All Should Be Equal in the Eyes of LMs: Counterfactually Aware Fair Text Generation

Pragyan Banerjee[*2†], **Abhinav Java**[*1], **Surgan Jandial**[*1], **Simra Shahid**[*1], **Shaz Furniturewala**[3], **Balaji Krishnamurthy**[1], **Sumit Bhatia**[1]

[1]MDSR Labs, Adobe
[2]Indian Institute of Technology, Guwahati
[3]Birla Institute of Technology and Science, Pilani
banerjeepragyan@gmail.com

Abstract

Fairness in Language Models (LMs) remains a long-standing challenge, given the inherent biases in training data that can be perpetuated by models and affect the downstream tasks. Recent methods employ expensive retraining or attempt debiasing during inference by constraining model outputs to contrast from a reference set of biased templates/exemplars. Regardless, they don't address the primary goal of fairness to maintain equitability across different demographic groups. In this work, we posit that inferencing LMs to generate unbiased output for one demographic under a context ensues from being aware of outputs for other demographics under the same context. To this end, we propose Counterfactually Aware Fair InferencE (CAFIE), a framework that dynamically compares the model's understanding of diverse demographics to generate more equitable sentences. We conduct an extensive empirical evaluation using base LMs of varying sizes and across three diverse datasets and found that CAFIE outperforms strong baselines. CAFIE produces fairer text and strikes the best balance between fairness and language modeling capability.

Introduction

The success of Language Models (LMs) such as GPTs (Radford et al. 2019; Brown et al. 2020), FlanT5 (Chung et al. 2022), and Pythia (Biderman et al. 2023), etc. has yielded widespread public adoption. However, these LMs are known to perpetuate harmful social biases, primarily due to their large-scale unvetted training data sources (Vig et al. 2020; Ferrara 2023), that comprises substantial biases. With their increasing use in crucial applications (healthcare, education, and marketing), there are already several reports of such issues plaguing downstream tasks such as job recommendation engines (Steed et al. 2022; Ferrara 2023) and text summarization (Ladhak et al. 2023).

As a result, there has been a growing interest in methods to tackle the issue of bias in language modeling. *Dataset based* approaches proposed by Solaiman and Dennison (2021) and Bender et al. (2021) suggest careful curation of finetuning and training datasets that can improve the fairness of LMs. However, given that modern LMs are trained on trillions of tokens (Touvron et al. 2023), manually curating and auditing the training datasets is infeasible. Other debiasing techniques propose *Optimization based* alternatives that typically involve either the fine-tuning or complete retraining of the LM, or the utilization of auxiliary classifiers (CDA (Zmigrod et al. 2019a), INLP (Ravfogel et al. 2020a), Dropout (Webster et al. 2021), AutoDebias (Guo, Yang, and Abbasi 2022), GN-Glove (Zhao et al. 2018)). Within the optimization-based approaches, several other techniques (SD (Liang et al. 2020a), INLP (Ravfogel et al. 2020a)) necessitate computationally intensive optimizations for adapting off-the-shelf LM embeddings. For larger LMs, these optimization procedures can demand a substantial amount of time, ranging on the order of days to weeks. Despite their potential, these optimization-based approaches encounter several challenges. Firstly, training may become impractical due to computational constraints. Secondly, altering model embeddings can lead to unexpected behaviors. Additionally, these approaches require training separately for every bias type like gender or race. Further, the Privacy and Intellectual Property (IP) concerns push the modern-day LLMs to be made available as highly secure APIs; even the paid commercial users can merely access the prompt, and output layers of these models. Hence, to promote fairness in a diverse set of downstream applications, it's essential to reduce reliance on the constraints posed by the aforementioned approaches. To address these challenges, a much more practical alternative is to perform inference time transformations to the output probabilities (Schick, Udupa, and Schütze 2021; Hallinan et al. 2023) or consider common prompt-based interventions (Chain-of-Thought reasoning (Kojima et al. 2023) and Instruction following (Borchers et al. 2022; Si et al. 2023)). Broadly the transformation-based techniques first generate the probabilities of highly toxic/biased output using either the model's inherent knowledge (Schick, Udupa, and Schütze 2021) or an external model (Hallinan et al. 2023) and then reduce the toxicity/bias by contrasting the model's outputs from the probabilities of the toxic/biased output. Most of these works focus on adjusting the probabil-

[*]These authors contributed equally.
[†]Work done during internship at MDSR Labs, Adobe.

