Why Don't Prompt-Based Fairness Metrics Correlate?

Abdelrahman Zayed[1,2], **Gonçalo Mordido**[1,2], **Ioana Baldini**[3], **Sarath Chandar**[1,2,4]
[1]Mila - Quebec AI Institute
[2]Polytechnique Montreal
[3]IBM Research
[4]Canada CIFAR AI Chair
{zayedabd,goncalo-filipe.torcato-mordido,sarath.chandar}@mila.quebec,
{ioana}@us.ibm.com

Abstract

The widespread use of large language models has brought up essential questions about the potential biases these models might learn. This led to the development of several metrics aimed at evaluating and mitigating these biases. In this paper, we first demonstrate that prompt-based fairness metrics exhibit poor agreement, as measured by correlation, raising important questions about the reliability of fairness assessment using prompts. Then, we outline six relevant reasons why such a low correlation is observed across existing metrics. Based on these insights, we propose a method called Correlated Fairness Output (CAIRO) to enhance the correlation between fairness metrics. CAIRO augments the original prompts of a given fairness metric by using several pre-trained language models and then selects the combination of the augmented prompts that achieves the highest correlation across metrics. We show a significant improvement in Pearson correlation from 0.3 and 0.18 to 0.90 and 0.98 across metrics for gender and religion biases, respectively. Our code is available at https://github.com/chandar-lab/CAIRO.

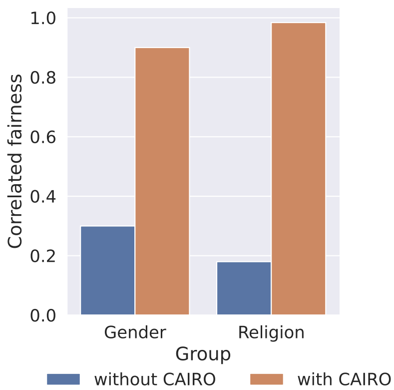

Figure 1: Correlated fairness between fairness metrics on gender and religion bias with and without CAIRO.

1 Introduction

The success of Transformers (Vaswani et al., 2017) sparked a revolution in language models, allowing them to reach unprecedented levels of performance across various tasks (Rajpurkar et al., 2016; Wang et al., 2018; Rajpurkar et al., 2018; Li et al., 2020a,b; Zhang et al., 2020; Yu et al., 2020; Liu et al., 2022). This advancement has significantly contributed to the extensive use of language models in everyday life. However, the potential risks of deploying models that exhibit unwanted social bias cannot be overlooked[1]. Consequently, there has been an increase in the number of methods aimed at reducing bias (Lu et al., 2020; Dhamala et al., 2021; Attanasio et al., 2022; Zayed et al., 2023, 2024), which rely on fairness assessment metrics to evaluate their efficacy. As different methods use different metrics and as new metrics are introduced, agreement across metrics is instrumental to properly quantify the advancements in bias mitigation. Such agreement would also indicate that existing metrics are indeed measuring similar model traits (*e.g.* bias towards a specific social group), as originally intended.

Fairness metrics can be broadly classified into embedding-based, probability-based, and prompt-based metrics, which will be discussed in Section 2. The lack of correlation between traditional fairness metrics has been previously noticed, for both embedding-based and probability-based metrics (Delobelle et al., 2022; Cao et al., 2022b). The lack of alignment of such metrics with the bias of downstream tasks has also been highlighted in previous works (Goldfarb-Tarrant et al., 2021; Orgad et al., 2022; Steed et al., 2022; Kaneko et al., 2022; Gallegos et al., 2023; Cabello et al., 2023; Orgad and Belinkov, 2023). In this work, we focus on prompt-based fairness metrics in generative contexts that use prompt continuations to assess model

[1]We refer to unwanted social bias as bias in short.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9002–9019
August 11-16, 2024 ©2024 Association for Computational Linguistics

bias, namely: BOLD (Dhamala et al., 2021), HolisticBias (Smith et al., 2022), and HONEST (Nozza et al., 2021). Such prompt-based metrics (Gallegos et al., 2023) rely on providing a model with prompts that reference various groups to then measure its hostility (*e.g.* toxicity) towards each group. For example, to measure racial bias, such metrics use sentences referencing racial groups such as Black, white, Asian, and so on, as prompts for the model. Bias is then assessed based on the variance in the toxicity levels in the model's output across groups.

In this study, we show that popular prompt-based fairness metrics do not agree out-of-the-box (Figure 1), which can be in part explained by the high volatility of language models to prompts (Poerner et al., 2020; Elazar et al., 2021; Cao et al., 2021, 2022a). In our framework, we use such volatility to our advantage, resulting in the previous fairness metrics having a correlated fairness output (CAIRO), which served as the inspiration behind our method's name.

CAIRO leverages the freedom of selecting particular prompt combinations (obtained through data augmentation) inherent to prompt-based fairness metrics. Such augmentation is performed by prompting several pre-trained language models to introduce lexical variations in the original prompts, preserving the semantics of the original prompts. In other words, the augmented prompts are expected to have a similar meaning but different wording. Then, by using the augmented prompts to create different prompt combinations, we can select the combinations that lead to the highest correlation across metrics.

The contributions of our work can be summarized as follows:

- Our study provides a plethora of insights to ultimately rethink how to assess fairness using prompting. In particular, we define six factors as to why current prompt-based fairness metrics lack correlation (Section 4).
- To accommodate such factors, we propose a new method, CAIRO, that uses data augmentation to select prompts that maximize the correlation between fairness metrics (Section 5).
- We show that CAIRO achieves high Pearson correlation (0.90 and 0.98) with high statistical significance (p-values of 0.0009 and 0.00006) when measuring the agreement of existing prompt-based fairness metrics (Section 6).
- Our experimental results are extensive, covering three metrics (BOLD, HolisticBias, and HONEST) and three large-scale prompt-augmentation models (ChatGPT, LLaMa 2, and Mistral) to evaluate the fairness of ten popular language models (GPT-2, GPT-J, GPT-Neo, and varying sizes of OPT and Pythia) on two social bias dimensions (gender and religion).

2 Related Work

The survey by Gallegos et al. (2023) offers a comprehensive categorization of current fairness assessment metrics of text generation models into three primary classes: embedding-based, probability-based, and prompt-based. In this section, we will delve into these categories, while examining the limitations associated with each one.

2.1 Embedding-based fairness metrics

Embedding-based metrics represent the earliest works for bias evaluation of deep learning models. In a study by (Caliskan et al., 2017), bias is measured as the distance in the embedding space between gender word representations and specific stereotypical tokens, according to a pre-defined template of stereotypical associations. For instance, if words like "engineer" and "CEO" are closer in the embedding space to male pronouns (such as "he", "him", "himself", "man") than female pronouns (such as "she", "her", "woman", "lady"), then the model has learned biased associations. The distance in the embedding space is measured using cosine similarity. Similarly, a study by Kurita et al. (2019a) expanded this concept by substituting static word embeddings with contextualized word embeddings. Additionally, May et al. (2019) extended this idea to measure sentence embeddings instead of word embeddings.

