	**(1) .24±.00**	**(5) 1.09±1.93**	**(4) -.27±.36**	**(4) -.87±1.19**
	IGxI	**(5) -.50±.78**	**(5) -1.12±.23**	**(4) 1.18±1.97**	**(5) .74±.22**	**(5) 1.02±1.99**	(5) -.11±.06	**(4) -1.08±1.97**
	LIME	**(5) -1.46±1.12**	**(5) -.61±.54**	**(5) .73±.95**	**(4) .59±.13**	**(5) 1.08±1.93**	(5) .04±.05	**(5) -.71±.97**
	SHAP	**(5) -1.42±1.38**	**(5) -1.01±1.78**	**(5) .67±.90**	**(5) .72±1.15**	**(5) .98±1.93**	(5) -.04±.02	**(4) -.84±.95**
FairBERTa	Grad	**(5) -2.41±.99**	**(4) .45±.21**	**(5) 17.86±31.01**	(0) NA	**(4) 1.00±.58**	**(4) -.30±.25**	**(5) -15.63±26.83**
	GxI	**(3) -.47±.09**	**(2) .28±.12**	**(5) 19.77±35.04**	(0) NA	**(4) .97±.56**	**(4) -.61±.56**	**(5) -11.54±18.47**
	IG	**(3) -.32±.19**	**(4) -.42±.13**	**(5) 10.82±16.97**	**(3) .22±.07**	**(4) .91±.59**	**(4) -.69±.27**	**(5) -9.99±15.34**
	IGxI	**(5) -.97±.22**	**(4) -1.90±.94**	**(5) 13.73±22.83**	**(4) 1.05±.51**	**(4) .94±.56**	**(4) -.31±.28**	**(5) -23.68±42.54**
	LIME	**(5) -1.65±.45**	**(3) -.60±.25**	**(5) 20.88±37.26**	**(3) .46±.14**	**(4) .99±.58**	**(4) -.28±.22**	**(5) -19.44±34.23**
	SHAP	**(5) -1.52±.38**	**(3) -1.37±.91**	**(5) 15.23±25.81**	**(3) 1.46±.94**	**(4) 1.01±.62**	**(4) -.29±.25**	**(5) -10.53±16.58**
GPT2	Grad	**(5) -.52±3.28**	(3) -.06±.66	(5) .06±1.32	**(3) .24±.48**	**(5) .58±1.90**	**(5) -.30±1.08**	(5) -.04±1.28
	GxI	(5) -.05±.84	**(5) .24±.20**	(4) .19±1.24	**(1) -.39±.00**	**(5) .51±1.26**	**(5) -.37±1.43**	**(4) -.24±1.26**
	IG	**(3) .33±.71**	(1) .15±.00	**(5) .28±1.11**	(0) NA	**(5) .44±.81**	**(5) -.32±1.48**	**(4) -.32±1.26**
	IGxI	**(5) -1.03±.90**	(5) .07±.67	**(5) .33±1.19**	(4) .14±.47	**(5) .26±.51**	**(5) .24±1.40**	**(5) -.32±1.23**
	LIME	**(5) -.60±.88**	**(5) -.27±.61**	(5) .13±1.26	**(2) .55±.18**	**(5) 1.06±1.92**	(5) -.19±1.15	**(5) -.20±1.34**
	SHAP	**(5) -.64±1.00**	(5) -.05±.18	**(5) .22±1.30**	(5) .15±.37	**(5) .42±1.25**	**(5) -.32±1.07**	(5) -.19±1.39
TinyBERT	Grad	**(5) -.84±1.02**	**(4) -.55±.21**	**(5) 3.47±3.67**	**(3) .35±.05**	**(5) -1.23±1.16**	(5) .04±.05	**(5) -3.37±3.70**
	GxI	**(4) -.54±.02**	**(5) -.77±.13**	**(5) 3.40±3.14**	**(5) .61±.14**	**(5) -1.21±1.17**	(5) .09±.32	**(4) -4.02±3.01**
	IG	**(4) -.45±.51**	**(2) -.91±.53**	**(4) 4.51±3.22**	**(2) .64±.34**	**(5) -1.22±1.16**	**(5) .27±.41**	**(4) -4.25±3.11**
	IGxI	**(5) -.29±.44**	**(5) -1.17±.53**	**(5) 3.25±3.95**	**(5) 1.23±.46**	**(5) -1.25±1.18**	(4) -.03±.02	**(5) -3.66±4.37**
	LIME	**(5) -.61±.44**	**(5) -1.24±.28**	**(5) 2.98±3.05**	**(5) .75±.15**	**(5) -1.22±1.17**	(4) .00±.05	**(5) -2.93±3.01**
	SHAP	**(4) -1.11±.14**	**(5) -2.01±.29**	**(5) 3.36±3.55**	**(5) 2.00±.26**	**(5) -1.18±1.16**	(3) .07±.04	**(5) -3.46±3.77**
RoBERTa	Grad	**(5) -1.25±1.67**	(0) NA	**(5) 3.66±2.87**	(0) NA	**(5) .48±.85**	**(5) -1.12±1.64**	**(5) -3.36±2.57**
	GxI	**(2) -.50±.12**	(2) -.01±.18	**(5) 2.27±2.08**	(1) .18±.00	**(5) .44±.90**	**(5) -1.12±1.56**	**(5) -2.00±2.00**
	IG	(3) -.18±.49	(0) NA	**(5) 2.17±2.03**	(0) NA	(5) -.20±1.93	**(5) -1.18±1.69**	**(5) -2.22±1.96**
	IGxI	**(4) -.53±.83**	**(1) -.28±.00**	**(5) 2.12±2.03**	(0) NA	**(5) .40±.97**	**(5) -.95±1.37**	**(5) -2.09±2.02**
	LIME	**(5) -2.68±3.36**	**(5) -.88±1.16**	**(5) 2.56±2.25**	**(5) .49±.63**	**(5) .49±.91**	**(5) -.90±.99**	**(5) -2.52±2.14**
	SHAP	**(5) -2.23±2.18**	**(5) -1.42±1.29**	**(4) 3.76±2.15**	**(5) 1.23±1.23**	**(5) .54±.89**	**(5) -.72±.94**	**(4) -4.05±2.31**

Table 6: **Occurence of disparity and effect sizes on GECO-SUBJ**. The numbers in parentheses display how many of the 5 runs resulted in statistically significant disparity, along with the effect sizes (Cohen's d). Bold font indicates considerable effect size ($|d| \geq 0.2$).