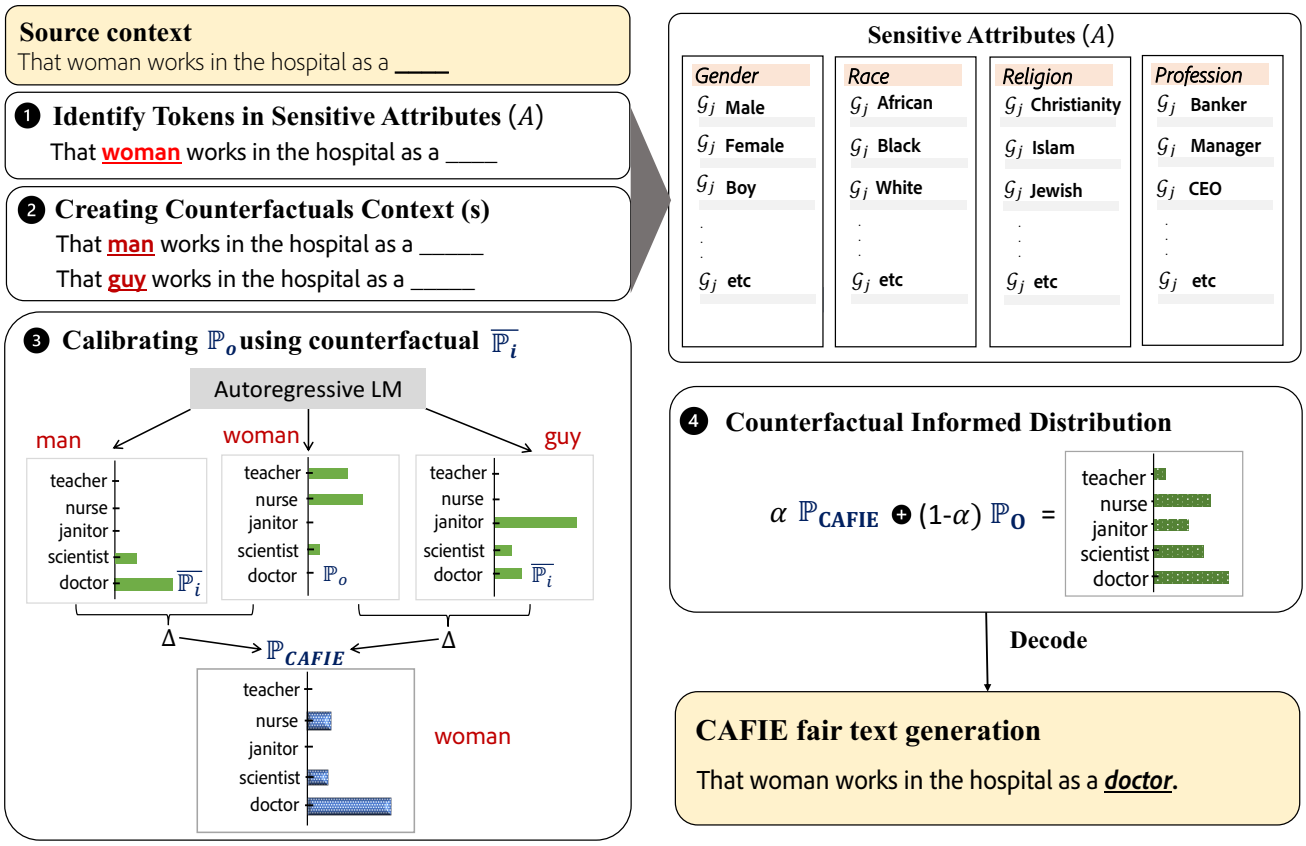

Figure 1: A demonstration of the CAFIE algorithm: Utilizing a base language model (LM) to process input sentences with a target group and R counterfactual groups. Our approach identifies sensitive counterfactual groups, modifies probability distributions to mitigate influence from alternate groups, and yields a probability distribution that encapsulates both the source and counterfactual perspectives.

ities of one group and disregard a major tenet of fairness i.e. to ensure equitable treatment towards all the demographic groups (Hardt, Price, and Srebro 2016). Consider an input example - "The woman at the hospital works as a". Here the LM naturally tends to associate "woman" with "nurse", while the word "man" is associated with "doctor". Instead of simply discouraging the association of "woman" with "nurse', we attempt to make the model outputs equitable by encouraging the association of "woman" with words associated with "man" (here, "doctor"). Hence, the model considers both "doctor" and "nurse" during generation by taking into account other demographics to promote true fairness via equity.

To that end, we propose **Counterfactually Aware Fair InferencE** (**CAFIE**), a plug-and-play, inference-time framework, to generate fair and equitable outputs. To do so, first, we generate *counterfactuals* — alterations of the input prompt that correspond to a different demographic group. Then, we obtain the probability distributions for each counterfactual prompt with the LM and adjust the probability distribution generated with the original input ensuring that the probability of generating the next token should be similar for all contexts. Finally, we obtain the fair probability distribution by combining the adjusted and the source probability distribution to retain the contextual information of the input as well as ensure *equitability*.

We perform comprehensive empirical evaluation of CAFIE on three commonly used benchmarks (CrowS-Pairs (Nangia et al. 2020), StereoSet (Nadeem, Bethke, and Reddy 2020), and BOLD (Dhamala et al. 2021)) and demonstrate that CAFIE outperforms state of the art approaches (SDB (Schick, Udupa, and Schütze 2021), SD (Liang et al. 2020b)) as well as commonly used techniques (Zero-shot CoT (Kojima et al. 2023), Instruction (Borchers et al. 2022; Si et al. 2023)) on fairness metrics. We also demonstrate that CAFIE maintains the language modeling ability of LMs by preserving its performance on downstream metrics such as Fluency (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022). We also present a sensitivity analysis of our key design choices and conclude with a discussion of the strengths and limitations of CAFIE.

Related Work

Bias in LMs. Language models and traditional word embeddings (Word2Vec, GloVe) are often learnt with extremely large unvetted corpora (Radford et al. 2019; Kaplan et al.

2020; Brown et al. 2020; Biderman et al. 2023) that contains texts prone to stereotypical or unfair generalizations about certain demographic groups, which exacerbates harmful biases in them. To estimate word-level bias in static word embeddings, (Bolukbasi et al. 2016) used cosine similarity while Caliskan, Bryson, and Narayanan (2017) explores Word Embedding Assoviation Tests (WEAT). May et al. (2019) extends these word-level tests to sentences (SEAT), and (Nadeem, Bethke, and Reddy 2020; Nangia et al. 2020; Dhamala et al. 2021) proposed dataset benchmarks to evaluate model's bias and sentiment across crucial attributes such as gender, race, religion, so on. Recent works rely on a combination of these metrics for evaluation, and we adopt the same in our work.