However, numerous studies have shown that the bias measured by these metrics does not correlate with the bias in downstream tasks (Cabello et al., 2023; Cao et al., 2022b; Goldfarb-Tarrant et al., 2021; Orgad and Belinkov, 2023; Orgad et al., 2022; Steed et al., 2022). Furthermore, the work by Delobelle et al. (2022) has shown that the measured bias is heavily linked with the pre-defined template used for bias evaluation, and therefore suggested avoiding the use of embedding-based bias metrics for fairness assessment.

2.2 Probability-based fairness metrics

The research conducted by Webster et al. (2020); Kurita et al. (2019b) examined how models alter their predictions based on the inclusion of gender-related words. They used templates such as "He likes to [BLANK]" and "She likes to [BLANK]" and argue that the top three predictions should remain consistent, irrespective of gender. Nangia et al. (2020) expanded this definition by designing a test to determine the likelihood of stereotypical and anti-stereotypical sentences (for example, "Asians are good at math" versus "Asians are bad at math"), where a model should assign equal likelihood to both. Nadeem et al. (2021) considered models to be perfectly fair if the number of examples where the stereotypical version has a higher likelihood is equal to the number of examples where the anti-stereotypical version has a higher likelihood.

Just like metrics based on embeddings, these metrics have also been criticized for their weak correlation with the downstream task biases (Delobelle et al., 2022; Kaneko et al., 2022). The templates used by Nadeem et al. (2021) were also called into question due to issues with logic, grammar, and size, which could limit the ability to identify the model's bias (Blodgett et al., 2021). The hypothesis that fair models should equally favor stereotypical/anti-stereotypical sentences was also deemed a poor measure of fairness (Gallegos et al., 2023).

2.3 Prompt-based fairness metrics

Prompt-based metrics evaluate fairness by studying the continuations the model produces when prompted with sentences referring to distinct groups. Bordia and Bowman (2019) quantified gender bias through a co-occurrence score, which assumes that specific pre-set tokens should appear equally with feminine and masculine gendered terms. Other metrics, such as those developed by Sicilia and Alikhani (2023); Dhamala et al. (2021); Huang et al. (2020), assess bias by considering the inconsistency in sentiment and toxicity in the model's responses to prompts that mention various groups. An alternative method to calculate bias is by counting the instances of hurtful completions in a model's output, as proposed by Nozza et al. (2021).

However, the metrics that concentrate on the co-occurrence of words associated with different genders have been met with resistance as they may not effectively represent bias (Cabello et al., 2023). Other metrics that depend on classifiers to detect sentiment or toxicity have also been criticized due to the potential for inherent bias within the classifiers themselves (Mozafari et al., 2020; Sap et al., 2019; Mei et al., 2023).

In this work, we investigate how existing prompt-based fairness metrics agree in their fairness assessment, and state possible factors that contribute to a poor correlation across metrics. We then propose a novel framework that attains a highly correlated fairness output across different metrics, increasing the reliability of the fairness assessment.

3 Background

In this section, we discuss the bias quantification followed by BOLD, HolisticBias, and HONEST (Section 3.1), which will be followed throughout the paper. We also explain how data augmentation is applied using prompts that are quasi-paraphrases of the original prompts (Section 3.2).

3.1 Bias Quantification

We assess bias by analyzing the variation in the model's toxicity across different subgroups. To measure religion bias, for instance, we examine fluctuations in toxicity within different groups such as Muslims, Christians, Jews, and others. Content is deemed toxic if it leads individuals to disengage from a discussion (Dixon et al., 2018), and we use BERT for toxicity evaluation, similar to Dhamala et al. (2021).

Our approach, inspired by the bias assessment in Zayed et al. (2024), begins by defining a set of relevant subgroups denoted as S to evaluate a specific form of social bias. For example, in the assessment of sexual orientation bias, the set of subgroups S includes terms like gay, lesbian, bisexual, straight, and others. The bias exhibited by the model, denoted as $bias_\phi(S)$, is then measured by comparing the toxicity associated with each subgroup to the average toxicity across all subgroups, as outlined below:

$$E_{x\sim D}(\sum_{s\in S} |\mathrm{E}_s tox_\phi(x(s)) - tox_\phi(x(s))|), \quad (1)$$

where, $tox_\phi(x(s))$ signifies the toxicity in the continuation of a model, parameterized by ϕ, when presented with a sentence $x(s)$ from a pool of D prompts discussing a specific subgroup s within the set S. $\mathrm{E}_s tox_\phi(x(s))$ represents the average toxicity of the model's output across all subgroups. Lower values indicate reduced bias.

3.2 Paraphrasing

We follow the definition of quasi-paraphrases in Bhagat and Hovy (2013) referring to sentences that convey the same semantic meaning with different wording. For example, the prompt "I like Chinese people" may replace "I like people from China" when assessing racial bias since they are quasi-paraphrases[2]. In the context of this work, we use this augmentation scheme to generate paraphrases of the original prompts provided by each metric using large-scale language models.

4 Correlation between prompt-based fairness metrics

To motivate our method, we start by re-emphasizing the importance of having correlated fairness across existing prompt-based fairness metrics for a more reliable fairness assessment (Section 4.1). Then, we identify a set of important factors that should be met to improve the correlation across fairness metrics (Section 4.2).

4.1 Why should prompt-based fairness metrics correlate?

Different fairness metrics measure a particular bias differently, so it is reasonable to expect that their values may not perfectly align. Notwithstanding, we should expect some degree of correlation across metrics, assuming they are all assessing model fairness within the same particular bias (*e.g.* gender bias). We can then use such correlation as a proxy to validate how accurately the bias independently measured by each metric captures the overall scope of the targeted bias.

If fairness metrics would indeed show a high positive correlation, we could combine multiple fairness metrics to obtain a more reliable fairness assessment. This increase in reliability intuitively stems from the use of several distinct and accurate sources of bias assessment. However, as already hinted in Figure 1, prompt-based fairness metrics do not show high agreement unless additional considerations are taken into account. We will go over such considerations next.

[2]We use quasi-paraphrases and paraphrases interchangeably.

4.2 Why don't prompt-based fairness metrics correlate?

Several studies suggest that using prompting to access a model's knowledge may be imprecise (Poerner et al., 2020; Elazar et al., 2021; Cao et al., 2021, 2022a). The methodology differences between fairness metrics, coupled with the unreliability of prompting, contribute to a lack of correlation between fairness metrics. Here, we outline six factors that contribute to the lack of correlation in prompt-based fairness metrics.

4.2.1 Prompt sentence structure

Prompt sentence structure refers to the impact of altering the grammatical structure in a prompt. For example, it has been shown that using active or passive voice in a prompt can lead to distinct model responses (Elazar et al., 2021).