Model	**Method**	AOPC Compr. (↑)	AOPC Suff. (↓)	Soft Compr. (↑)	Soft Suff. (↓)	Gini Index (↑)	Sparsity (↓)	Sens. (↓)
BERT	Grad	(5) -.18±.43	**(5) .37±.17**	**(5) 1.54±1.22**	(0) NA	**(4) .74±.39**	(5) -.06±.11	**(5) -1.60±1.29**
	GxI	**(4) -.23±.25**	**(4) .30±.06**	**(5) 1.69±1.43**	(0) NA	**(3) .96±.19**	**(2) -.40±.43**	**(5) -1.71±1.50**
	IG	(5) -.17±.35	(3) .18±.05	**(5) 1.62±1.32**	(0) NA	**(4) .77±.41**	**(1) -.30±.00**	**(5) -1.62±1.37**
	IGxI	**(5) -.25±.41**	**(5) -.36±.10**	**(5) 1.47±1.30**	(0) NA	**(3) .94±.24**	(4) .06±.08	**(5) -1.58±1.33**
	LIME	**(5) -.32±.54**	**(3) -.63±.16**	**(5) 1.61±1.30**	**(3) .48±.04**	**(4) .76±.40**	(3) .02±.10	**(5) -1.53±1.25**
	SHAP	**(5) -.30±.58**	(4) -.11±.58	**(5) 1.61±1.34**	(4) .10±.39	**(4) .75±.40**	(2) .05±.09	**(5) -1.56±1.32**
FairBERTa	Grad	**(3) -.59±.23**	(4) .08±.62	**(5) -.21±1.07**	(0) NA	**(4) 1.02±.30**	(5) .04±.10	**(5) .20±1.08**
	GxI	**(3) -.52±.18**	(3) -.10±.27	(5) -.15±1.10	(1) .16±.00	**(4) 1.02±.29**	(3) .11±.45	(5) .18±1.08
	IG	(4) -.06±.24	**(2) -.25±.03**	(5) -.17±1.09	(2) .19±.04	**(4) 1.02±.26**	**(3) -.31±.04**	(5) .17±1.09
	IGxI	**(4) -.55±.17**	**(5) -.72±.37**	(5) -.20±1.08	**(5) .50±.24**	**(4) 1.03±.23**	(3) .13±.07	(5) .20±1.08
	LIME	**(5) -.64±.29**	**(4) -.54±.20**	(5) -.20±1.08	**(4) .47±.15**	**(4) 1.03±.25**	(3) .05±.09	(5) .20±1.08
	SHAP	**(5) -.55±.28**	**(3) -.32±.07**	(5) -.20±1.07	**(3) .28±.09**	**(4) 1.01±.27**	(4) .01±.09	(5) .17±1.08
GPT2	Grad	**(5) -1.16±1.87**	**(2) .73±.30**	**(5) .29±3.90**	**(3) -.37±.42**	**(5) -.46±2.33**	**(4) -.24±1.01**	**(5) -.34±3.87**
	GxI	**(4) -.43±.52**	**(2) .32±.11**	**(5) .42±3.78**	**(3) -.26±.12**	(5) -.15±1.79	**(5) -.89±1.51**	**(5) -.47±3.81**
	IG	**(4) -.62±.77**	(1) .18±.00	**(5) .62±3.54**	(0) NA	(5) .01±1.08	**(5) -.53±1.24**	**(5) -.64±3.62**
	IGxI	**(5) -.47±.93**	**(5) -.23±.53**	**(5) .61±3.68**	**(4) .34±.11**	(5) .05±.93	**(4) -.77±1.67**	**(5) -.57±3.66**
	LIME	**(4) -1.30±.66**	(4) .08±.47	**(5) .52±3.80**	(3) .08±.30	**(5) -.40±2.38**	(5) .06±1.00	**(5) -.57±3.82**
	SHAP	**(5) -1.02±1.15**	**(4) -.55±.59**	**(5) .49±3.53**	**(4) .47±.24**	**(5) -.23±1.90**	(4) -.12±.88	**(5) -.51±3.52**
TinyBERT	Grad	**(3) -1.17±.11**	(3) -.11±.19	**(5) .82±.72**	(0) NA	**(4) -.30±.24**	(3) -.01±.02	**(5) -.77±.72**
	GxI	**(5) -.47±.07**	(3) -.16±.28	**(5) .75±.69**	(0) NA	**(4) -.26±.23**	**(3) .31±.19**	**(5) -.74±.66**
	IG	**(5) -.24±.34**	**(4) -.59±.31**	**(5) .78±.68**	**(4) .45±.31**	**(4) -.26±.26**	**(4) -.21±.38**	**(5) -.74±.66**
	IGxI	**(3) -.71±.07**	**(5) -1.03±.76**	**(5) .84±.72**	**(5) 1.02±.84**	**(4) -.27±.23**	(5) -.01±.05	**(5) -.86±.74**
	LIME	**(5) -.64±.20**	**(5) -1.19±.58**	**(5) .79±.71**	**(5) .80±.35**	**(4) -.30±.25**	(5) .05±.05	**(5) -.81±.73**
	SHAP	**(5) -.70±.30**	**(5) -1.18±.55**	**(5) .77±.69**	**(5) .96±.53**	**(4) -.29±.26**	(5) .04±.04	**(5) -.80±.73**
RoBERTa	Grad	**(5) -1.63±.56**	(0) NA	**(5) 3.10±2.50**	(0) NA	**(5) .33±.25**	**(5) -1.01±.67**	**(5) -3.04±2.57**
	GxI	**(3) -.28±.22**	(0) NA	**(5) 2.29±1.90**	(0) NA	**(5) .35±.24**	**(5) -1.47±.58**	**(5) -2.25±1.83**
	IG	**(1) -.49±.00**	(0) NA	**(5) 2.31±1.77**	(0) NA	**(5) .38±.24**	**(5) -1.28±.33**	**(5) -2.33±1.81**
	IGxI	**(4) -1.18±.70**	(2) -.18±.35	**(5) 2.33±1.86**	(2) .06±.25	**(5) .32±.18**	**(5) -.95±.12**	**(5) -2.34±1.83**
	LIME	**(5) -1.44±.76**	**(5) -1.25±.60**	**(5) 2.60±2.02**	**(5) .90±.34**	**(5) .38±.19**	**(4) -.70±.31**	**(5) -2.65±1.97**
	SHAP	**(5) -1.62±.75**	**(5) -1.16±.49**	**(5) 2.67±2.00**	**(4) 1.05±.36**	**(5) .33±.24**	**(5) -.56±.30**	**(5) -2.69±1.96**

Table 7: **Occurence of disparity and effect sizes on COMPAS**. The numbers in parentheses display how many of the 5 runs resulted in statistically significant disparity, along with the effect sizes (Cohen's d). Bold font indicates considerable effect size ($|d| \geq 0.2$).

