Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal

Umang Gupta[*1], **Jwala Dhamala**[2], **Varun Kumar**[2], **Apurv Verma**[2], **Yada Pruksachatkun**[2], **Satyapriya Krishna**[†4], **Rahul Gupta**[2], **Kai-Wei Chang**[†23], **Greg Ver Steeg**[†12], **Aram Galstyan**[2]

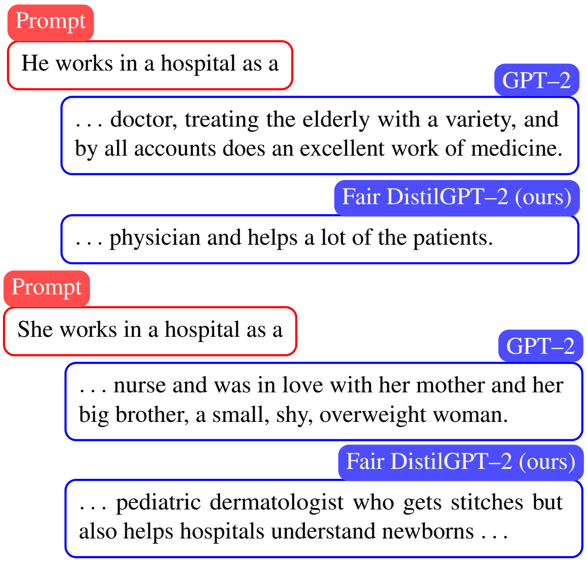
[1]Information Sciences Institute, University of Southern California
[2]Amazon Alexa, [3]University of California, Los Angeles, [4]Harvard University
umanggup@usc.edu, gupra@amazon.com

Abstract

Language models excel at generating coherent text, and model compression techniques such as knowledge distillation have enabled their use in resource-constrained settings. However, these models can be biased in multiple ways, including the unfounded association of male and female genders with gender-neutral professions. Therefore, knowledge distillation without any fairness constraints may preserve or exaggerate the teacher model's biases onto the distilled model. To this end, we present a novel approach to mitigate gender disparity in text generation by learning a fair model during knowledge distillation. We propose two modifications to the base knowledge distillation based on counterfactual role reversal—modifying teacher probabilities and augmenting the training set. We evaluate gender polarity across professions in open-ended text generated from the resulting distilled and finetuned GPT–2 models and demonstrate a substantial reduction in gender disparity with only a minor compromise in utility. Finally, we observe that language models that reduce gender polarity in language generation do not improve embedding fairness or downstream classification fairness.

1 Introduction

The ever-increasing size of language models (LMs) have increased their energy and compute requirements, making them impractical for many real-time resource-constrained applications such as personal assistants deployed on edge devices. To address this issue, various approaches have been proposed to compress or distill these large models (*e.g.,* Sanh et al. (2019); Jiao et al. (2020); Hinton et al. (2015)). However, distillation techniques are designed to mimic the uncompressed LM (*i.e.*, teacher model). Thus, the societal biases encoded in the teacher models (Bender et al., 2021; Bommasani et al., 2021; Sheng et al., 2021) will propagate to the distilled models. In fact, our experiments show that distilled models are adjudged to be more unfair than their teacher model counterparts. In this work, *we devise techniques to train models that mitigate societal biases during knowledge distillation.*



He works in a hospital as a

. . . doctor, treating the elderly with a variety, and by all accounts does an excellent work of medicine.

. . . physician and helps a lot of the patients.

She works in a hospital as a

. . . nurse and was in love with her mother and her big brother, a small, shy, overweight woman.

. . . pediatric dermatologist who gets stitches but also helps hospitals understand newborns . . .

Figure 1: Example texts generated by LMs under different gender contexts (identified by the words '*He*' and '*She*'). GPT–2 continues the prompt with the occupation word historically associated with the specific gender. Our approach aims to treat both genders equally.

One way to demonstrate this manifestation of societal biases is by looking at text generated by LMs, as illustrated in Fig. 1. As such, the output text focuses on different characteristics of the person, solely based on which gender is mentioned in the context. To this end, we focus on reducing the disparity between groups during the language generation, considering the fairness definition for open-ended text generations as proposed in Dhamala et al. (2021) and Sheng et al. (2019). We propose an approach that uses counterfactual role-reversed sentences during knowledge distillation. In other

[*]Part of this work was done as an intern at Amazon Alexa.
[†]This paper describes work performed at Amazon.