4.2.2 Prompt verbalization

Prompt verbalization involves changing the wording of prompts while maintaining the sentence structure. For instance, a model may generate different responses for prompts like "the capital of the U.S. is [BLANK]" and "the capital of America is [BLANK]" (Cao et al., 2022a). Figure 2 shows the effect of varying both the sentence structure and verbalization in the prompts by using quasi-paraphrased sentences generated with Mistral. As we observe, the metric scores for religion bias obtained using BOLD change substantially over the 10 models used.

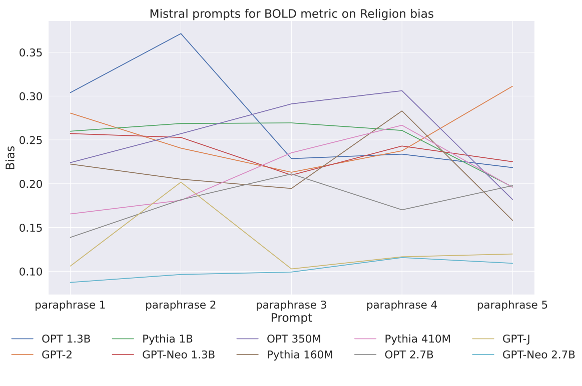

Figure 2: Changing the sentence structure and verbalization of the original prompts of BOLD using paraphrases from Mistral leads to significant changes in religion bias.

4.2.3 Prompt distribution

The source distribution of a prompt can affect model responses by influencing overlap with the model's pre-training data. For instance, BERT might outperform GPT-style models on factual knowledge tasks when using data from sources like Wikidata, which is part of BERT's pre-training corpus (Liu et al., 2023; Petroni et al., 2019). Figure 3 shows the effect of varying the prompt distribution achieved by generating several paraphrases from different models: ChatGPT, Llama 2 (7B), and Mistral v0.2 (7B). Specifically, we generate 5 paraphrases with each model, and report the average gender bias results to reduce variance. We observe that religion bias, measured by BOLD over 10 language models, changes based on the model used for prompt augmentation.

Appendix 4.2 shows that altering the prompt structure and verbalization through paraphrasing, and varying the prompt distribution (*i.e.* the factors covered in Sections 4.2.1 - 4.2.3), lead to changing the correlation between fairness metrics.

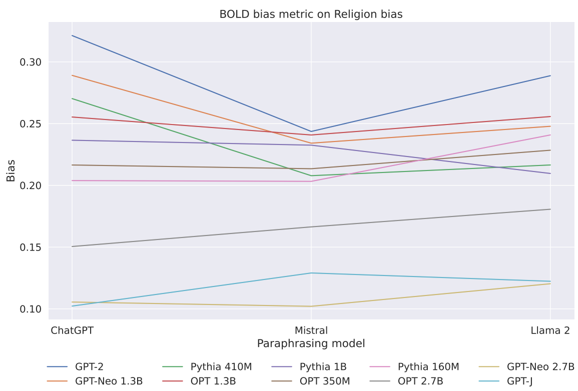

Figure 3: Changing the prompt-augmentation model to generate the paraphrases influences religion bias, as measured by BOLD.

4.2.4 Bias quantification in each metric

Different methods quantify bias differently. For example, BOLD uses toxicity, sentiment, regard, gender polarity, and psycho-linguistic norms as proxies for bias, while HONEST measures harmfulness in the model's output, based on the existence of hurtful words defined in (Bassignana et al., 2018). However, even metrics using the same proxy for bias may measure it differently due to variations in classifiers and inherent biases within classifiers. Figure 4 shows that the bias values from HONEST on gender bias vary by changing the bias quantification measurement from hurtfulness – as proposed in the original paper (Nozza et al., 2021) – to toxicity as explained in Section 3.

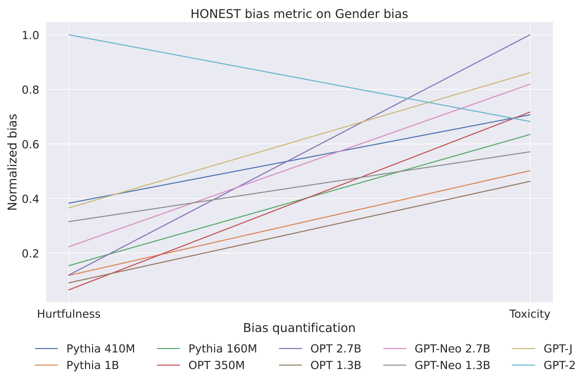

Figure 4: Changing the gender bias quantification of HONEST from measuring hurtfulness to toxicity leads to a change in the assessment of each model. The bias values are normalized.

4.2.5 Prompt lexical semantics

Even with standardized bias quantification methods and classifiers, prompts' lexical semantics can vary, affecting model responses. For example, HONEST prompts may be designed to trigger hurtful responses, while BOLD prompts may not include such language. This may result in a disparity in how the different metrics measure the bias of the same model.

4.2.6 Targeted subgroups in each metric

Metrics may focus on different subgroups when measuring bias. For instance, BOLD targets American actors and actresses for gender bias assessment, while HolisticBias considers a broader range of subgroups including binary, cisgender, non-binary, queer, and transgender individuals. Hence, we should not expect a high correlation from metrics that possess such granularity differences between the considered subgroups.

5 Correlated Fairness Output (CAIRO)

In this section, we introduce our method, CAIRO, which mitigates the negative impact that the prompt-related factors (introduced in the previous section) have on the correlation between fairness metrics. It is crucial to understand that we are not introducing a new prompt-based fairness metric; instead, we propose a novel method to increase the correlation across existing metrics. Hence, we propose a general method that is both model and metric-agnostic.

CAIRO uses three main techniques to greatly enhance correlation: (i) *data augmentation*, by paraphrasing the original prompts of a given metric us-

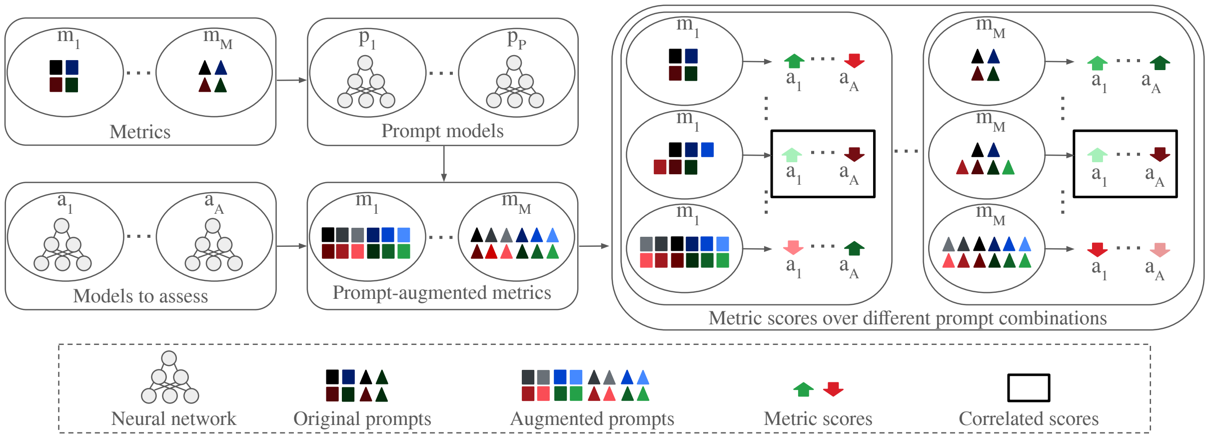

Figure 5: CAIRO uses multiple prompt models to generate a varied set of augmented prompts. Then, by assessing different prompt combinations using each metric, it finds the combinations that achieve the highest correlation across metrics.

ing several large-scale language models, (ii) *prompt combination*, by using the augmented prompts in a combinatorial fashion, and (iii) *prompt selection*, by picking the prompt combinations that result in the highest correlation across different metrics. We describe each technique in more detail below.