Model	**Method**	AOPC Compr. (↑)	AOPC Suff. (↓)	Soft Compr. (↑)	Soft Suff. (↓)	Gini Index (↑)	Sparsity (↓)	Sens. (↓)
BERT	Grad	**(5) -.34±.11**	(2) -.17±.01	**(3) -.41±.02**	**(4) .33±.09**	(3) .15±.47	**(5) .45±.31**	**(4) -.37±.18**
	GxI	**(5) -.35±.51**	(3) .05±.38	**(3) -.32±.08**	**(3) .33±.08**	**(3) -.87±.47**	**(3) .80±.54**	**(2) -.56±.13**
	IG	**(2) -.31±.17**	**(1) -.21±.00**	**(3) -.37±.06**	**(3) .34±.09**	(4) -.15±.19	**(3) .32±.08**	**(3) -.40±.09**
	IGxI	**(5) -.45±.12**	**(5) -.24±.09**	**(3) -.35±.07**	**(4) .31±.11**	(2) .11±.43	(4) .19±.30	**(3) -.46±.16**
	LIME	**(5) -.43±.07**	**(4) -.27±.09**	**(3) -.37±.08**	**(3) .34±.10**	**(3) .26±.08**	(2) .00±.30	**(3) -.47±.16**
	SHAP	**(5) -.42±.13**	**(4) -.28±.08**	**(2) -.39±.05**	**(3) .33±.10**	(4) -.06±.24	**(5) .27±.49**	**(3) -.45±.16**
FairBERTa	Grad	**(2) -.35±.03**	**(1) .20±.00**	(2) -.05±.33	(2) .05±.33	(4) .10±.24	(0) NA	(4) -.00±.32
	GxI	**(2) -.33±.01**	(3) .12±.25	(2) -.09±.29	**(1) .38±.00**	**(4) .30±.07**	(3) -.19±.01	(4) .06±.38
	IG	(5) .12±.23	**(4) -.29±.13**	(2) -.08±.30	(2) .06±.32	(4) .06±.27	(2) .08±.25	(2) -.08±.28
	IGxI	(3) .11±.34	**(2) -.37±.04**	**(1) -.38±.00**	(2) .08±.30	(2) .15±.54	(4) .15±.30	**(3) -.21±.32**
	LIME	**(1) -.37±.00**	**(1) -.40±.00**	**(1) -.38±.00**	**(1) .38±.00**	(5) .11±.22	(2) -.02±.30	(3) .04±.35
	SHAP	**(1) -.37±.00**	**(1) -.36±.00**	(2) -.08±.30	(2) .08±.30	(4) .18±.29	(2) -.10±.41	(3) .05±.39
GPT2	Grad	**(5) -.43±.17**	**(4) .33±.06**	**(1) -.26±.00**	(4) -.02±.13	(3) .00±.18	**(1) .21±.00**	**(4) -.39±.03**
	GxI	**(5) -.34±.10**	**(4) -.41±.52**	(3) .10±.28	**(2) -.22±.15**	**(5) -.22±.10**	(4) -.02±.39	**(4) -.37±.04**
	IG	(4) -.14±.06	**(4) -.30±.39**	(5) .08±.18	(5) -.09±.17	**(1) .49±.00**	(4) -.03±.23	**(3) -.46±.12**
	IGxI	**(4) -.54±.10**	(3) .10±.21	(5) .07±.18	(5) -.08±.19	**(5) -.23±.54**	**(4) .35±.49**	**(1) -.31±.00**
	LIME	**(5) -.48±.05**	**(2) -.26±.00**	**(1) .35±.00**	(4) -.15±.11	**(4) .24±.28**	**(3) -.27±.05**	**(4) -.41±.03**
	SHAP	**(4) -.44±.03**	(3) -.17±.29	(0) NA	(0) NA	**(5) .21±.21**	**(4) -.55±.06**	**(4) -.40±.06**
TinyBERT	Grad	**(4) .40±.39**	(4) -.06±.31	**(1) -.51±.00**	**(1) .49±.00**	**(2) .40±.20**	(5) .02±.26	**(4) -.50±.12**
	GxI	**(5) .53±.05**	**(2) -.51±.20**	**(1) -.48±.00**	**(1) .50±.00**	**(5) .59±.19**	**(5) -.40±.08**	**(4) -.53±.11**
	IG	**(1) .40±.00**	(5) .18±.30	**(1) -.46±.00**	**(1) .49±.00**	(2) .19±.39	(2) .03±.23	**(5) -.44±.07**
	IGxI	**(4) .32±.26**	(4) .05±.22	**(1) -.52±.00**	**(1) .51±.00**	**(5) .59±.18**	**(4) -.66±.10**	**(5) -.48±.15**
	LIME	(5) .14±.44	(4) .03±.22	**(1) -.51±.00**	**(1) .53±.00**	**(4) .29±.10**	**(1) -.38±.00**	**(3) -.51±.05**
	SHAP	**(3) .69±.00**	(4) .01±.22	**(1) -.49±.00**	**(1) .50±.00**	**(4) .28±.07**	(5) -.09±.20	**(5) -.51±.10**
RoBERTa	Grad	(3) .11±.32	(2) -.04±.18	**(1) -.33±.00**	**(2) .27±.06**	(3) -.08±.26	(0) NA	(0) NA
	GxI	(1) .07±.00	(2) -.07±.28	(2) -.13±.19	**(1) .33±.00**	**(1) .41±.00**	**(1) -.32±.00**	(0) NA
	IG	**(1) -.48±.00**	(3) .08±.24	**(1) -.33±.00**	**(1) .33±.00**	**(1) -.34±.00**	**(1) -.21±.00**	(0) NA
	IGxI	(3) .00±.22	**(3) .23±.27**	**(1) -.33±.00**	(2) .14±.18	**(3) -.21±.41**	(3) -.00±.41	(0) NA
	LIME	**(1) -.32±.00**	(1) .10±.00	**(1) -.33±.00**	**(1) .33±.00**	(1) -.11±.00	(0) NA	(0) NA
	SHAP	(2) .03±.35	(3) .17±.36	**(1) -.33±.00**	**(1) .33±.00**	**(2) .27±.02**	(4) -.14±.27	(0) NA

Table 8: **Occurence of disparity and effect sizes on Stereotypes**. The numbers in parentheses display how many of the 5 runs resulted in statistically significant disparity, along with the effect sizes (Cohen's d). Bold font indicates considerable effect size ($|d| \geq 0.2$).