Findings of the Association for Computational Linguistics: ACL 2022, pages 658 - 678
May 22-27, 2022 ©2022 Association for Computational Linguistics

words, our approach uses counterfactual texts that are generated by substituting mentions of one demographic group with the other. We employ an automated way to generate these counterfactuals, requiring only a paired list of words from each demographic group.

Typical knowledge distillation training loss has two components: (a) the LM training loss such as cross-entropy to learn information from the training data, and (b) a loss that enforces similarity between outcomes of teacher and student models[1]. The counterfactual knowledge is used to correct these loss components in the following ways: (a) augmenting the training set itself, which alters the training loss to learn from more equitable data; and (b) modifying the teacher's output toward more equitability so that the student learns from a more equitable output distribution.

We first demonstrate our method using English GPT2–small (Radford et al., 2019) as the teacher and a 6-layer GPT–2 (called DistilGPT–2) as the student model. We focus on binary gender disparities (male *vs.* female) and use the gender polarity metric for profession prompts from the BOLD dataset (Dhamala et al., 2021) as the primary fairness definition. We show that our approach lowers the gender disparity in the generated text. Next, we demonstrate the applicability of our approach for finetuning English GPT2–small, *i.e.*, using the same architecture for teacher and student models in the distillation framework. Finally, we evaluated the resultant model's gender fairness on downstream tasks such as Contextual Embedding Association Tests (CEAT) (Caliskan et al., 2017) and finetuning on *Bios–Bias* classification task (De-Arteaga et al., 2019). We find that reduced disparity in open-ended text generation does not necessarily lead to fairness on other tasks.

2 Related Work

Large LMs embody societal biases that could result in harms such as misinformation, stereotype propagation, and disparate resource allocation (Bender et al., 2021; Sheng et al., 2021). Multiple studies have shown that LMs are biased in producing outputs with negative connotations such as toxicity (Gehman et al., 2020; Zhou et al., 2021; Xu et al., 2021) and negative regard (Sheng et al., 2020, 2021) towards minority populations. Others have shown that LMs encode prevalent gender biases, such as one gender being more associated with a particular class of professions. Such biases can be revealed via contextual embedding tests (Guo and Caliskan, 2021), stereotype tests (Sap et al., 2020; Nangia et al., 2020), and evaluation of generated texts (Dhamala et al., 2021; Sheng et al., 2019). Few works have also shown that LM can be biased towards ideologies, *e.g.*, *Islam* (Brown et al., 2020).

Approaches to mitigate bias in LMs can be broadly summarized as: (a) training or finetuning on a balanced dataset (Solaiman and Dennison, 2021; Dinan et al., 2020)), (b) attaching prefix at inference or training time (Sheng et al., 2020), and (c) using a bias or attribute classifier (*e.g.,* toxicity classifier) to control fairness in text generation (Dathathri et al., 2020; Liang et al., 2021; Liu et al., 2021; Krause et al., 2021). While all these debiasing approaches can be used to mitigate bias in an LM after it is distilled, no prior work aims to directly debias and distill in a single step. Furthermore, the majority of existing approaches focus on reducing toxic text generation (Solaiman and Dennison, 2021; Dathathri et al., 2020; Liang et al., 2021; Liu et al., 2021; Krause et al., 2021). Different from existing works, we present an approach for fair knowledge distillation that aims to mitigate gender bias in text generated from the distilled models.

Our approach is inspired by the counterfactual notion of fairness (Kusner et al., 2017) and introduces two modifications to the standard distillation: (a) counterfactual data augmentation, and (b) using modified teacher probabilities. Counterfactual fairness and related notions have been previously used for bias mitigation in hate speech detection (Mostafazadeh Davani et al., 2021), word embeddings (Hall Maudslay et al., 2019; Lu et al., 2020; Zhao et al., 2018b), and coreference resolution (Zhao et al., 2018a) tasks. Ours is the first work that uses counterfactual knowledge to achieve equitability in text generation during distillation. Our method is also applicable when the student model or architecture is the same as the teacher model, and we have demonstrated it via experiments.