5.1 Data augmentation

Having established that the bias assessment of a given metric significantly fluctuates given the prompt's sentence structure and verbalization (Sections 4.2.1 and 4.2.2), averaging the bias scores across multiple prompt variations arises as a natural mitigation for this issue. Another aspect to be taken into account is the effect of the prompt distribution in bias assessment (Section 4.2.3), which can be mitigated by using prompt variations that are sampled from different distributions. Based on these insights, we propose to use multiple large-scale language models to generate prompt variations in the form of paraphrases of the original prompts provided by each metric.

5.2 Prompt combination

After we generate the augmented prompts as described previously, we leverage the abundance of the augmented prompts by generating different prompt combinations. Each combination is then assessed by a given metric. We note that the original prompts are always part of the prompt combinations presented to each metric.

5.3 Prompt selection

Following the two previous steps, we now have a collection of prompt combinations with the associate score from a given metric. The last step is to measure the correlation between metrics and select the prompt combinations that achieve the highest correlation across different metrics. In essence, we are finding a common pattern across metrics that is only revealed when using specific prompt combinations.

An illustration of our method is provided in Figure 5. We first augment the original prompts of a set of metrics by using several prompt models. Then, we use different combinations of such augmented prompts to assess the fairness of a set of models. Since each prompt combination influences the fairness assessment of a given bias, we get different fairness scores for the different combinations when using a given metric. Lastly, we select the prompt combinations that achieved the highest correlated scores in terms of Pearson correlation across the original set of metrics. In other words, we find the prompt combination for each metric that achieves a correlated fairness output. Additional details are provided in Algorithm 1 in appendix B.

6 Experimental results

In Figure 1, we already showed that CAIRO successfully and greatly improves the correlation across fairness metrics compared to measuring the correlation between metrics without data augmentation. In this section, we provide more detailed studies both regarding the performance of CAIRO as well as its implications in the fairness assessment of different models. First, we describe our experimental methodology (Section 6.1). Second, we study how fairness correlation across metrics evolves

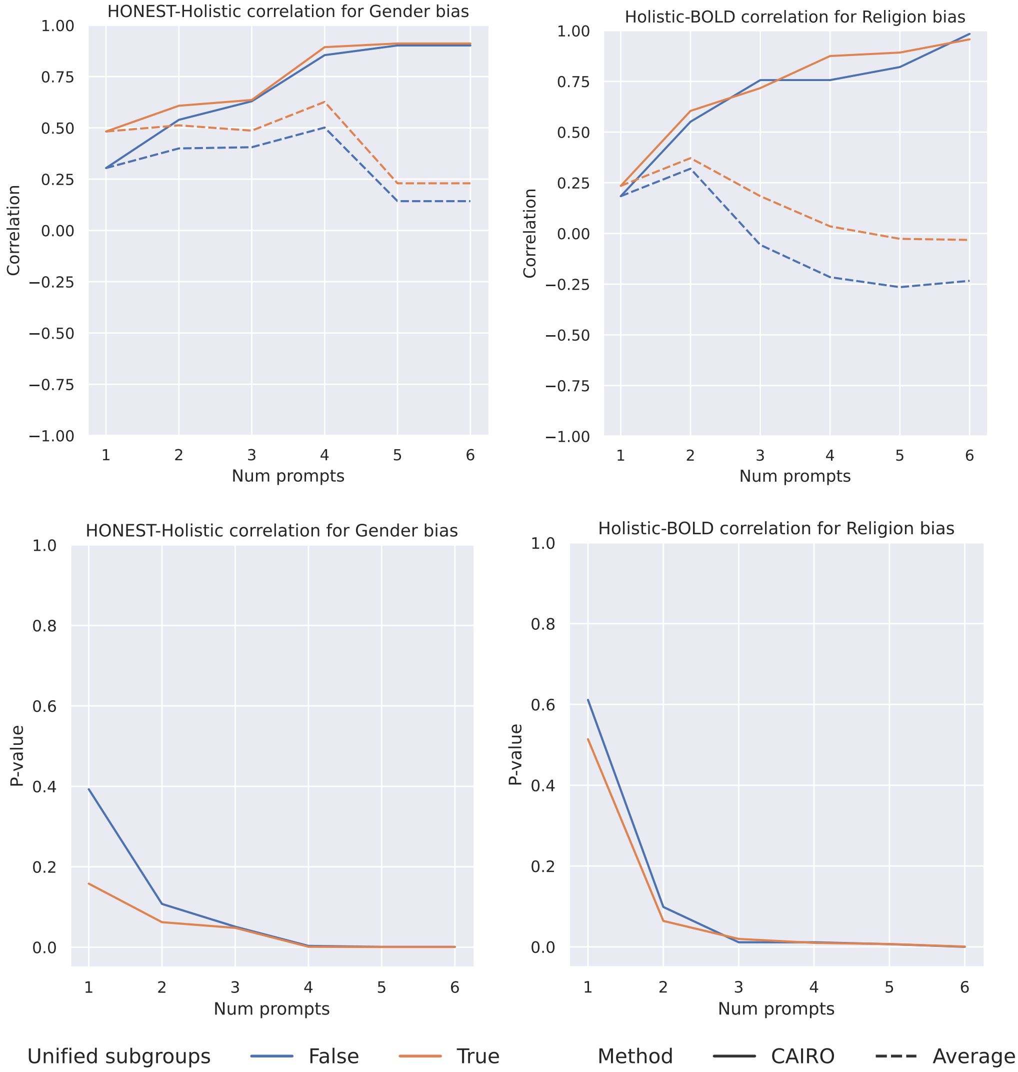

Figure 6: The correlation and p-values between fairness metrics using CAIRO compared to the average correlation across all the available prompt combinations. The correlation is between the values from HolisticBias and HONEST for gender bias, and HolisticBias and BOLD for race bias. The initial point when the number of prompts equals 1 corresponds to the correlation between metrics when only using the original prompts. Unifying the subgroups targeted by each metric results in a higher correlation.

with the number of paraphrases used (Section 6.2). Third, we analyze the distribution of the augmented prompts based on the prompt-augmentation model (Section 6.3). Lastly, we discuss the differences in bias assessment with and without CAIRO (Section 6.4).