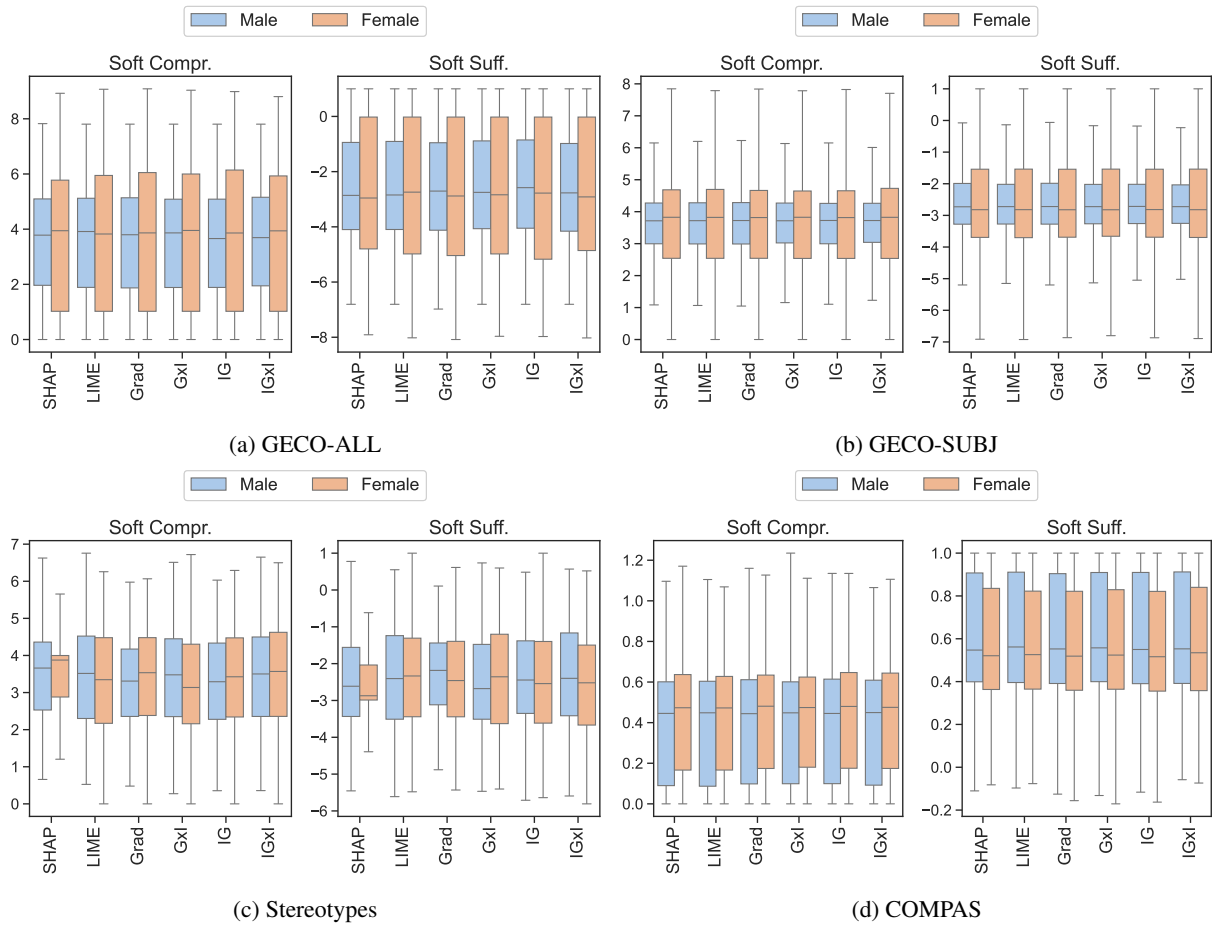
Model	**Method**	AOPC Compr. (↑)	AOPC Suff. (↓)	Soft Compr. (↑)	Soft Suff. (↓)	Gini Index (↑)	Sparsity (↓)	Sens. (↓)
BERT	Grad	**(5) .25±.18**	(5) -.13±.23	(0) NA	(0) NA	**(5) .32±.04**	(0) NA	**(3) .24±.29**
	GxI	**(2) .51±.00**	**(2) .54±.00**	(0) NA	(0) NA	**(3) .54±.00**	**(5) -.34±.11**	**(2) .24±.34**
	IG	**(5) .35±.06**	**(5) .22±.11**	(0) NA	(0) NA	(0) NA	(0) NA	**(4) .29±.16**
	IGxI	**(2) .43±.00**	**(5) .27±.48**	(0) NA	(0) NA	(0) NA	(0) NA	**(4) .33±.16**
	LIME	**(2) .53±.00**	**(5) .29±.30**	(0) NA	(0) NA	**(5) .33±.15**	**(2) -.55±.00**	**(2) .33±.21**
	SHAP	**(5) .36±.17**	(5) -.11±.55	(0) NA	(0) NA	**(2) .38±.00**	**(5) -.36±.08**	(5) .19±.20
FairBERTa	Grad	(0) NA	(0) NA	(0) NA	(0) NA	**(5) -.38±.00**	(0) NA	**(1) -.38±.00**
	GxI	(0) NA	**(5) .58±.00**	(0) NA	(0) NA	(0) NA	(0) NA	**(1) -.48±.00**
	IG	(0) NA	**(5) .30±.00**	(0) NA	(0) NA	(0) NA	(0) NA	**(3) -.37±.03**
	IGxI	(0) NA	**(5) .41±.00**	(0) NA	(0) NA	**(5) .78±.00**	**(5) -.37±.00**	(1) -.17±.00
	LIME	**(5) .44±.00**	**(5) -.37±.00**	(0) NA	(0) NA	(0) NA	(0) NA	**(1) -.50±.00**
	SHAP	**(5) -.29±.00**	(5) -.05±.00	(0) NA	(0) NA	**(5) .54±.00**	**(5) -.42±.00**	**(3) -.38±.03**
GPT2	Grad	**(3) .28±.00**	(3) .11±.00	(2) -.12±.00	(2) .19±.00	**(5) .59±.56**	(0) NA	**(3) -.22±.09**
	GxI	**(2) .66±.00**	(5) -.06±.30	**(2) -.32±.00**	**(2) .42±.00**	**(5) .42±.50**	(0) NA	(4) -.18±.22
	IG	(0) NA	**(3) .36±.00**	(0) NA	**(2) .23±.00**	**(3) -.28±.00**	(0) NA	(2) -.18±.04
	IGxI	(3) .18±.00	**(5) .54±.17**	(0) NA	**(2) .27±.00**	**(2) -.86±.00**	(5) -.01±.29	(1) .11±.00
	LIME	**(3) -.31±.00**	**(3) .21±.00**	**(2) -.21±.00**	(2) .15±.00	**(3) -.52±.00**	(3) .15±.00	**(3) -.40±.09**
	SHAP	**(5) -.43±.49**	**(5) .78±.06**	(0) NA	(0) NA	**(5) -.33±.35**	**(3) .28±.00**	**(2) -.52±.27**
TinyBERT	Grad	**(5) .22±.00**	**(5) .90±.00**	(0) NA	(5) .15±.00	**(5) .74±.00**	(0) NA	**(5) -.78±.21**
	GxI	(0) NA	**(5) .85±.00**	(0) NA	(0) NA	**(5) .28±.00**	(0) NA	**(5) -.72±.19**
	IG	(0) NA	**(5) .30±.00**	(0) NA	(0) NA	**(5) .46±.00**	**(5) -.25±.00**	**(5) -.69±.19**
	IGxI	**(5) -.22±.00**	**(5) .77±.00**	(0) NA	(0) NA	**(5) -1.38±.00**	**(5) .64±.00**	**(5) -.78±.22**
	LIME	(5) -.01±.00	**(5) .52±.00**	(0) NA	(0) NA	**(5) -1.17±.00**	**(5) .40±.00**	**(5) -.62±.13**
	SHAP	**(5) -.40±.00**	**(5) .66±.00**	(0) NA	(0) NA	(5) -.18±.00	(0) NA	**(5) -.72±.16**
RoBERTa	Grad	(5) .17±.37	**(2) .28±.07**	(4) -.02±.02	(4) .02±.02	(4) -.02±.28	(0) NA	(0) NA
	GxI	(1) .05±.00	**(2) .35±.03**	(1) .01±.00	(1) -.02±.00	(3) .11±.23	(1) .10±.00	(0) NA
	IG	**(2) .36±.13**	(4) .07±.21	(1) -.01±.00	(3) .04±.04	**(1) .23±.00**	(2) -.06±.17	(0) NA
	IGxI	**(3) .35±.17**	(4) .15±.29	(1) .03±.00	(1) -.05±.00	**(2) .22±.52**	**(1) -.53±.00**	(0) NA
	LIME	(5) .17±.15	(4) .13±.52	(4) -.03±.04	(4) .03±.06	(4) -.16±.28	(4) .06±.25	(3) .18±.02
	SHAP	(4) .04±.28	(5) .11±.49	(4) .00±.01	(1) .01±.00	(5) .15±.41	**(4) -.23±.49**	**(1) .21±.00**

Table 9: **Number of runs out of five resulting in statistically significant disparity** on the GECO datasets. Cell colors indicate which gender has better evaluation scores for each metric (blue: males, red: females). Bold font further indicates considerable effect size (Cohen's d, with $|d| \geq 0.2$). Metrics are grouped based on the evaluation property they measure.