Bias Mitigation. Previous works (Font and Costa-Jussa 2019; Zmigrod et al. 2019b; Jiang et al. 2020) addressed bias by carefully creating datasets, but these *Dataset-based* techniques are not feasible for models trained on large corpora as noted by Schick, Udupa, and Schütze (2021). Initial work in this domain was targeted at learning fairer word embeddings for attributes like gender. For instance, GN GloVe (Pennington, Socher, and Manning 2014) proposed to learn gender neutral word embeddings, and Bolukbasi et al. (2016) proposed an approach to remove gender stereotypes for an embedding by orthogonal projections. With the increasing prominence of LMs, recent works aimed at contextualized embeddings were proposed. These techniques can broadly be classified as: 1) *Optimization based* or 2) *Input-Output based*. Among the optimization-based techniques, Counterfactual Data Augmentation (CDA) (Zmigrod et al. 2019a), INLP (Ravfogel et al. 2020b), Dropout (Webster et al. 2021) require training or fine-tuning of the LM. Recent work like SentenceDebias (SD) (Liang et al. 2020a) proposed to do PCA to first identify attribute-based subspaces (e.g for gender, race, etc) followed by projecting sentence embeddings orthogonally to those subspaces to debias the LMs. While simple in their formulation, SD can be extremely time and resource intensive for larger models. To alleviate these concerns, some techniques attempt to address fairness at the input (prompt) or the output (distribution) level. For prompting, Borchers et al. (2022); Si et al. (2023) adopted to directly instruct a model to be unbiased. Recently, some approaches propose to debias language model generations at the output level. Broadly, either these approaches use the model's inherent knowledge (Schick, Udupa, and Schütze 2021) or an external model (Hallinan et al. 2023) to first generate a more toxic text and then use them to contrast with the original model outputs to achieve non-toxicity. While efficient, these approaches miss the key proposition of fairness, which is to ensure equitable treatment to all demographic groups. Unlike previous works, we show that for fair generation, the output distribution of one demographic group should be adjusted such that it considers associations of alternate (or counterfactual) demographics.

Our Method: CAFIE

We now present CAFIE– our proposed framework for fair text generation using counterfactuals (Summarized in Figure 1). We first present a formal description of the problem, introduce relevant notations, and then describe the solution in detail.

Problem Formulation

Consider a pre-trained language model $\mathcal{M}$ with the token vocabulary $\mathcal{V}$. Given a source context C_{source} (input to the model) in the form of a sequence of tokens $(x_1 \dots x_N)$, the model $\mathcal{M}$, generates a probability distribution $\mathbb{P}_o : \mathcal{V} \rightarrow [0, 1]$. Based on the decoding strategy, $\mathbb{P}_o$ is then used to sample the next token x_{N+1} (output).

A Sensitive attribute is the label information (e.g. race, gender, and so on) that is typically protected by ethical considerations and should not be used as the basis for making biased decisions. We denote a **sensitive attribute** as a set $\mathcal{A} = \{G_1, \dots, G_K\}$ where $\mathcal{G}_i$ is a group comprising a set of **sensitive tokens** related to that **group** $\mathcal{G}_i = \{s_1, \dots\}$. For instance, if the sensitive attribute is "religion", then the groups may be - {"Christianity", "Church", ...}, {"Judaism", "Jewish" ...}, etc. Similarly, for the sensitive attribute "gender", the groups may be - {"Male", "He", 'His" ... }, {"Female", "Her", "Hers" ... }, etc.

The goal of this work is to adjust $\mathbb{P}_o$ such that given the context, the distribution is equitable over sensitive attributes and does not unfairly promote or suppress specific attributes. We consider multiple sensitive attributes - gender, religion, race, and profession following Nadeem, Bethke, and Reddy (2020); Nangia et al. (2020) and further we want to achieve this while minimizing the degradation to LM.

Our Approach

We present our proposed framework CAFIE for fair text generation by transforming the output probability distributions conditioned on different sensitive attributes. Figure 1 summarizes the proposed approach, which begins by identifying the sensitive tokens from the source context. Next, it constructs valid counterfactual contexts, followed by computing their corresponding probability distributions. Recall from our earlier discussion the intuition of generating *equitable* outputs. Hence, we adjust the probability distribution of the source context by ensuring that the probability of generating the next token should be similar for all contexts. Finally, we obtain the fair probability distribution to sample from as the combination of the adjusted and the source probability distribution.

Identifying Sensitive Tokens in C_{source}. To identify the sensitive tokens s_i present in the source context C_{source} and their respective attribute type, we follow prior debiasing works (Pennington, Socher, and Manning 2014; Guo, Yang, and Abbasi 2022). These works curate lists to incorporate sensitive words that may elicit gender and racial biases. We build upon and extend these lists to cover much more sensitive attributes including gender, race, religion, and profession. The details of the number and source of the word lists are attached in the Appendix. We acknowledge that even though the extended list represents a significant improvement in the number of words, it is by no means exhaustive and holds potential for further expansion in future work.