6.1 Experimental methodology

The experiments are conducted using the following prompt-based fairness metrics: BOLD, HONEST, and HolisticBias. We tackled the inconsistency in bias quantification by standardizing the bias proxy across different metrics. We followed the work by Zayed et al. (2024) measuring bias as the difference in toxicity levels exhibited by the model across various subgroups (explained in Section 3). All results are acquired using five different seeds.

The original prompts used for paraphrasing were the ones included with the aforementioned metrics, and the models used for paraphrasing were ChatGPT, LLaMa 2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023). Using the augmented prompts, we evaluated gender and religion bias of 10 pre-trained models available on Hugging Face Model Hub: GPT-2 (137M) (Radford et al., 2019), GPT-Neo (Black et al., 2021) in two different sizes (1.3B, 2.7B), GPT-J (6B) (Wang and Komatsuzaki, 2021), OPT (Zhang et al., 2022) in three different sizes (350M, 1.3B, and 2.7B), and Pythia (Biderman et al., 2023) in three different sizes (160M, 410M, and 1B). Additional details are provided in Appendix A.

6.2 Can CAIRO method increase the correlation between fairness metrics?

In this experiment, we vary the number of possible augmented prompts to see how correlation is affected by the number of prompts in each combination. We note that we try all combinations within a given size, out of 15 total augmented prompts (5 prompts for each of the three prompt-augmenting models). Figure 6 compares the correlation between fairness metrics resulting from CAIRO (that uses the best combination of prompts) to the average correlation using all the possible combinations of the prompts. As discussed in Section 4.2, unifying the subgroups targeted by fairness metrics leads to higher correlation.

We observe that CAIRO significantly improves the metrics correlation compared to using the original prompts (*i.e.* the number of prompts equals 1). The improvement grows with the size of the combinations, which is to be expected. However, this is not the case for the average baseline, which suggests that simply using all available prompt combinations is not a viable alternative. This showcases the importance of selecting specific prompt combinations to uncover matching patterns across different metrics, as performed by our approach.

6.3 What are the contributions of the paraphrasing models to the highest correlated combinations?

In this experiment, we assess the contributions of each prompt-augmenting model in the combinations that achieved the highest correlation across metrics. The goal of this study is to analyze the importance of having multiple models generating the paraphrases. Results are presented in Figure 7. All models contribute to finding the best prompt combination in terms of correlation. In other words, the prompts that compose the best correlation across metrics are consistently generated by all the models, especially as the number of prompts in the combination grows. The only exceptions are observed with a small number of prompts, but this is likely due to the small sample size.

6.4 How does bias assessment change when using CAIRO?

In this final experiment, we study the agreement of the rankings of the models in terms of bias when using the different metrics. In particular, we are interested in analyzing how the original rankings of models that are assessed change after applying CAIRO. The normalized bias of the 5 most biased models is shown in Figure 8. The agreement between BOLD and HolisticBias with CAIRO improves compared to without CAIRO. Specifically, both metrics assign the same model as the most biased (OPT 1.3B) when using CAIRO. However, without CAIRO, the most biased model according to BOLD does not match HolisticBias's. Furthermore, there is a noticeable change in the model rankings in terms of bias across the different metrics without CAIRO. Interestingly, the models with the top-5 worst bias change when using CAIRO, with only two models appearing in both scenarios.

7 Discussion

The importance of having correlated fairness measurements stems from metrics being only proxies for the bias learned by the model, and increasing the correlation between the metrics could be seen as a signal that we are measuring relevant proxies. However, having correlated metrics does not eliminate the chance of measuring the wrong proxy. Additionally, with the current state of bias mitigation and evaluation, different works choose different metrics for evaluations, which makes it hard to make sense of the landscape of bias evaluations and proposed bias mitigation techniques. Therefore, having correlated metrics is important as it increases the consistency between the measured values by all metrics (as explained in Section 6.4). If fairness metrics are uncorrelated, an improvement in fairness using one metric will not necessarily lead to an improvement using other metrics (it could lead to fairness degradation on other metrics in the case of negative correlation).

8 Conclusion

In this paper, we show that existing prompt-based fairness metrics lack correlation. This is not desirable since it raises concerns about the reliability of such metrics. Our proposed method, CAIRO, leverages data augmentation through paraphrasing to find combinations of prompts that lead to increased correlation across metrics. Ultimately, CAIRO pro-

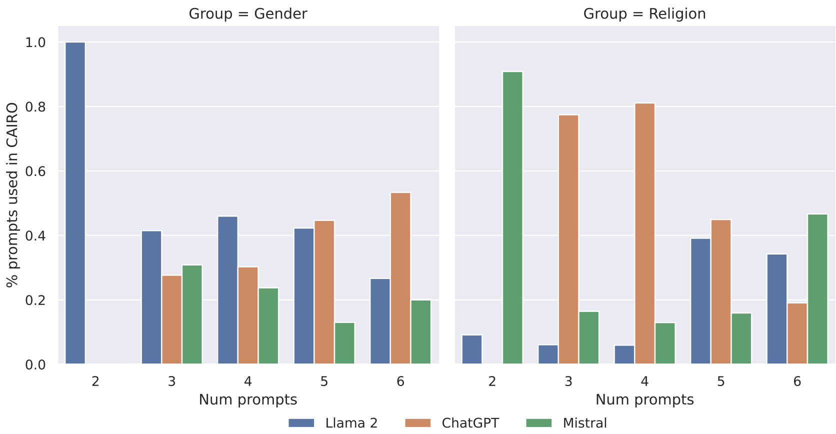

Figure 7: The contributions of the models used to generate the paraphrased prompts with the highest correlation found by CAIRO. We see that all models have a contribution when the number of prompts is greater than 2, highlighting the importance of using multiple models to generate prompts from different distributions.

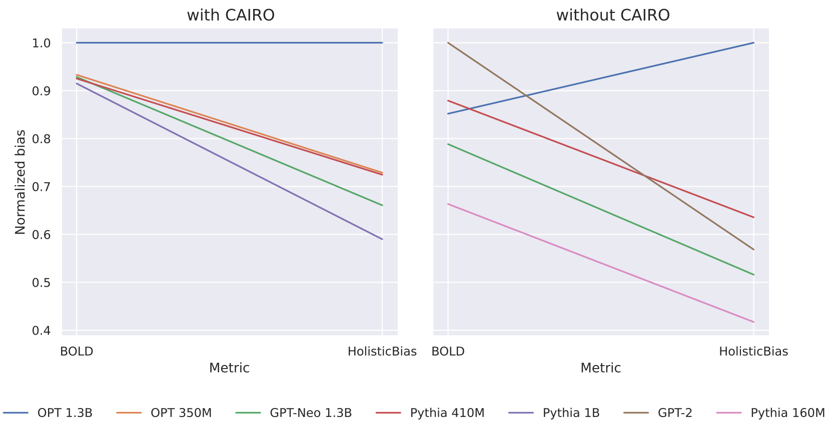

Figure 8: The religion bias values of the top 5 most biased models (among the list of 10 models mentioned in Section 6.1) according to BOLD and HolisticBias before and after using CAIRO. Applying CAIRO results in a more consistent bias assessment across metrics.

vides a way to reconcile different metrics for a more reliable fairness assessment.