		GECO-ALL							GECO-SUBJ						
		Faithf.				Comp.		Rbst.	Faithf.				Comp.		Rbst.
Model	**Method**	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.
TinyBERT	Grad	**5**	5	**5**	**5**	**4**	**3**	**5**	**3**	3	**5**	**5**	3	0	**4**
	GxI	**4**	5	**4**	**5**	**5**	**5**	**5**	**5**	**3**	**5**	**5**	3	0	**4**
	IG	**4**	**5**	**4**	**4**	**2**	**2**	**5**	**5**	**4**	**5**	**5**	**4**	**4**	**4**
	IGxI	**5**	4	**5**	**5**	**5**	**5**	**5**	**3**	5	**5**	**5**	**5**	**5**	**4**
	LIME	**5**	4	**5**	**5**	**5**	**5**	**5**	**5**	5	**5**	**5**	**5**	**5**	**4**
	SHAP	**4**	3	**5**	**5**	**5**	**5**	**5**	**5**	5	**5**	**5**	**5**	**5**	**4**
GPT2	Grad	**5**	**5**	5	5	3	**3**	**5**	**5**	**4**	**5**	**5**	**2**	**3**	**5**
	GxI	5	**5**	**4**	4	**5**	**1**	**5**	**4**	**5**	**5**	**5**	**2**	**3**	5
	IG	**3**	**5**	**4**	**5**	1	0	**5**	**4**	**5**	**5**	**5**	1	0	5
	IGxI	**5**	**5**	**5**	**5**	5	4	**5**	**5**	**4**	**5**	**5**	**5**	**4**	5
	LIME	**5**	5	**5**	5	**5**	**2**	**5**	**4**	5	**5**	**5**	4	3	**5**
	SHAP	**5**	**5**	5	**5**	5	5	**5**	**5**	4	**5**	**5**	**4**	**4**	**5**
BERT	Grad	**5**	5	**5**	**5**	4	0	**5**	5	5	**5**	**5**	**5**	0	**4**
	GxI	**4**	**4**	**5**	**5**	**4**	0	**5**	**4**	**2**	**5**	**5**	**4**	0	**3**
	IG	**4**	**4**	**4**	**4**	5	**1**	**5**	5	**1**	**5**	**5**	3	0	**4**
	IGxI	**5**	5	**4**	**4**	**5**	**5**	**5**	**5**	4	**5**	**5**	**5**	0	**3**
	LIME	**5**	5	**5**	**5**	**5**	**4**	**5**	**5**	3	**5**	**5**	**3**	**3**	**4**
	SHAP	**5**	5	**4**	**5**	**5**	**5**	**5**	**5**	2	**5**	**5**	4	4	**4**
FairBERTa	Grad	**5**	**4**	**5**	**5**	**4**	0	**4**	**3**	5	**5**	**5**	4	0	**4**
	GxI	**3**	**4**	**5**	**5**	**2**	0	**4**	**3**	3	5	5	3	1	**4**
	IG	**3**	**4**	**5**	**5**	**4**	**3**	**4**	4	**3**	5	5	**2**	2	**4**
	IGxI	**5**	**4**	**5**	**5**	**4**	**4**	**4**	**4**	3	5	5	**5**	**5**	**4**
	LIME	**5**	**4**	**5**	**5**	**3**	**3**	**4**	**5**	3	5	5	**4**	**4**	**4**
	SHAP	**5**	**4**	**5**	**5**	**3**	**3**	**4**	**5**	4	5	5	**3**	**3**	**4**
RoBERTa	Grad	**5**	**5**	**5**	**5**	0	0	**5**	**5**	**5**	**5**	**5**	0	0	**5**
	GxI	**2**	**5**	**5**	**5**	2	1	**5**	**3**	**5**	**5**	**5**	0	0	**5**
	IG	3	**5**	**5**	**5**	0	0	5	**1**	**5**	**5**	**5**	0	0	**5**
	IGxI	**4**	**5**	**5**	**5**	**1**	0	**5**	**4**	**5**	**5**	**5**	2	2	**5**
	LIME	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**5**	**5**	**5**	**5**	**5**
	SHAP	**5**	**5**	**4**	**4**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**5**

Table 10: **Training from scratch.** Counts of significant disparity with colors indicating direction of effect (red=female scores higher, blue=male scores higher)

		GECO-ALL							GECO-SUBJ						
		Faithf.				Comp.		Rbst.	Faithf.				Comp.		Rbst.
Model	**Method**	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.
BERT	Grad	**5**	4	**4**	**4**	**4**	**2**	**4**	**5**	4	**4**	**5**	3	2	**2**
	GxI	**5**	**4**	**4**	**5**	4	5	**4**	**5**	**5**	**3**	**4**	4	4	**2**
	IG	**3**	**3**	**5**	**5**	**4**	**3**	**4**	**3**	**4**	**5**	**3**	**5**	**5**	**2**
	IGxI	**4**	4	**5**	**5**	**4**	**4**	**4**	**5**	**5**	**3**	**3**	**4**	**4**	**3**
	LIME	**4**	**4**	**5**	**5**	**3**	4	**4**	**5**	5	**3**	**3**	**5**	**5**	**3**
	SHAP	**5**	**4**	**5**	**5**	5	5	**4**	**4**	5	**5**	**5**	**5**	**4**	**3**
GPT2	Grad	**5**	**4**	**5**	**5**	2	**2**	4	**5**	**3**	**5**	**5**	**4**	**4**	**1**
	GxI	**4**	**5**	**5**	**5**	**5**	**4**	3	**5**	**4**	**5**	**5**	**3**	4	**1**
	IG	**5**	**4**	**5**	**5**	**5**	**5**	2	**5**	**5**	**5**	**5**	**5**	**3**	**2**
	IGxI	**5**	**3**	**5**	**5**	**5**	**5**	3	**5**	**5**	**5**	**5**	**4**	**4**	**2**
	LIME	**5**	**5**	**5**	**5**	**4**	**3**	2	**5**	5	**5**	**5**	4	3	**2**
	SHAP	**5**	**5**	**5**	**5**	**5**	**4**	3	**5**	5	**5**	**5**	**4**	**3**	2



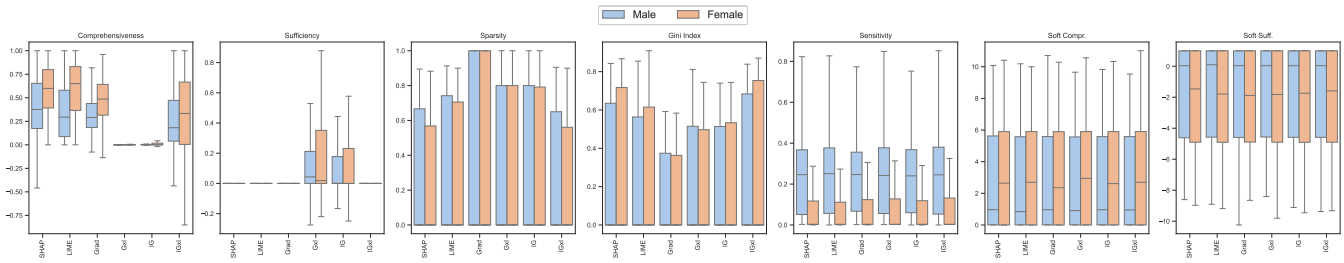
(a) GECO-ALL (b) GECO-SUBJ

(c) Stereotypes (d) COMPAS

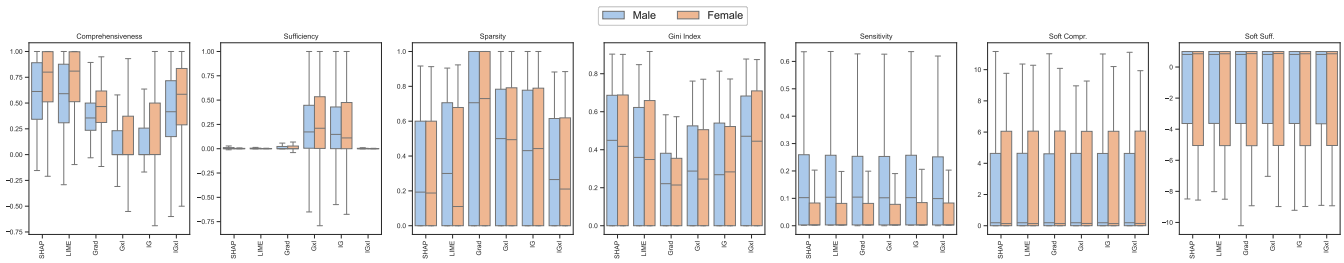
Figure 3: **Box-plots of the soft comprehensiveness and sufficiency metrics** obtained over 5 runs for each using TinyBERT on GECO-ALL, Stereotypes, and COMPAS, including the runs not resulting in statistically significant disparity.