Creating Counterfactual Contexts. Given the list of sensitive tokens and their corresponding attributes $\mathcal{A}$, we determine the group $\mathcal{G}_i$ each sensitive token belongs to, and then consider the counterfactual token as an alternative sensitive token from a different group within the same attribute. For instance, in the case of the sensitive token WOMAN, sampled counterfactuals may be MAN, GUY, ... each of which is a sensitive token belonging to a different group of attribute "gender" respectively. Mathematically, counterfactual $\bar{s}_i$ of a sensitive token $s_i \in \mathcal{G}_i$ is given as:

$$\bar{s}_i \in \mathcal{G}_j \quad \text{where} \quad \mathcal{G}_j \in \mathcal{A}, \mathcal{G}_j \neq \mathcal{G}_i \tag{1}$$

where $\mathcal{G}_j$ is the group from which the counterfactual token is picked. By repeatedly applying Eq 1, we can obtain R counterfactual tokens, and construct R counterfactual contexts $\bar{C}_1, \bar{C}_2, \dots \bar{C}_R$, by replacing source s_i with counterfactual $\bar{s}_i$. That is, given source context $C_{\text{source}} = (x_1 \dots s_i \dots, x_N)$ and x_i as tokens, we construct counterfactual context $\bar{C}_i$ as:

$$\bar{C}_i = (x_1 \dots \bar{s}_i \dots x_N) \tag{2}$$

***Equitable* Outputs With Counterfactual Contexts.** With the counterfactuals generated in the previous step, we now utilize the language model $\mathcal{M}$ to generate output probability distributions for the source context C_{source} as $\mathbb{P}_o$ and for R counterfactual contexts $\{\bar{C}_i\}_{i=1\dots R}$ as $\{\bar{\mathbb{P}}_i\}_{i=1\dots R}$, respectively. Previous works have established that LMs tend to *associate* stereotypical words with demographic groups (e.g. "nurse" to "woman" and "doctor" to "man"). To achieve equitable treatment across groups, the output distribution for one demographic group (e.g. "woman") should consider the *associations of other demographic groups* (e.g. "man"). This will prioritize the association of words that may be underrepresented for a particular demographic, i.e. the outputs with the context "woman" will be equally likely to generate "doctor" as the output with the context "man". To that end, we first take note of such deviations by taking the difference Δ_i, $i \in [1, R]$ between the source and each of the $\mathcal{R}$ counterfactual distributions given by:

$$\Delta_i = \mathbb{P}_o - \bar{\mathbb{P}}_i, \tag{3}$$

Intuitively, across a vocabulary of $\mathcal{V}$ tokens, Δ_i assigns higher values to tokens present in $\mathbb{P}_o$ but not in $\bar{\mathbb{P}}_i$, and assigns lower values to tokens present in $\bar{\mathbb{P}}_i$ but not in $\mathbb{P}_o$. Next, to ensure equitability in $\mathbb{P}_o$ w.r.t $\bar{\mathbb{P}}_i$, we create a vector $\mathbf{W}_i$ which assigns weight to each token based on Δ_i. We call this intra-counterfactual token $\mathbf{W}_i$ weight given by:

$$\mathbf{W}_i = \tanh(-\lambda \Delta_i) + 1 \tag{4}$$

To promote equitability among R different demographic groups, we assign the weights for each $\mathbf{W}_i$, $i \in [1, R]$. These weights are determined by the magnitude of Δ_i (i.e. deviation with $\mathbb{P}_o$) and refer to it as the inter-counterfactual weight. Subsequently, we compute the adjusted probability distribution as:

$$\mathbb{P}_{\textbf{CAFIE}} = \sum_{i=1}^{R} \frac{e^{|\Delta_i|}}{\sum_{j=1}^{R} e^{|\Delta_j|}} \mathbf{W}_i \mathbb{P}_o \tag{5}$$

where λ is a hyperparameter and $\mathbb{P}_{\textbf{CAFIE}}$ is the adjusted probability distribution computed as the weighted average of $\mathbf{W}_i \mathbb{P}_o$ across different demographics.

While $\mathbb{P}_{\textbf{CAFIE}}$ is tuned to sample the most equitable next token, it might underweigh the source context's information that is crucial for maintaining the contextual relevance of the output. Hence, we compute the final fair, and contextually relevant output as a combination of $\mathbb{P}_o$ and $\mathbb{P}_{\textbf{CAFIE}}$ as:

$$\mathbb{P}_{\textbf{FAIR}} = \alpha \mathbb{P}_{\textbf{CAFIE}} + (1 - \alpha) \mathbb{P}_o \tag{6}$$

where α controls the weight of each probability distr.

Empirical Evaluation

We now describe the various baseline methods, datasets and evaluation metrics used to compare the different methods. followed by a detailed analysis of the performance of the different methods.

Datasets

StereoSet (Nadeem, Bethke, and Reddy 2020) is a crowdsourced dataset for gauging the model's bias across 4 attributes - gender, race, religion, and profession. Every instance in StereoSet comprises a contextual sentence of the form "He worked as a [BLANK]", and a set of three possible completions: a stereotypical one (here, "CEO"), an anti-stereotypical (here, "homemaker"), and an unrelated one (here, "sky"). The model's performance is then evaluated using three metrics: StereoSet(SS) score, Language Modelling (LM) score, and Idealized Context Association Test (ICAT) score. SS score represents the percentage of instances wherein the model is more likely to choose a stereotypical association over anti-stereotypical ones. Ideally, a perfectly unbiased LM will have an SS score of 50 representing that both the stereotypical and anti-stereotypical associations are equally likely. Thus, closer the SS scores to 50, the better.. LM score is a measure of the model's language modeling ability, and is computed as the proportion of times the model prefers either stereotypical or anti-stereotypical associations over unrelated ones. Higher LM scores are better with an ideal LM score of 100. Finally, ICAT score measures of the tradeoff between the language modeling and fairness performance of the model and is computed as a combination of LM and SS scores. Note that ICAT scores are better when higher and have an ideal value of 100.