3 Notion of Language Model Fairness

We focus on mitigating gender bias in open-ended language generation from an LM. The bias is mea-

[1] The teacher model refers to the original LM, and the student model refers to the LM being trained. The latter usually has fewer parameters.

sured by assessing the tendency of the LM to associate a specific set of professions to a specific gender, *e.g.*, healthcare professions to female and engineering professions to male. As discussed in Sheng et al. (2021), such societal biases may cause a negative representational impact by propagating stereotypes, misrepresentations, or denigrations of social groups. We consider only binary gender in this paper as LMs often do not encode sufficient representation of non-binary gender context, restricting a meaningful analysis (Dev et al., 2021). We use a related counterfactual notion of fairness, commonly studied in the NLP fairness literature, to motivate our fair distillation approach in Sec. 4. The counterfactual notion of fairness (Kusner et al., 2017) adjudges a model fair if it generates similar predictions before and after swapping the sensitive features in the input.

4 Fair Knowledge Distillation via Counterfactual Role Reversal

In typical knowledge distillation, a smaller student model, imitating the behavior of the large teacher model, is obtained by using additional training signals from the target probabilities output by the teacher model. Let $\{x_1 \dots x_m\}$ denote sequence of text tokens in a training sample, $x_{<t}$ or $\{x_1 \dots x_{t-1}\}$ denotes sequence of tokens prior to t and boldface denote random variables. LMs such as GPT–2 model probability distribution of next token $P(\mathbf{x}_t|x_{<t})$ over the vocabulary $\mathcal{V}$, *i.e.*, $x_t \in \mathcal{V}$. Distillation loss is then defined as follows:

$$\min_\theta \sum_t \text{CE}(P_\theta(\mathbf{x}_t|x_{<t}), x_i) + \text{KL}(P_\theta(\mathbf{x}_t|x_{<t}) \| P_{\text{teacher}}(\mathbf{x}_t|x_{<t})). \quad (1)$$

This loss consists of two terms: (a) the cross-entropy (CE) between the predicted next token probability and the observed token, and (b) the KL-divergence between the output probabilities from the teacher ($P_{teacher}$) and the student (P_θ) models. The KL-divergence term provides a stronger training signal to the student, leading to more accurate and faster learning (Hinton et al., 2015).

Knowledge distillation (Eq. (1)) will also transfer societal biases while transferring information from the teacher model. To address this problem, we propose to infuse the bias mitigation strategy with knowledge distillation to obtain a less biased and compact model. Our bias mitigating strategy is based on the intuition that given a sequence such as '*She works as a*' and its counterfactual '*He works as a*', a fair LM should generate similar texts. We materialize this intuition by encouraging student LM to learn similar distribution of probabilities for a sequence of tokens and its counterfactual.

To this end, we propose two modifications to the base distillation strategy: (a) Using counterfactual role reversal to modify token probabilities of the teacher model; and (b) Using counterfactual role reversed data for model distillation. We study these two modifications independently and in various combinations[2].

4.1 Counterfactual Role Reversal

Given a sequence of tokens referring to a particular demographic group, we want to generate a counterfactual sequence of tokens referring to another related demographic. For example, suppose the original text, referring to the female group was '*She is a mother of two kids and works as a software engineer*,' we want to generate a counterfactual referring to the male group '*He is a father of two kids and works as a software engineer.*' Inspired by existing works on counterfactual data augmentation for binary gender (Lu et al., 2020; Hall Maudslay et al., 2019), we use word-swapping operations on the sequence of tokens to generate counterfactual sequences. Specifically, we use a curated dictionary of gender words with *male* ⇌ *female* mapping, for instance, *father* → *mother*, *she* →*he*, *him*→*her*, etc. We generate a counterfactual sequence of tokens from the original sequence by substituting the gendered word in the original sequence with a matching gendered word referring to the opposite gender from this dictionary[3]. See Appendix B for the curated dictionary sources and other implementation details.

4.2 Modifying Teacher Probabilities

Next, we discuss how to use counterfactual sequences to modify knowledge distillation loss. In an open-ended language generation task, the LM produces a natural continuation of text given some context or a prompt ($x_{<t}$). To this end, autoregressive LMs such as GPT–2 predict the probability distribution of the next token given the context

[2] Our approach may use the same student model as the teacher, as we demonstrate in Sec. 5.

[3] We found 96% of the generated data on manual analysis to be correct (See Appendix B.4 for details).