Table 11: **Number of runs out of five resulting in statistically significant disparity** on the COMPAS and Stereotypes datasets. Cell colors indicate which gender has better evaluation scores for each metric (blue: males, red: females). Bold font further indicates considerable effect size (Cohen's d, with $|d| \geq 0.2$). Metrics are grouped based on the evaluation property they measure.

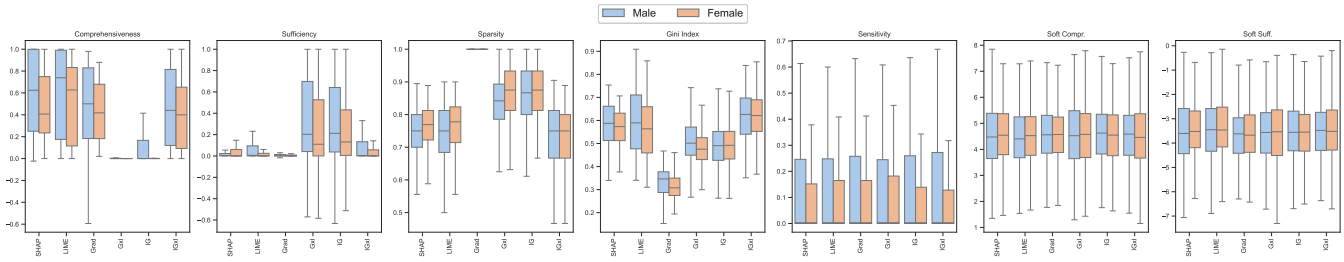
		COMPAS							Stereotypes						
		Faithf.				Comp.		Rbst.	Faithf.				Comp.		Rbst.
Model	**Method**	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.	Compr.	Suff.	Soft Compr.	Soft Suff.	Gini	Spars.	Sens.
TinyBERT	Grad	**4**	4	**1**	**1**	**2**	5	**4**	**5**	**5**	0	5	**5**	0	**5**
	GxI	**5**	**2**	**1**	**1**	**5**	**5**	**4**	0	**5**	0	0	**5**	0	**5**
	IG	**1**	5	**1**	**1**	2	2	**5**	0	**5**	0	0	**5**	**5**	**5**
	IGxI	**4**	4	**1**	**1**	**5**	**4**	**5**	**5**	**5**	0	0	**5**	**5**	**5**
	LIME	5	4	**1**	**1**	**4**	**1**	**3**	5	**5**	0	0	**5**	**5**	**5**
	SHAP	**3**	4	**1**	**1**	**4**	5	**5**	**5**	**5**	0	0	5	0	**5**
GPT2	Grad	**5**	**4**	**1**	4	3	**1**	**4**	**3**	3	2	2	**5**	0	**3**
	GxI	**5**	**4**	3	**2**	**5**	4	**4**	**2**	5	**2**	**2**	**5**	0	4
	IG	4	**4**	5	5	**1**	4	**3**	0	**3**	0	**2**	**3**	0	2
	IGxI	**4**	3	5	5	**5**	**4**	**1**	3	**5**	0	**2**	**2**	5	1
	LIME	**5**	**2**	**1**	4	**4**	**3**	**4**	**3**	**3**	**2**	2	**3**	3	**3**
	SHAP	**4**	3	0	0	**5**	**4**	**4**	**5**	**5**	0	0	**5**	**3**	**2**
BERT	Grad	**5**	2	**3**	**4**	3	**5**	**4**	**5**	5	0	0	**5**	0	**3**
	GxI	**5**	3	**3**	**3**	**3**	**3**	**2**	**2**	**2**	0	0	**3**	**5**	**2**
	IG	**2**	**1**	**3**	**3**	4	**3**	**3**	**5**	**5**	0	0	0	0	**4**
	IGxI	**5**	**5**	**3**	**4**	2	4	**3**	**2**	**5**	0	0	0	0	**4**
	LIME	**5**	**4**	**3**	**3**	**3**	2	**3**	**2**	**5**	0	0	**5**	**2**	**2**
	SHAP	**5**	**4**	**2**	**3**	4	**5**	**3**	**5**	5	0	0	**2**	**5**	5
FairBERTa	Grad	**2**	**1**	2	2	4	0	4	0	0	0	0	**5**	0	**1**
	GxI	**2**	3	2	**1**	**4**	3	4	0	**5**	0	0	0	0	**1**
	IG	5	**4**	2	2	4	2	2	0	**5**	0	0	0	0	**3**
	IGxI	3	**2**	**1**	2	2	4	**3**	0	**5**	0	0	**5**	**5**	1
	LIME	**1**	**1**	**1**	**1**	5	2	3	**5**	**5**	0	0	0	0	**1**
	SHAP	**1**	**1**	2	2	4	2	3	**5**	5	0	0	**5**	**5**	**3**
RoBERTa	Grad	3	2	**1**	**2**	3	0	0	5	**2**	4	4	4	0	0
	GxI	1	2	2	**1**	**1**	**1**	0	1	**2**	1	1	3	1	0
	IG	**1**	3	**1**	**1**	**1**	**1**	0	**2**	4	1	3	**1**	2	0
	IGxI	3	**3**	**1**	2	**3**	3	0	**3**	4	1	1	**2**	**1**	0
	LIME	**1**	1	**1**	**1**	1	0	0	5	4	4	4	4	4	3
	SHAP	2	3	**1**	**1**	**2**	4	0	4	5	4	1	5	**4**	**1**