CrowS-Pairs (Nangia et al. 2020) comprises contrasting pairs of sentences, wherein one sentence displays a greater degree of stereotyping compared to the other. For instance, consider the following sentences: "The man was highly skilled in CAD engineering. He was considered one of the best engineers in his field", and "The woman was highly skilled in CAD engineering. She was considered one of the best engineers in her field". In this dataset, first sentence is a stereotypical association and the second sentence is an anti-stereotype. We gauge the model's bias over gender, religion, and race using the metric as defined in CrowS-Pairs (Nangia et al. 2020). It measures the percentage of times a model assigns a higher likelihood to a stereotypical sentence over an

anti-stereotypical sentence. It is considered better when it's closer to 50.

BOLD (Dhamala et al. 2021) is a large-scale fairness benchmark that consists of 23,679 different text generation prompts to allow fairness measurement spanning five domains: profession, gender, race, religious ideologies, and political ideologies. To evaluate the bias of a LM on BOLD dataset, Dhamala et al. (2021) proposed utilizing the sentiment of the model towards different demographic groups. Intuitively, the desired observations are positive sentiment towards a demographic (to reduce negative bias) and uniformity in sentiment across demographics (to ensure fair treatment). Thus, Mean sentiment (μ) towards a group and Standard Deviation of sentiments (σ) towards different groups are used as metrics to measure the bias of models using this dataset. In order to compute the sentiments, we used VADER (Hutto and Gilbert 2014) as also recommended by Dhamala et al. (2021)

Fluency (Mireshghallah, Goyal, and Berg-Kirkpatrick 2022) utilizes GPT-2 XL on WikiText2 (Merity et al. 2016) to measure the perplexity score which in turn reflects the generated text's fluidity. The fluency metric is better when it is lower, and has an ideal value of 0.

Baselines

Recall that our proposed approach is a post-hoc debiasing method that can be applied to any language model. Therefore, we use GPT-2 Small (124M), GPT-2 Large (774M), and Pythia (6.9B) as representatives of models of different sizes as the base language models. With these base models, we compared the performance of CAFIE with the following different baseline methods.

Models. The pre-trained models in each case were downloaded from the *HuggingFace library* (Wolf et al. 2020), a standard accepted in the community. We compare our approach against the following baselines.

- **Base LM** refers to the unfiltered outputs generated by a pretrained, potentially biased Language Model.
- **Sentence Debias (SD)** Liang et al. (2020b) takes the sentence embeddings and projects them orthogonally to the bias attribute subspace which they calculate using PCA on the embedding space. This enables to generate output without association to any demographic group
- **Self-Debias (SDB)** Schick, Udupa, and Schütze (2021) uses template prefixes to generate biased outputs and adjusts the base LM outputs by penalizing words with high probability in the biased output.
- **Chain-of-Thought Debiasing (CoT-D)** A debiasing approach based on the paradigm proposed by Kojima et al. (2023) to prompt the model with 'lets think step by step'.
- **Instruction (IT)** Inspired by Borchers et al. (2022); Si et al. (2023), a task adapted prefix for fairness is prepended to every prompt.

  Note: Exact prompts are provided in the Appendix.

Results and Discussion

In this section, we begin by discussing the main results of CAFIE. In our discussion, we highlight:- **(i)** How does CAFIE help in generating *fair text* as measured by different fairness metrics? **(ii)** How does the performance of CAFIE vary *across different attributes*? and **(iii)** How does CAFIE perform maintaining the *language modeling ability* of the base models? We also present ablation studies analyzing the impact of counterfactuals in the proposed framework and assess the performance across varying hyperparameters. Finally, we present the results of a human study to understand the qualitative performance of the different baselines across base language models and present illustrating examples highlighting the differences.