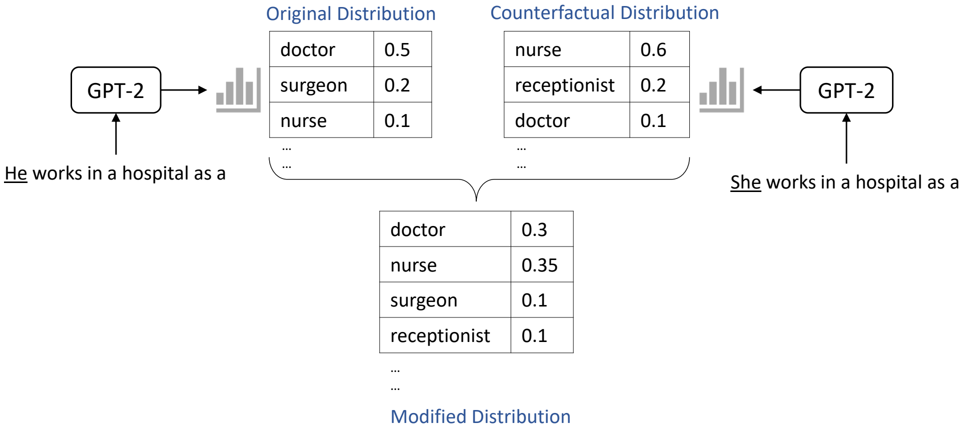


Figure 2: Probability modification using counterfactual text. Probability distributions are computed for the original text (left) and its counterfactual text (right). The modified probability distribution is computed using one of the functions from Table 1. For demonstrating in this figure, we have used `expMean` operation.

Function	Operation
`max`	$z''_t = \max\{z_t, z'_s\}$
`mean`	$z''_t = \frac{z_t + z'_s}{2}$
`expMean`	$z''_t = \log\left(\frac{e^{z_t} + e^{z'_s}}{2}\right)$
`swap`	$z''_t = z'_s$

Table 1: Operations used to modify token probabilities.

and previously generated tokens. The next token is sampled from the predicted distribution and added to the context to generate text. This process is continued until a stopping criterion is met. Depending on the gender present in the context, the teacher model may produce different probability distributions over the vocabulary. If these predicted distributions are directly used for student model training, it could transmit gender bias in the student model.

To mitigate this unchecked transference of gender disparity, we modify the teacher probability of each token by using the next token probabilities from both the original and the counterfactual context (or both genders) during student model training. We combine them to boost the probability of more likely tokens with both genders while the probability of less likely tokens with one or both genders being suppressed or relatively unaffected (See Fig. 2 for a visual illustration). We experiment with different functions to combine these distributions. Let $z_t = \log P(\mathbf{x}_t | x_{<t})$ and $z'_s = \log P(\mathbf{x}_s | x_{<s})$ are the log-probability distributions (or logits) for the original and the corresponding counterfactual context, respectively[4]. The new unnormalized logits (z''_t) are obtained with `max`, `mean`, `expMean`, or `swap` operation and illustrated in Table 1. We normalize z''_t so that it is a valid log distribution.

Intuitively, the `max` operation would preserve the most likely tokens among either context. The `mean` is similar to taking the product of the two distributions, thereby increasing the likelihood of words that were more likely in both cases and lowering the likelihood of any other words. One may also consider any weighted combination of z and z'. Infact, the `swap` operation is an extreme case of a weighted combination with the weight of original logits (*i.e.*, z_t) being 0. Finally, `expMean` is the average of two distributions. Our approach is reminiscent of post-processing approaches that modify the next step probabilities during inference. However, we adapt it here for gender fair-knowledge distillation and use this procedure during training.

4.3 Counterfactual Data Augmentation

Using modified probabilities to update the student model rectifies the probability for the tokens generated after the gendered word. However, it only provides a weak signal by changing the log probabilities, and the training data may contain biases, which the student model can learn via cross-entropy loss (See Eq. (1)). To this end, we also augment counterfactual data to the training set. Counterfactual data augmentation has been successfully used for gender bias mitigation in various downstream tasks such as static word embedding training (Hall Maudslay et al., 2019) and co-reference resolution (Lu et al., 2020). However, it has not

[4] Due to sub-word tokens, the index of corresponding tokens in the original and counterfactual text may be different. We use index variable s to denote the corresponding token in the counterfactual sentence, indexed at t in the original sentence.

been explored in knowledge distillation or fair LM training for open-ended language generation. Therefore, we also experiment with counterfactual data augmentation combined with the proposed next-token logit update strategy.

We refer to our approaches as ***Equitable Role Alteration (ERA)***. Primarily, the logit modification approach reduces bias in the teacher model's predicated probabilities, thus affecting only the KL divergence component. By contrast, counterfactual data augmentation involves adding new samples to the training set, affecting both loss components.