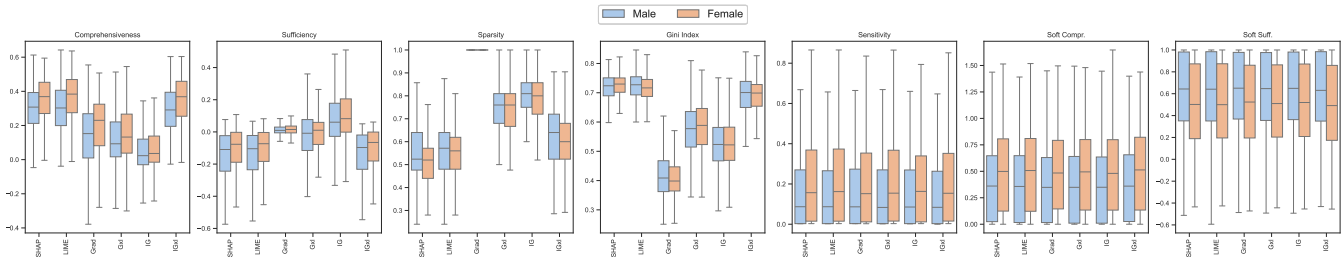
(a) GECO-ALL



(b) GECO-SUBJ

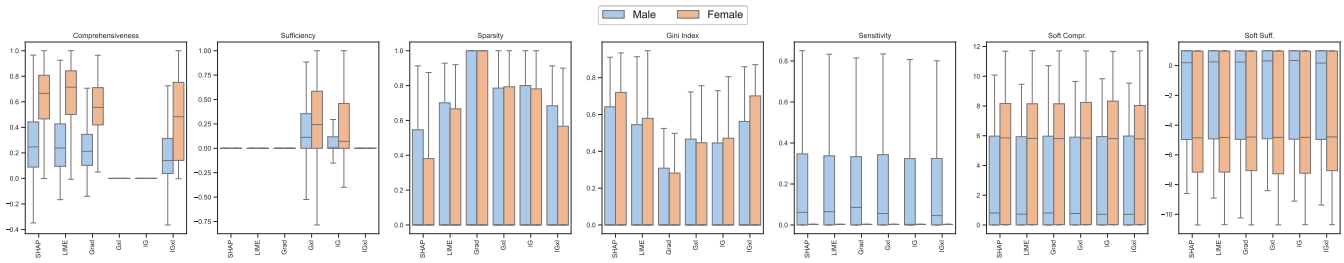


(c) Stereotypes

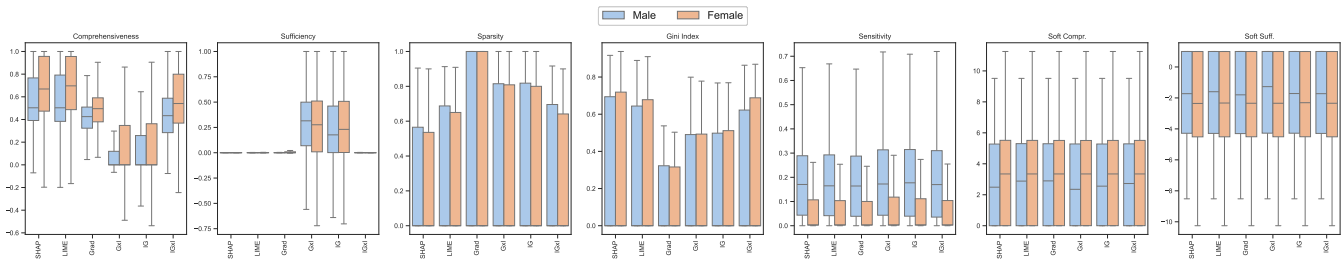


(d) COMPAS

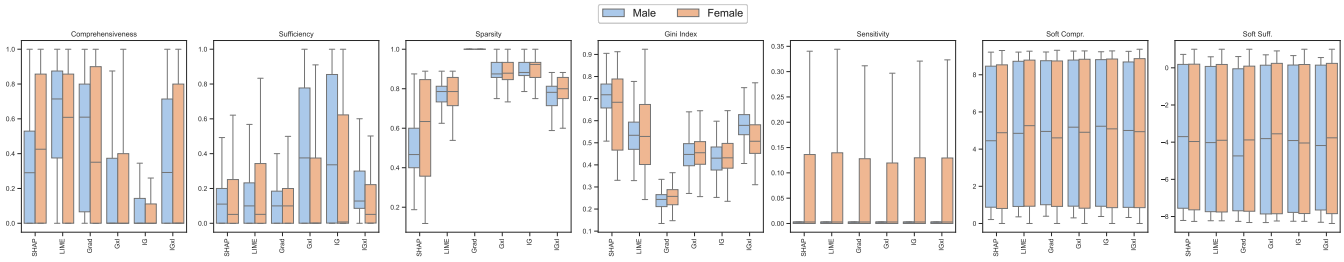
Figure 4: **Box-plots of evaluation scores** obtained over 5 runs for each using BERT on our datasets, including the runs not resulting in statistically significant disparity.



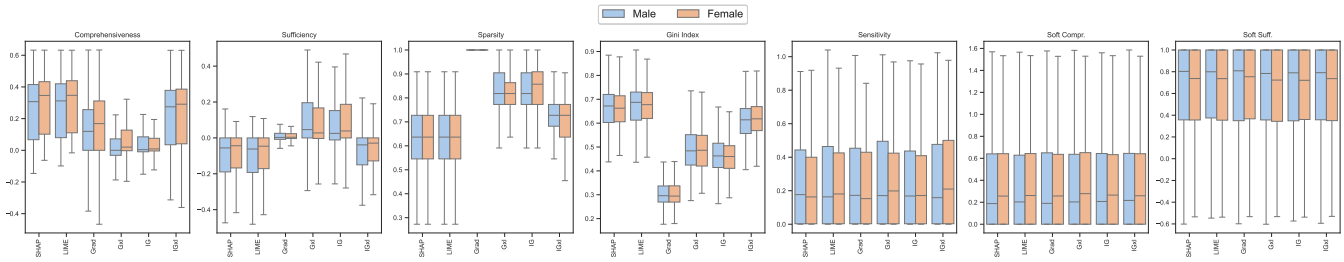
(a) GECO-ALL



(b) GECO-SUBJ

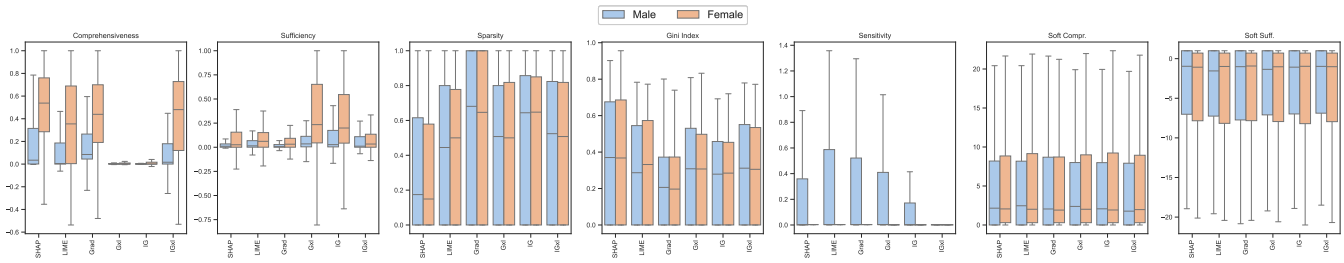


(c) Stereotypes

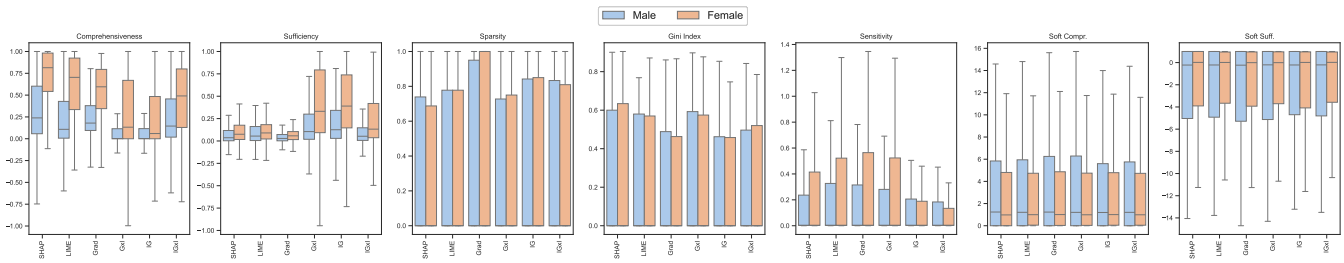


(d) COMPAS

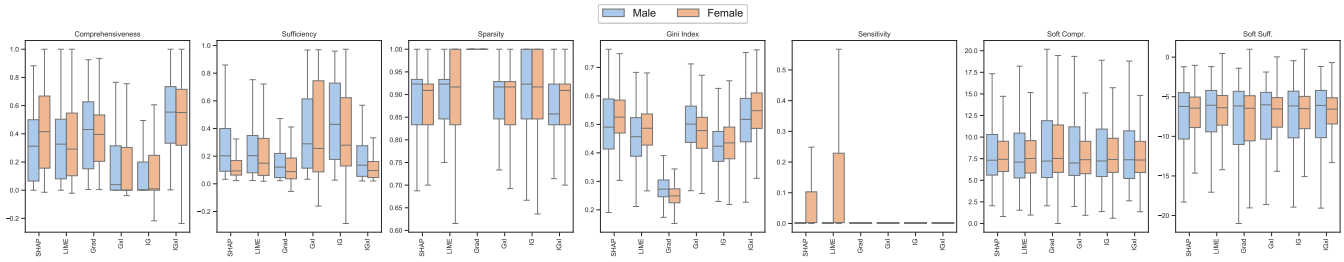
Figure 5: **Box-plots of evaluation scores** obtained over 5 runs for each using FairBERTa on our datasets, including the runs not resulting in statistically significant disparity.