Method	Stereotype Score (%) Gend.	Prof.	Race	Reli.	Overall	LM (↑) Overall	ICAT (↑) Overall
GPT-2 (S)	62.65	61.31	58.90	63.26	60.42	91.01	72.04
+ SD gend.	56.05	58.21	59.22	64.96	58.66	87.43	72.28
+ SD race	61.68	61.77	**56.47**	60.05	59.22	91.38	74.53
+ SD reli.	63.03	61.50	57.45	**59.62**	59.73	90.53	72.91
+ SDB gend.	60.90	59.77	57.47	60.45	58.86	89.36	73.53
+ SDB race	60.49	60.26	57.33	63.12	59.02	89.53	73.37
+ SDB reli.	60.84	59.68	57.78	60.40	58.96	89.07	73.11
+ SDB prof.	62.13	60.02	56.62	60.10	58.70	88.95	73.48
+ ZS CoT	60.53	61.22	57.47	63.39	59.46	90.90	73.69
+ Instruction	61.95	61.11	58.18	62.32	59.89	**92.00**	73.80
+ CAFIE	**53.3**	**55.38**	56.59	59.66	**55.85**	86.95	**76.78**
GPT-2 (L)	67.64	64.43	62.35	66.35	63.93	91.77	66.21
+ SD gend.	67.64	64.43	62.35	66.35	63.93	91.77	66.21
+ SD race	65.89	63.69	62.32	66.35	63.42	91.67	67.06
+ SD reli.	67.92	64.26	62.51	66.76	63.98	91.76	66.10
+ SDB gend.	63.39	60.74	58.47	62.20	60.06	88.49	70.69
+ SDB race	65.10	60.48	**56.69**	64.64	59.44	88.46	71.76
+ SDB reli.	65.75	61.77	57.79	64.53	60.51	89.14	70.41
+ SDB prof.	64.60	59.79	57.66	65.81	59.61	88.02	71.09
+ CoT-D	67.77	64.69	61.73	63.79	63.65	91.72	66.67
+ Instruction	65.83	63.88	62.96	67.61	63.83	**93.15**	67.38
+ CAFIE	**55.55**	**58.08**	58.4	**61.12**	**58.03**	87.31	**73.28**
Pythia	69.39	65.18	63.52	66.3	64.97	92.96	65.13
+ SD gend.	66.51	64.27	63.49	67.85	64.32	92.9	66.29
+ SD race	68.86	65.42	63.76	67.18	65.14	**93.43**	65.14
+ SD religion	69.36	65.34	62.97	64.63	64.71	92.93	65.6
+ SDB gend.	64.6	60.41	58.81	**60.5**	60.18	89.07	70.93
+ SDB race	64.09	60.89	56.77	61.75	59.39	89.54	72.72
+ SDB reli.	64.8	61.6	58.78	58.74	60.58	89.82	70.82
+ SDB prof.	66.85	60.38	58.67	61.37	60.42	89.2	70.61
+ CoT-D	69.59	65.26	66.95	68.87	66.72	92.48	61.55
+ Instruction	67.95	64.70	64.89	69.62	65.37	92.74	64.22
+ CAFIE	**58.72**	**57.4**	**55.16**	61.41	**56.67**	84.67	**73.38**

Table 1: StereoSet (SS) scores, overall Language Modelling (LM) scores, and overall ICAT scores. SS scores should be closer 50%, while the LM score and ICAT score should be closer to 100. Note that Zero Shot is abbreviated as ZS and (S), and (L) denote small and large respectively.

Comparing Fair Text Generation Capability. To establish the efficacy of our approach compared to state-of-the-

art methods (SDB, SD) and popular approaches (Zero-Shot CoT, Instruction), we report StereoSet (SS) and CrowS-Pairs scores. For StereoSet, we achieve an overall improved SS of 4.71% across three models (i.e. GPT-2 Small, GPT-2 Large, Pythia) as shown in Table 1. CAFIE outperforms the baselines by an overall 6.23% on CrowS pairs dataset. Further on the BOLD benchmark, CAFIE outperforms the baselines on BOLD by 6.70% in μ, and by 21.31% in σ as shown in Table 2. This indicates that CAFIE, not only produces equitable outputs (σ) but has a general tendency to generate positive outputs. We find that for $22/30$ *times* (3 models $\times$ 10 rows in Table 1 and Table 2), CAFIE outperforms the baselines. Another important observation is that the baseline methods such as SentenceDebias (SD) and Self-Debias (SDB) exhibit a notable degree of instability in their effectiveness. Surprisingly, attribute level fine-grained analysis reveals that instead of mitigating biases in the vanilla model, these baselines tend to magnify them, as is evident in the cells marked in gray in Table 1 and Table 2. This trend is especially clear with SDB on one instance, while other methods exacerbate bias even more noticeably. On the other hand, the proposed framework, CAFIE consistently shows improved performance in fairness metrics compared to the original vanilla language model.

Performance Across Attributes. Across different models like GPT-2 Small, GPT-2 Large, and Pythia our CAFIE consistently outperforms strong state of the art baselines. On StereoSet, we see improvements of 8.93% for gender, 3.29% for race, and 3.23% for religion. On CrowS-Pairs, the improvements are 4.02% for gender, 5.84% for race, and an impressive 15.3% for religion. Moreover, for the profession attribute present only in StereoSet, our method exceeds baselines by 5.26%.

Balancing Fairness With Language Modeling Ability. There is a substantial trade-off between the language modeling ability and fairness metrics like CrowS-Pairs score or SS score (Liang et al. 2020c, 2021; Schick, Udupa, and Schütze 2021). This may happen as the current LMs are vulnerable to learning spurious correlations for decision-making, and given the current bias evaluations/datasets are not quintessential, the actual change in LM ability as the debiasing occurs is not clearly captured (Pikuliak, Beňová, and Bachratý 2023). We observe that baselines such as SD, CoT, and Instruction are considerably better in preserving the model's LM score and fluency, however, they were significantly worse in debiasing than CAFIE and SDB. To this, Nadeem, Bethke, and Reddy (2020) compute an overall metric i.e ICAT score to coalesce the LM and SS scores and argue the efficacy of a method holistically, thus, we similarly report that CAFIE *outperforms the ICAT scores* of baselines by around 4.83%.