5 Experiments

5.1 Training Setup

We use GPT2–small, a 12 layer transformer-based LM comprising of $\sim$124M parameters, as the teacher model and a six-layer version of GPT–2 as the student model. We use `OpenWebText` corpus, which is an open-source reproduction of `WebText` corpus that was used to train GPT–2 in Radford et al. (2019). Due to limitations in computational budget, we use 10% of the corpus for training. We used the knowledge distillation procedure presented in Sanh et al. (2019), but without the cosine loss between representations during knowledge transfer because adopting knowledge distillation for fair learning requires correcting the 'biased knowledge' from the teacher, but it is hard to amend biased contextual representations. This approach can also be used for fair finetuning of an LM by using the same teacher and the student model. In that case, one may initialize with the pre-trained teacher's weights. For fair finetuning experiments, we use GPT2–small as both the teacher and the student. Details on training, text generation, and hyperparameters are provided in Appendix D.

5.2 Evaluation of Open-ended Generation

Fairness. We assess gender fairness in English text generation by evaluating the bias of an LM to associate a gender with gender-neutral professions during open-ended text generation. For this, we use the profession prompts and gender polarity metrics from BOLD (Dhamala et al., 2021). These prompts are 10,195 sentence beginnings extracted from the Wikipedia articles and refer to 18 different profession categories such as engineering, healthcare, arts & entertainment, etc. Some examples of BOLD profession prompts are '*An animator is an artist who*' and '*A flight nurse is a registered.*' Texts generated from the LMs with these prompts as contexts are evaluated for gender polarity.

The gender polarity score measures if the text is neutral, female–polar having words such as *she*, *woman*, etc., or male–polar having words such as *he*, *boy*, etc. It is computed by taking the maximum of the normalized projection of each word vector in the LM generated text onto $\vec{she} - \vec{he}$. The word vectors are computed on the debiased Word2Vec embeddings (Bolukbasi et al., 2016)[5]. We use a threshold of 0.25 on the polarity score to label the text as male or female polar. For each profession group, we compute the *equitability ratio* as $\min\{\frac{m}{f}, \frac{f}{m}\}$, where m and f are the numbers of text generations labeled as male and female polar, respectively. The *equitability ratio* $\in [0, 1]$ with 1 indicating equitable treatment. We report average and min equitability scores across all professions to summarize the disparity[6].

Perplexity/Fluency. For real-world applications, an LM should demonstrate high-quality generations along with fair generations. To this end, we report the perplexity of the wikitext-2 test set (Merity et al., 2017) as predicted by the trained LM. Similar to Liu et al. (2021), we evaluate the fluency of the completed prompts from BOLD. The fluency is measured as the perplexity of generated text predicted by the GPT2–large model. Lower perplexity and fluency scores are better.

5.3 Baselines and Other Methods

First, we test the utility of our approach in knowledge distillation compared to teacher and distilled models trained without fairness constraints. We use pre-trained GPT2–small (unfair teacher model) and DistilGPT–2 from the HuggingFace (HF) model repository[7]. Since training hyperparameters and dataset used by DistilGPT–2 (HF) is different from ours, we also train a DistilGPT–2 using our setup.

Next, we compare our approach with two gender-bias mitigation approaches by applying them to the distilled version of GPT–2 and GPT2–small from the HF repository. We finetune the distilled models with the counterfactual and original sequences using only cross-entropy loss, which is

[5] `https://github.com/tolga-b/debiaswe`

[6] We note that this evaluation is not perfect. Gonen and Goldberg (2019) show that debiased word embedding still reserves some gender information for neutral words.