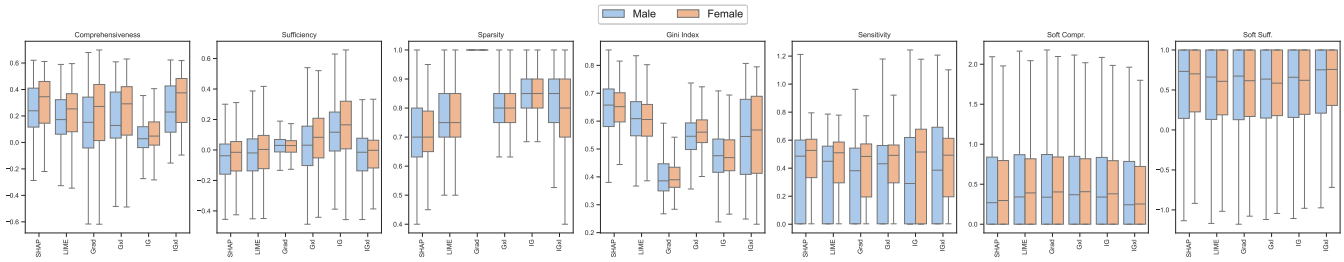
(a) GECO-ALL



(b) GECO-SUBJ

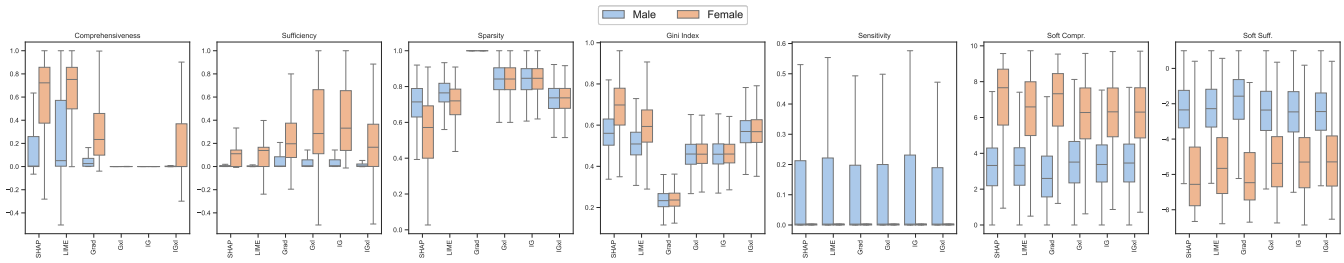


(c) Stereotypes

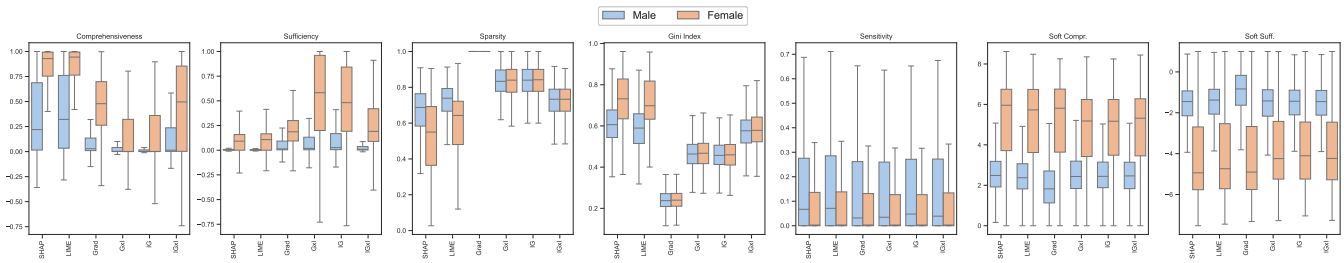


(d) COMPAS

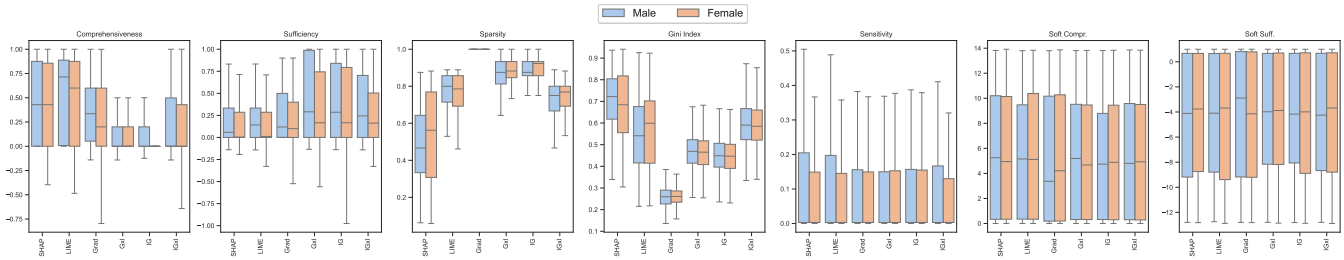
Figure 6: **Box-plots of evaluation scores** obtained over 5 runs for each using GPT-2 on our datasets, including the runs not resulting in statistically significant disparity.



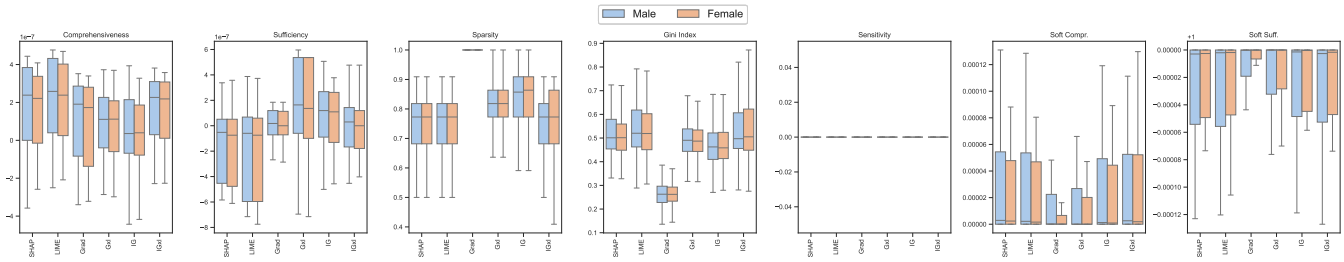
(a) GECO-ALL



(b) GECO-SUBJ



(c) Stereotypes



(d) COMPAS

Figure 7: **Box-plots of evaluation scores** obtained over 5 runs for each using RoBERTa on our datasets, including the runs not resulting in statistically significant disparity.