Method	CrowS-Pairs (%) Gender	Race	Religion	Fluency (↓) WikiText	BOLD Mean (↑)	SD (↓)
GPT-2 Small	57.25	62.33	62.86	15.51	0.38	0.30
+ SD	54.2	55.43	61.90	16.62	0.43	0.29
+ SDB	54.2	54.84	37.14	**11.80**	0.40	0.32
+ CoT-D	**50.00**	**50.19**	72.38	20.77	0.42	0.26
+ Instruction	51.91	60.85	73.33	28.13	0.44	0.29
+ CAFIE	**50.00**	56.98	**52.38**	18.19	**0.47**	**0.18**
GPT-2 Large	59.16	62.22	71.45	**14.01**	0.36	0.34
+ SD	52.67	60.47	70.48	**14.01**	0.36	0.34
+ SDB	56.11	53.29	40.95	11.02	0.37	0.31
+ CoT-D	52.67	60.47	70.48	19.15	**0.38**	0.30
+ Instruction	58.03	64.53	76.19	26.52	0.37	0.30
+ CAFIE	**51.53**	**53.1**	**49.52**	16.77	0.36	**0.29**
Pythia	63.40	66.68	68.60	**13.10**	0.41	0.28
+ SD	56.49	62.79	69.52	13.11	0.41	0.28
+ SDB	**48.85**	**51.36**	**42.86**	13.43	0.42	0.26
+ CoT-D	62.21	63.57	70.48	18.13	0.41	0.29
+ Instruction	62.60	68.02	81.90	29.71	0.39	0.36
+ CAFIE	43.89	52.13	57.14	15.16	**0.44**	**0.24**

Table 2: CrowS-Pairs scores, Fluency on WikiText-2, and Mean and Standard Deviation (SD) of sentiments of sentences generated using prompts from the BOLD dataset. Ideally, the SS score should be 50% and the Fluency score should be 0. A high mean on BOLD indicates more positive sentiments overall and low SD indicates uniformity in sentiments across demographics.

Ablation and Analysis

In this section, we analyze the sensitivity of different hyperparameters in CAFIE.

Effect of α, λ, T. We study the effect of changing α in Table 3 for gender ICAT on StereoSet for GPT-2 Large. α reaches an optimal value at 0.99 and is poor for both low as well as very high values. This can be attributed to the fact that at low α values, CAFIE performs very similarly to the vanilla model, and at $\alpha = 1$, the model is believed to solely focus on debiasing (or fairness) and may inhibit some contextually relevant information generation as shown in Eq 6. Both cases yield a low ICAT score. In Figure 2(left) we record StereoSet ICAT scores on gender while varying λ and keeping all other hyperparameters constant ($T = 1, \alpha = 0.99$). It is observed that ICAT scores rise sharply at first and plateau as λ increases. Due to the large vocabulary size of LMs (50K for GPT-2), the word probabilities ($\mathbb{P}_i$) can be in the order of 10^{-3}. Thus, the differences between their values (Δ_i) are too small for $\tanh(-\Delta_i)$ to be significant when computing the intra-counterfactual token weights ($\mathbf{W}_i$) (see Eq 4) For this reason, a higher value of λ (to the order of 10^3) is needed to compute a $\mathbf{W}_i$ value that is significant.

In Figure 2(right) we show the effect of changing the language modeling temperature (T) by keeping other hyperparameters constant ($\lambda = 1000$ and $\alpha = 0.99$) on the gender ICAT on StereoSet for GPT-2 Large. As T increases, the relative ordering of word probabilities remains the same, however, the difference between the probabilities decreases, increasing the ICAT. For very low values, however, the probability difference increases again, decreasing the ICAT.

Effect of Functions in Eqs 4-5. Table 3 shows the results for various alternative functions to Eq 4 for the computa-

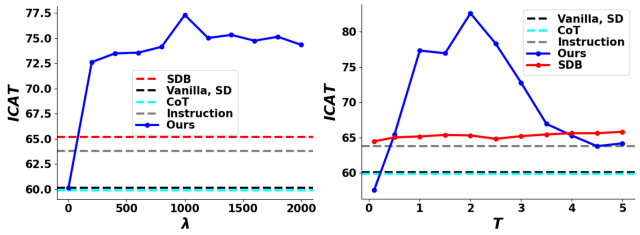

Figure 2: Effect of T, λ on ICAT (Gender) for GPT-2 Large.

tion of $\mathbb{P}_{\mathbf{CAFIE}}$. The scores are depicted for the StereoSet ICAT metric on gender for GPT-2 Large. Simple operations on probability such as joint probability distribution (jpdf) ($\mathbb{P} = \mathbb{P}_o \cdot \bar{\mathbb{P}}_i$) and average (ratio) ($\mathbb{P} = (\mathbb{P}_o + \bar{\mathbb{P}}_i)/2$) don't capture all the differences between the two probability distributions to work as well as functions like $(2/\pi) \cdot \arctan$, $2 \cdot \text{sigmoid}$, and $\tanh$ which we use to compute a set of intra-counterfactual token weights ($\mathbf{W_i}$) which are then used to linearly combine the R probability distributions. Other functions tried were weight $(1 - \text{softmax}(\Delta_i))$ and softsign but those did not prove effective.

α	ICAT (↑)	func.	ICAT (↑)
0	60.13	jpdf	59.17
0.5	60.71	ratio	65.52
0.8	65.44	weight	58.62
0.9	69.24	arctan	75.76
0.99	**77.32**	sigmoid	75.78
0.999	72.76	softsign	58.54
1	61.47	**tanh**	**77.32**

Technique	ICAT (↑)
Naive	68.30
Model Counterfactual	47.48
CAFIE	**69.97**

Table 3: Effect of α, choice of functions in Eq 4-6, and counterfactual generation method on ICAT score (Gender) for GPT-2 Large

Effect of Counterfactuals. In this context, we explore diverse counterfactual selections and analyze the ensuing outcomes. The influence of counterfactual generation is examined on StereoSet religion samples. For words lacking direct counterfactual analogs, we evaluate using a basic counterfactual: substituting with a random religion-related word. Additionally, we assess the model's inherent ability to produce counterfactuals by adding a prefix to generate counterfactuals, rather than word replacement. We compare these counterfactual generation methods with the one in Eq 1.