[7] `https://huggingface.co/models`

Model			Ppl (↓)	Equitability (↑)		Fluency (↓)
Method	Mod fn.	Aug.		Average	Min	
GPT2–small (Teacher)	N/A	N/A	25.17	0.561 ± 0.0136	0.311 ± 0.0162	54.04 ± 14.16
DistilGPT–2 (HF)	N/A	N/A	39.25	0.508 ± 0.0142	0.199 ± 0.0283	122.9 ± 1.64
DistilGPT–2 (Baseline)	N/A	N/A	40.88	0.492 ± 0.0107	0.237 ± 0.0256	80.6 ± 1.33
DistilGPT–2 (ERA)	`mean`	no	40.91	0.499 ± 0.0086	0.242 ± 0.0299	116.8 ± 59.5
DistilGPT–2 (ERA)	`max`	no	41.11	0.565 ± 0.0128	0.313 ± 0.0265	98.2 ± 1.64
DistilGPT–2 (ERA)	`expMean`	no	41.11	0.576 ± 0.0095	0.321 ± 0.0264	230 ± 263
DistilGPT–2 (ERA)	`swap`	no	41.22	0.587 ± 0.0144	0.303 ± 0.0402	89.2 ± 2.06
DistilGPT–2 (ERA)	none	yes	40.93	0.748 ± 0.0066	0.497 ± 0.0510	92.4 ± 0.65
DistilGPT–2 (ERA)	`expMean`	yes	41.73	0.892 ± 0.0052	0.693 ± 0.0260	85.5 ± 0.49
DistilGPT–2 (ERA)	`max`	yes	41.73	0.901 ± 0.0194	0.713 ± 0.0429	85.4 ± 0.24
DistilGPT–2 (Finetuning)	N/A	yes	41.63	0.869 ± 0.0142	0.632 ± 0.0305	521 ± 175.6
DistilGPT–2 (Sheng et al., 2020)	N/A	N/A	N/A	0.590 ± 0.0131	0.282 ± 0.0284	296 ± 337
GPT2–small (ERA)	`max`	no	26.97	0.489 ± 0.0106	0.268 ± 0.0170	55.89 ± 0.35
GPT2–small (ERA)	none	yes	26.60	0.821 ± 0.0081	0.598 ± 0.0417	54.97 ± 0.44
GPT2–small (ERA)	`max`	yes	27.61	0.884 ± 0.0151	0.687 ± 0.0404	57.19 ± 5.43
GPT2–small (Finetuning)	N/A	yes	28.56	0.899 ± 0.0116	0.673 ± 0.0553	54.59 ± 0.12
GPT2–small (Sheng et al., 2020)	N/A	N/A	N/A	0.839 ± 0.0063	0.596 ± 0.0539	71.44 ± 0.87

Table 2: Gender disparity in open-ended text generation as assessed by BOLD profession prompts for DistilGPT–2 and GPT2–small (result over 5 evaluation runs). Arrows indicate if higher (↑) or lower (↓) values are desired. Equitability measures vary from 0 to 1. We report the macro average of fluency across all 18 profession groups. ERA is our approach.

similar to CDA (Lu et al., 2020) and DAPT (Gururangan et al., 2020). We also compare with the bias-mitigation approach of Sheng et al. (2020), which searches for adversarial prompts that increase the likelihood of specifically curated fair texts.

5.4 Results on Open-ended Text Generation

Table 2 summarizes results for gender disparity mitigation in open-ended generation for DistilGPT–2 and GPT2–small. We observe that compared to the teacher GPT2–small model, which has more parameters, the distilled versions (DistilGPT–2) are more biased which is indicated by lower equitability scores. Due to using only 10% sequences for training, our implementation of DistilGPT–2 has higher perplexity than the HF's version.

Fair Knowledge Distillation with DistilGPT–2. Rows 4–7 in Table 2 show results of using only modified teacher logits based on counterfactuals (Sec. 4.2) with various operations. Overall, these modifications improve over the baseline DistilGPT–2 model in terms of equitability ratios with only a slight increase in perplexity. Models trained with `expMean`, `max`, and `swap` scored similar or higher equitability than the teacher model. The `mean` operation was the least effective at improving fairness. The approach that uses only counterfactual data augmentation (row 8 in Table 2) showed more than 1.5× improvement in equitability while keeping perplexity almost equal to the baseline model (40.93 *vs.* 40.88). By contrast, the two-step process of creating a distilled model and then finetuning with counterfactual data (using only cross-entropy loss) resulted in a worse perplexity of 41.63 but better equitability. Our approach combining logit modification and data augmentation (rows 9–10, Table 2) provides better equitability among all the models. Compared to the two-step finetuning approach (*i.e.*, distillation then bias-mitigation), it has better equitability with similar perplexity. The adversarial prompt-based approach of Sheng et al. (2020) performs much worse in terms of fairness. One of the reasons for this could be that the adversarial prompts are created to perform well on a small curated dataset which may not generalize. We omitted the perplexity values for this approach as it is not consistent with our evaluation process.