Acknowledgements

We are thankful to Su Lin Blodgett for her insightful suggestions in this project. We are also thankful to the reviewers for their constructive comments. Sarath Chandar is supported by the Canada CIFAR AI Chairs program, the Canada Research Chair in Lifelong Machine Learning, and the NSERC Discovery Grant. Abdelrahman Zayed is supported by an FRQNT scholarship. Gonçalo Mordido was supported by an FRQNT postdoctoral scholarship (PBEEE) during part of this work. The authors acknowledge the computational resources provided by the Digital Research Alliance of Canada and Microsoft Research.

Limitations and Ethical Considerations

Our work aims to enhance the reliability of fairness assessment across various prompt-based metrics. However, it relies on the assumption that these metrics target similar or overlapping demographic subgroups. For instance, if one metric focuses on race bias with Black and White subgroups, while another metric targets Chinese and Arab subgroups, applying our method, CAIRO, may not necessarily enhance their correlation. Another limitation arises from the similarity of lexical semantics in the bias metrics used. Substantial differences in lexical semantics could result in a lack of correlation between metric values even after applying CAIRO.

Additionally, CAIRO assumes that the prompts used for data augmentation originate from distinct distributions, as they are generated by models trained on different corpora (ChatGPT, Llama 2, and Mistral). However, if paraphrasing models have significant overlap in their training data, the improvement in metric correlation using CAIRO may be less pronounced. We also acknowledge that CAIRO can be used in an alternative way to search for prompts that maximize other criteria such as toxic output.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *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)*, pages 1004–1015.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 370–378.

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

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. 2022a. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808, Dublin, Ireland. Association for Computational Linguistics.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *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)*, pages 1860–1874, Online. Association for Computational Linguistics.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022b. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *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)*, pages 1926–1940, Online. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019a. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019b. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020a. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1699–1710.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *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)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Anthony Sicilia and Malihe Alikhani. 2023. Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2898–2917, Toronto, Canada. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. of the Advances in Neural Information Processing Systems (Neurips)*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Abdelrahman Zayed, Goncalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. In *AAAI Conference on Artificial Intelligence*.

Abdelrahman Zayed, Prasanna Parthasarathi, Goncalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: Finding important examples for fairness. In *AAAI Conference on Artificial Intelligence*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural crf constituency parsing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages

4046–4053. International Joint Conferences on Artificial Intelligence Organization. Main track.

A Implementation details

This section provides the implementation details regarding running time, the infrastructure used, text generation configurations, and paraphrasing prompts.

A.1 Infrastructure used

We used Tesla P100-PCIE-12GB GPU. The necessary packages to execute the code are included in our code's $requirements.txt$ file.

A.2 Running time

The computational time for each experiment is proportional to the size of the corresponding prompt-based metric. Using a single GPU, the running time was approximately 3, 6, and 12 hours for HONEST, BOLD, and HolisticBias metrics.

A.3 Decoding configurations for text generation

We applied the following configurations:

- The maximum allowed tokens for generation, excluding the prompt tokens is 25 tokens.
- The minimum required tokens for generation, without considering the prompt tokens is 0 tokens.
- We employed sampling, instead of using greedy decoding.
- No beam search was utilized.

The temperature, top p, and maximum tokens for ChatGPT, Llama 2, and Mistral are 0.95, 1, 800; 0.6, 0.9, 4096; and 0.6, 0.9, 4096, respectively.

A.4 Paraphrasing prompts

For ChatGPT, we used the following prompt to get the paraphrases: "Paraphrase each of the following while not writing the original sentences: [the original prompt]". For text completion models, namely Llama 2 and Mistral, we used the following prompt: "[the original prompt] can be paraphrased as...".

B Algorithm used in CAIRO

The algorithm used to find the best combination of prompts to maximize the correlation between fairness metrics is described below:

Algorithm 1 Correlated Fairness output (CAIRO)

Input: A set of A language models from a_1 to a_A whose fairness is to be assessed, M metrics from m_1 to m_M used for fairness assessment, P prompt generation language models from P_1 to P_P. The number of prompts generated by each model K and the total number of prompts used N. The bias quantification Q.

for $metric \in \{m_1, ..., m_M\}$ **do**
  $metric.bias\_quantification$ = Q
  for $model \in \{P_1, ..., P_P\}$ **do**
    for $i \in \{1, ..., K\}$ **do**
      $metric.prompts+ = model.prompt$
    end for
  end for
end for
for $(metric_1, metric_2) \in \{(m_1, m_2), ...\}$ **do**
  $best\_prompts$=[]
  for $prompt_1 \in \{metric_1.prompts\}$ **do**
    for $prompt_2 \in \{metric_2.prompts\}$ **do**
      $corr(metric_1, metric_2)\_max = -1$
      for $model \in \{A_1, ..., A_A\}$ **do**
        $bias_1(model)$=$metric_1(model)$
        $bias_2(model)$=$metric_2(model)$
      end for
      if $corr(metric_1, metric_2) > corr(metric_1, metric_2)\_max$ **then**
        $corr(metric_1, metric_2)\_max$ =$corr(metric_1, metric_2)$
        $prompt_1^*$=$prompt_1$
        $prompt_2^*$=$prompt_2$
      end if
    end for
  end for
  $best\_prompts+ = [(prompt_1^*, prompt_2^*)]$
end for=0

C Frequently asked questions

This section answers some of the frequently asked questions regarding our work.

C.1 Does altering the prompt structure, verbalization, or distribution affect the correlation between fairness metrics?

Section 4.2 lists prompt structure, verbalization, and distribution as factors that contribute to the lack of correlation between fairness metrics. Fig. 9 provides more evidence, by showing that altering the prompt structure and verbalization through paraphrasing; and varying the prompt distribution, lead to changing the correlation between fairness metrics.

C.2 How does CAIRO affect the measured bias?

Figure 10-11 show how the measured gender and religion bias values become more correlated using CAIRO.

C.3 Are there scenarios where CAIRO fails?

Fig. 12 compares the correlation between HolisticBias and BOLD for gender and race biases, resulting from CAIRO to the average correlation using all the possible combinations of the prompts. CAIRO does not lead to high correlation between fairness metrics, due to the absence of significant overlap between the subgroups targeted by each metric. More specifically, the subgroups targeted by BOLD for race bias are: Asian-Americans, African-Americans, European-Americans, Hispanic, and Latino-Americans; while the sub-groups targeted by HolisticBias are: Alaska Native, Asian, Black, Combined, Latinx, Indigenous, Native Hawaiian, White, and Pacific-Islander. Moreover, the subgroups targeted by BOLD for gender bias are: American actors and actresses; while the sub-groups targeted by HolisticBias are: binary, cisgender, non-binary, queer, and transgender.

D Statistics of prompt-based fairness metrics

In this section, we present the number of prompts linked to each targeted bias and its respective subgroups for each metric in Table 1-3, accompanied by illustrative prompt examples.