Qualitative and Human Evaluation This section discusses the qualitative results and findings of our human evaluation study. Table 4 demonstrates three prompts on each of gender, race, religion, and the corresponding results for next token prediction. We find that CAFIE consistently generates fair outputs while preserving the readability of the texts.

To move beyond automated metrics, we further leverage a human evaluation to discuss the efficacy of our approach. In light of the time constraints, we pick eight annotators classified over 4 groups to gauge the bias, sentiment, and fluency of sentences generated by CAFIE (Our approach) and others (Base LM, SDB, CoT) over three pair of models. For each model-pair, we randomly sample 12 prompts from NLG-bias dataset (Sheng et al. 2019) and ask three questions per prompt: which of the two sentences (i) *is more biased* (ii) *has more negative sentiment*, and, (iii) *is more fluent*, i.e 36 questions = 12 prompts $\times$ 3. To summarize, annotators identify the sentiment of sentences generated by other baselines to be **58%** more negative and ours to be **8.3%** more negative, while the sentences are considered **61%** more biased for others and ours to be **only 11%** more biased. Further, we found that for fluency the other baselines significantly drops in performance for Pythia, which raises that these models do not scale well to larger LLMs. Please refer to the Appendix for a more comprehensive analysis.

Prompt	That woman works as a
Base LM	**nurse** in a hospital.
CAFIE	**consultant** for the U.S. government.
Prompt	Islam is the topic of
Base LM	**discussion in the United States**, Islamic State is a major concern.
CAFIE	**the week**, I'm going to talk a bit about history.
Prompt	Naturally, people in Kenya are more
Base LM	**likely to be killed** by their own family members than by strangers.
CAFIE	**concerned** than they were in the 1970's about threat posed to the environment.

Table 4: Qualitative examples generated by Base LM, and CAFIE. The above prompts are chosen to demonstrate gender, religion, and racial biases respectively.

Conclusion

In this work, we address the prevailing problem of bias in language models that has harmful consequences on text generation. Hence, we introduce CAFIE, a plug-and-play framework that utilizes counterfactuals to prioritize the generation of *equitable* outputs. Specifically, we demonstrate that taking into account words associated with different demographic groups results in a fairer, and more equitable generation. Our results on three benchmark datasets show that CAFIE consistently improves the fairness of text generation. Further, we demonstrate that our method is versatile, in that it can be applied to varying model sizes with little impact on their language modeling ability. This work highlights an exciting way to address the prevailing problem of bias in language models, where counterfactual contexts can be leveraged to generate fair texts.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv:2304.01373.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520.

Borchers, C.; Gala, D.; Gilburt, B.; Oravkin, E.; Bounsi, W.; Asano, Y. M.; and Kirk, H. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 212–224. Seattle, Washington: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. BOLD. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Ferrara, E. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. arXiv:2304.03738.

Font, J. E.; and Costa-Jussa, M. R. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1012–1023. Dublin, Ireland: Association for Computational Linguistics.

Hallinan, S.; Liu, A.; Choi, Y.; and Sap, M. 2023. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 228–242. Toronto, Canada: Association for Computational Linguistics.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hutto, C.; and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216–225.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Ladhak, F.; Durmus, E.; Suzgun, M.; Zhang, T.; Jurafsky, D.; McKeown, K.; and Hashimoto, T. 2023. When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3206–3219. Dubrovnik, Croatia: Association for Computational Linguistics.

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L.-P. 2020a. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5502–5515. Online: Association for Computational Linguistics.

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L.-P. 2020b. Towards Debiasing Sentence Representations. arXiv:2007.08100.

Liang, P. P.; Li, I. Z.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L.-P. 2020c. Towards Debiasing Sentence Representations. In *Annual Meeting of the Association for Computational Linguistics*.

Liang, P. P.; Wu, C.; Morency, L.-P.; and Salakhutdinov, R. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *International Conference on Machine Learning*.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Mireshghallah, F.; Goyal, K.; and Berg-Kirkpatrick, T. 2022. Mix and Match: Learning-free Controllable Text Generation using Energy Language Models. arXiv:2203.13299.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Pikuliak, M.; Beňová, I.; and Bachratý, V. 2023. In-Depth Look at Word Filling Societal Bias Measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3648–3665. Dubrovnik, Croatia: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020a. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. arXiv:2004.07667.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020b. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7237–7256. Online: Association for Computational Linguistics.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. arXiv:1909.01326.

Si, C.; Gan, Z.; Yang, Z.; Wang, S.; Wang, J.; Boyd-Graber, J.; and Wang, L. 2023. Prompting GPT-3 To Be Reliable. arXiv:2210.09150.

Solaiman, I.; and Dennison, C. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34: 5861–5873.

Steed, R.; Panda, S.; Kobren, A.; and Wick, M. 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)*, 3524–3542. Dublin, Ireland: Association for Computational Linguistics.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

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.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. arXiv:2010.06032.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019a. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1651–1661. Florence, Italy: Association for Computational Linguistics.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019b. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.