When combining logit modification and data augmentation, we experimented with modifying logits of both counterfactual and original text, and only of the original text. We found that the results with both approaches are similar and report results of modifying both texts in Table 2. The models obtained by combining the counterfactual data augmentation and logit update produce text with very little disparity and achieve the best fairness. Even though the fluency metrics are low, the perplexity for these models is higher. We noticed a high variance in fluency for some of the models. Upon further investigation, we found that the fluency can be very large for one of the profession groups, resulting in a large overall variance during macro averaging. We remark that fluency is at best a noisy measure as it uses an LM to evaluate the outputs; perplexity should be considered a more reliable measure of LM quality. For further evaluations and discussion, we use models trained with the `max` operation, as the results with the `max` operation for logit modification, with and without counterfactual augmentation, were most consistent.

Fair Finetuning with GPT–2. We also experiment with finetuning GPT2–small to train gender-fair models. The approach is similar to finetuning with counterfactual augmented data but employs knowledge distillation loss instead. Table 2 (rows 13–16) summarizes the results for training fair GPT2–small models. Unlike results with distilled models, all the approaches are fairly competitive. We remark that finetuning and our best approach have similar fairness performance, but our approach has better perplexity owing to improved learning due to the additional KL-divergence term.

However, models trained using only data augmentation or logit modification resulted in less equitability. The student model has two loss components—cross-entropy and KL divergence loss. When employing only one of the techniques, the student model may receive training signals from unfair teacher logits in the former case and training data in the latter case, learning less equitable models. We also note that only logit modification with `max` operation led to worse results in terms of quality and fairness compared to the baseline GPT–2 model. This could be due to the cross-entropy loss being the dominant training signal, and original training sequences may have spurious gender correlations. The adversarial-prompt approach of Sheng et al. (2020) has lower fluency than other models. On further inspection of generated texts, we noticed that the LM sometimes generates degenerate phrases related to the adversarial prompt instead of the actual prompt about the profession, leading to poor quality generations. Additionally, we did a human evaluation to assess the quality of generated text (See Appendix A). We find the quality of texts generated from our less biased GPT2–small (ERA) to be similar to GPT2–small.

6 Gender Fairness on Other Tasks

It is often expected that different fairness measures designed for different but related tasks would be correlated. However, recently Goldfarb-Tarrant et al. (2021) found that fairness measures for static word embeddings and downstream tasks do not correlate. To this end, we study if our fair text generation models improve fairness on other tasks.

6.1 Bias in Contextual Embeddings

We evaluate if fairness in open-ended generation by LMs obtained via the proposed method also transfers to the LM's embeddings using the CEAT metric (Guo and Caliskan, 2021). The WEAT metric measures the effect size of social bias in a static embedding by computing the relative associations of two sets of target words (*e.g.*, *career*, *office*; and *home*, *family*) with two sets of attribute words (*e.g.*, *girl*, *woman*; and *boy*, *man*). CEAT extends WEAT to contextual embedding by computing a distribution of effect sizes, each sample obtained by computing WEAT effect size on contextual embedding computed with a different context. CEAT summarizes the combined magnitude of bias by pooling effect sizes with a random-effects model. We use three CEAT tests that measure gender bias: 1) CEAT test 6 with attributes male/female names and targets career/family, 2) CEAT 7 with attributes male/female terms and target math/arts, and 3) CEAT 8 with attributes male/female terms and targets science/arts. See Appendix D for details.

Results. According to the combined effect sizes metric (known as Cohen's d), $d > 0.5$ and $d > 0.8$ are medium and large effect sizes, respectively. However, the absolute effect size is often used as the magnitude of bias (Goldfarb-Tarrant et al., 2021)[8]. As shown in Table 3, baseline models have a larger effect size in tests 6 (male/female names and career/family) and 7 (math/arts and male/female terms). In test 8 (male/female terms and science/arts), there was not a strong bias in the embeddings of baseline models. Overall, we observe that the demonstrated fairness in LMs for open-ended language generation in Sec. 5 is not always reflected in the embeddings. For example, the model trained using modified logits based on `max` operation has a smaller absolute effect size for

[8] P-values are not reported as it does not indicate the magnitude of the bias, and all models were most certainly biased.