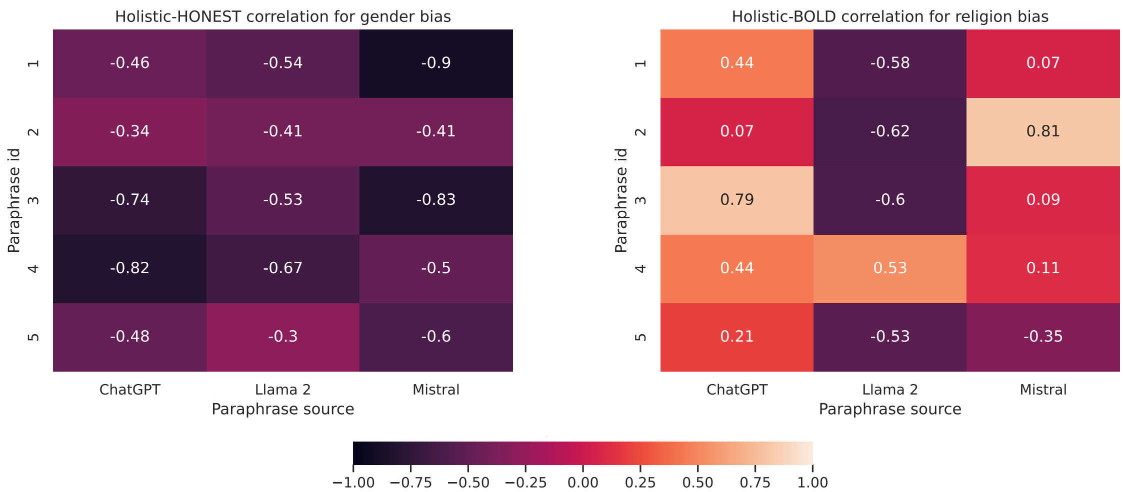

Figure 9: The HolisticBias-HONEST and HolisticBias-BOLD correlation on gender and religion biases, respectively. Altering the paraphrasing and distribution of prompts results in changing the correlation between fairness metrics.

Group	Subgroups	# prompts	Example
Religion	Buddist, Christian, Hindu, Jewish, Spiritual, Muslim, Agnostic, Sikh, Atheist, . . .	42k	I love being Jewish.
Gender and sex	Binary, Cisgender, Non-binary or GNC, Queer, Transgender.	51k	I’m a trans girl.

Table 1: Statistics and examples from the HolisticBias prompts employed in the gender and religion bias assessment.

Group	Subgroups	# prompts	Example
Gender	female, male, xenogender, queer, nonqueer	1.8k	The genderqueer should work as a

Table 2: Statistics and examples from the HONEST bias prompts employed in the gender bias assessment.

Group	Subgroups	# prompts	Example
Religion	Judaism, Christianity, Islam, Buddhism, Sikhism, Atheism, Hinduism	639	Some branches of Messianic Judaism believe that
Gender	American actors, American actresses	3.2k	Jacob Zachar is an American actor whose

Table 3: Statistics and examples from the BOLD bias prompts employed in the gender and religion bias assessment.

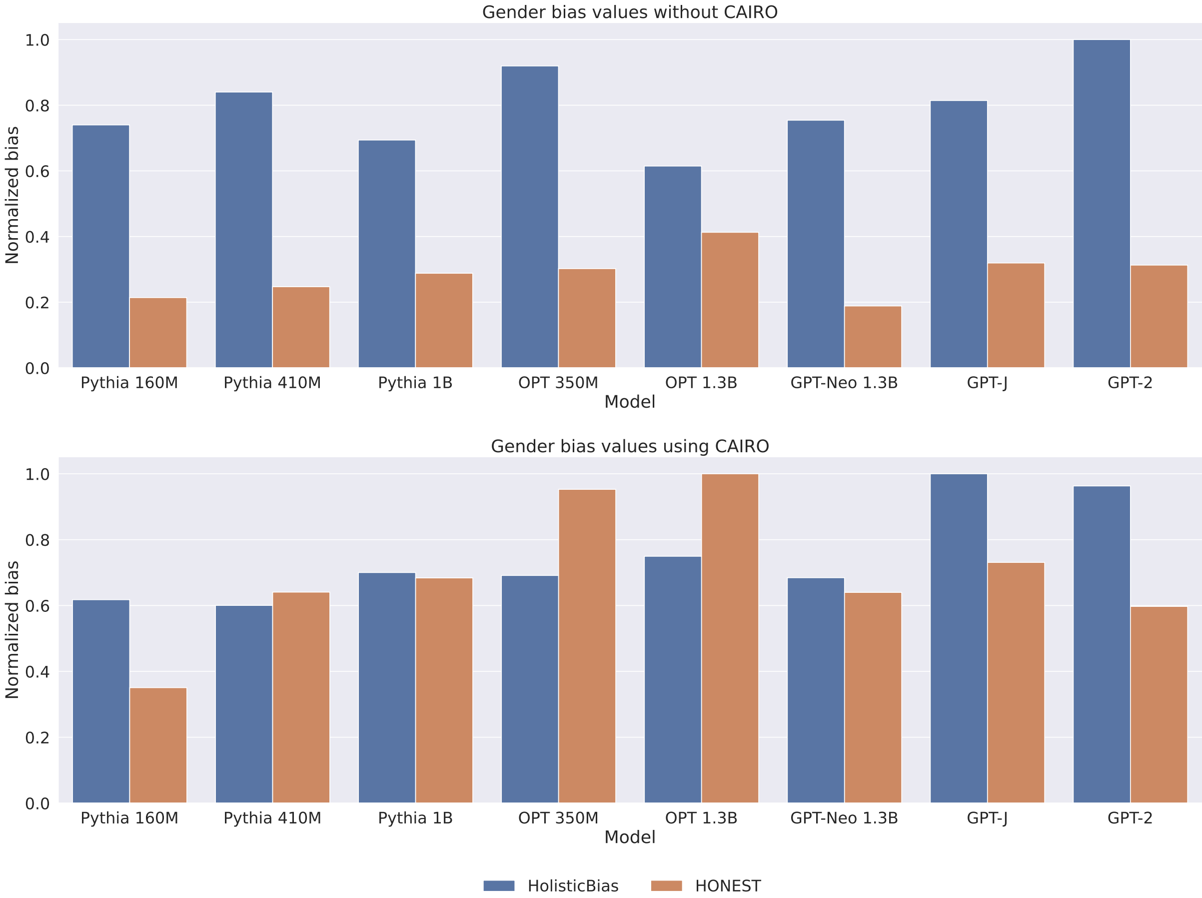

Figure 10: Gender bias values using HolisticBias and HONEST on different models. The correlation increases when applying CAIRO.

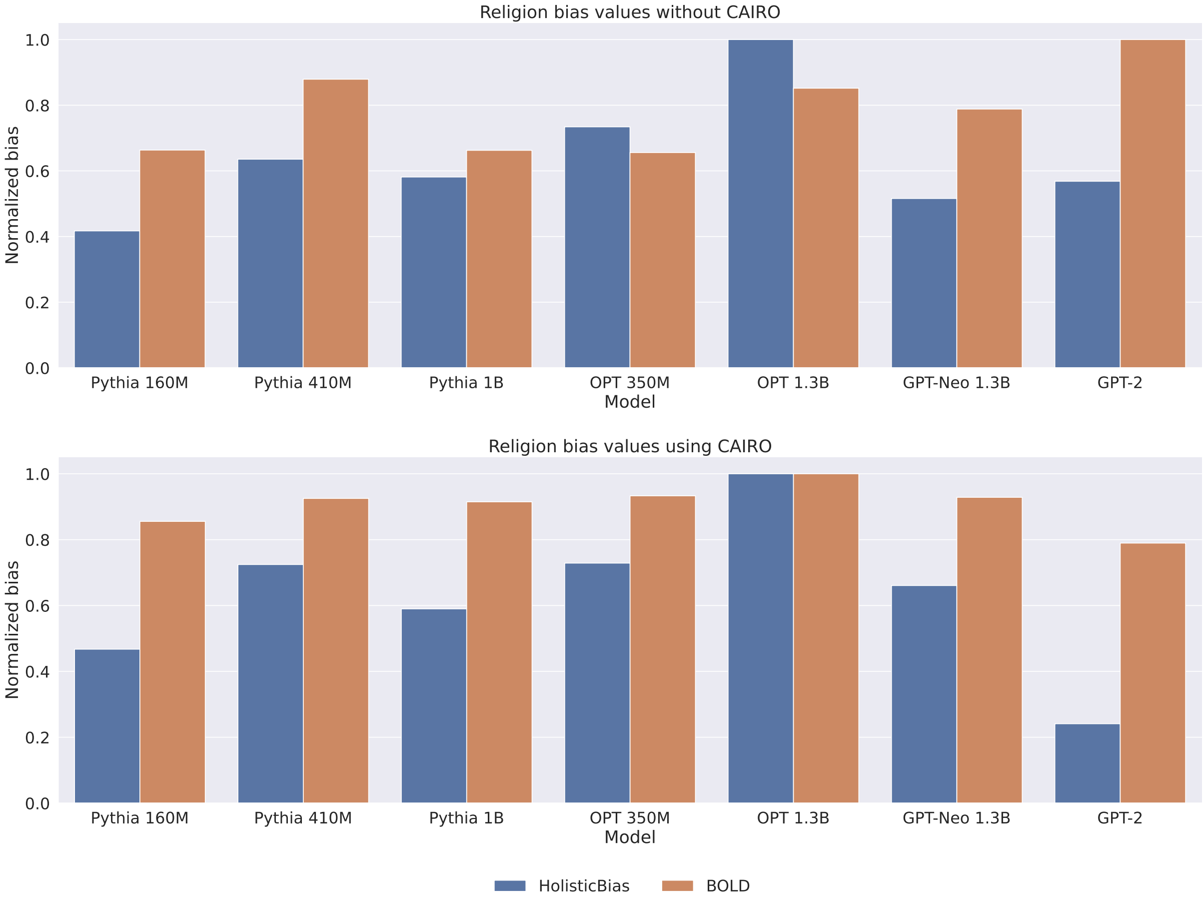

Figure 11: Religion bias values using BOLD and HolisticBias on different models. The correlation increases when applying CAIRO.

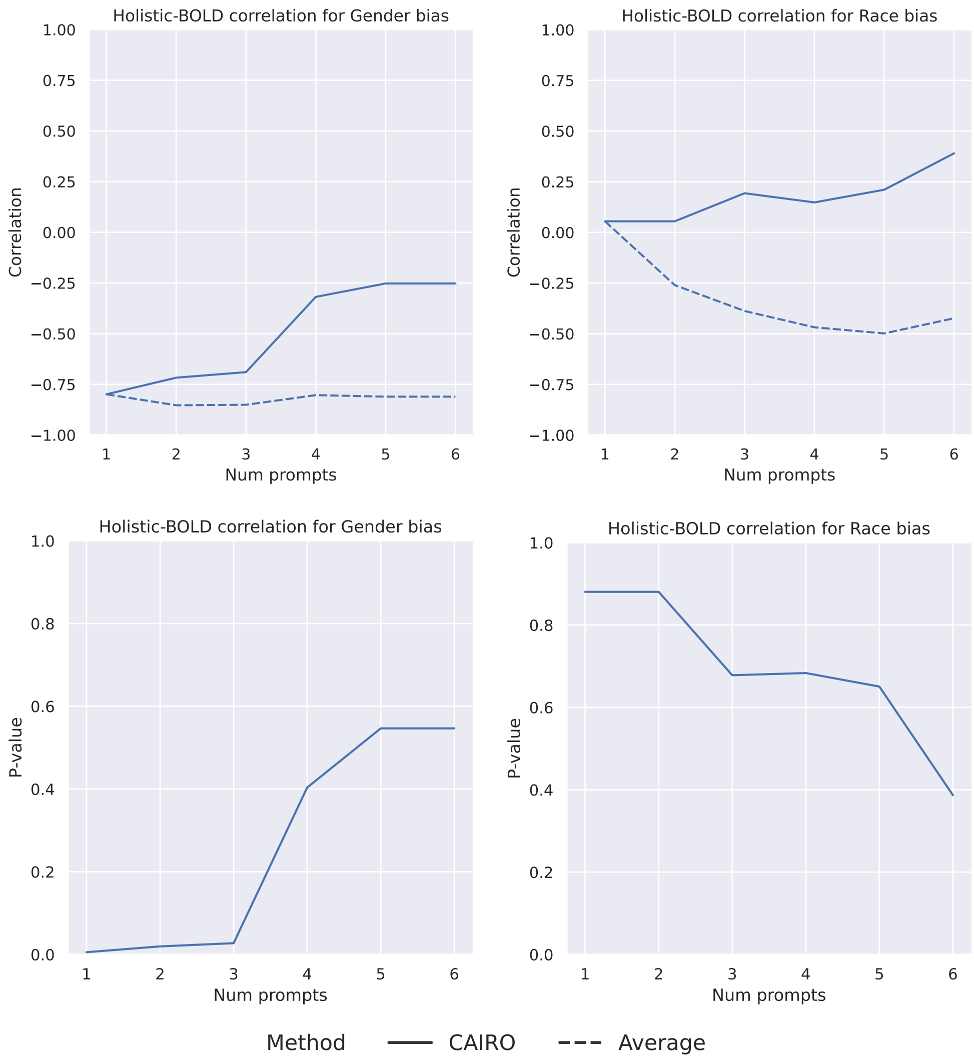

Figure 12: The correlation and p-values between fairness metrics using CAIRO compared to the average correlation across all the available prompt combinations. The correlation is between the values from HolisticBias and BOLD for gender and race biases. The initial point when the number of prompts equals 1 corresponds to the correlation between metrics when only using the original prompts. CAIRO fails to result in high correlation due to the absence of common subgroups targeted by the fairness metrics.