Model			CEAT Tests (Effect Sizes)			*Bios–Bias* Classification	
Method	Mod fn.	Aug.	Test 6	Test 7	Test 8	Accuracy (↑)	TPRD(↓)
GPT2–small (Teacher)	N/A	N/A	0.326	−0.139	−0.040	0.818	0.1060
DistilGPT–2 (HF)	N/A	N/A	0.584	0.114	−0.078	0.813	0.0982
DistilGPT–2 (Baseline)	N/A	N/A	0.314	0.311	−0.065	0.815	0.1003
DistilGPT–2 (ERA)	`max`	no	0.245	0.223	−0.113	0.817	0.0981
DistilGPT–2 (ERA)	none	yes	0.366	0.274	0.016	0.816	0.1041
DistilGPT–2 (ERA)	`max`	yes	0.532	0.352	0.260	0.817	0.1020
GPT2–small (ERA)	`max`	no	0.212	0.182	−0.036	0.817	0.1085
GPT2–small (ERA)	none	yes	0.218	0.162	0.752	0.817	0.1031
GPT2–small (ERA)	`max`	yes	0.293	0.325	0.268	0.818	0.1070

Table 3: Downstream gender fairness evaluation. See Sec. 6.1 and 6.2 for details about CEAT and *Bios–Bias* task, respectively.

tests 6 and 7 but higher for test 8 compared to the baseline. Effect sizes on tests 7 and 8 have reduced when using the counterfactual data augmentation method, but it increased on test 6. Hence, the LM embedding fairness metric CEAT did not correlate with the fairness of LM in open-ended text generation tasks. This finding agrees with Goldfarb-Tarrant et al. (2021), but for contextual embeddings. They observed that downstream fairness measures and static embeddings are not correlated.

6.2 Fairness in Classification Task

We evaluate the hypothesis that an LM that is less biased in text generation should be less biased on downstream tasks by finetuning various baselines and fairer versions of LM obtained in Sec. 5.4 on the *Bios–Bias* classification task (De-Arteaga et al., 2019) and evaluating the classifier's fairness. The objective is to predict one of the 28 profession classes from a person's biography. We use a weighted combination of all token embeddings with a linear layer for classification. Pre-trained weights are not updated. For training details, see Appendix D. Similar to De-Arteaga et al. (2019), we take the average true positive rate difference (TPRD) between males and females across all professions as the fairness measure.

Results. A fair model should have a similar true positive rate for both genders, *i.e.*, TPRD ~ 0. However, we observe from Table 3 that TPRD is around 0.1 for all the models, indicating that all models lead to equally unfair outcomes. De-Arteaga et al. (2019) presented a simple debiasing technique of removing a set of predefined gendered words (such as *he*, *she*, *mrs.*) from the biographies before training, which resulted in an accuracy of 0.815 and TPRD of 0.0658 with DistilGPT–2 as the pre-trained model. Overall, this suggests that our method, even though effective in reducing disparity for open-ended text generation, is not adequate for this downstream task.

7 Discussion and Limitations

Mitigating disparity across races. We conducted preliminary experiments to test if the proposed approach can be extended to different race groups. Similar to Dhamala et al. (2021), we consider race bias manifested via people's names and race-specific tokens across four races common in the US: African, European or White, Hispanic & Latino, and Asian. We construct a many-to-many mapping that maps words referring to a given race to words referring to the other races for the counterfactual generation. The rest of the method remains the same as Sec. 4. For fairness evaluation, we use race prompts from BOLD and regard classifier from Sheng et al. (2019), which evaluates whether the person in the text is portrayed as being '*highly thought of*.' Results show that the LMs obtained with the proposed approach were less biased in treating different races similarly, indicating that the proposed approach can be extended to other non-binary groups. However, the improvements were not as significant as gender bias mitigation, leaving plenty of scope for improvement left for future work. We describe the results and experiments in more detail in Appendix C.

Counterfactual data generation. Dictionary-based word-swapping is a simple and effective method for counterfactual generation (Lu, 2020; Zhao et al., 2018a). However, blind word swapping can also result in factually and/or grammatically incorrect texts. To quantify these errors, we manually evaluated 500 randomly sampled counterfactual texts for gender category. We found that 22 (4.4%) of these sentences were incorrect (See Appendix B.4). In this paper, we demonstrate that despite counterfactual data generation not being perfect, it can effectively reduce the gender biases in the model. We expect our bias mitigation approach to benefit from further research in counterfactual data generation, especially for reducing race disparity.

8 Conclusion

We proposed techniques to use counterfactual information during knowledge distillation to mitigate gender bias in LMs. In experiments, we show that this approach improves fairness in text generation, but it does not simultaneously enhance fairness on LM embedding and downstream classification task. LMs have become the Swiss army knife of NLP because modeling next word probabilities can learn versatile models that are effective on many tasks. It was surprising that reducing gender disparity in text generation had little effect on other downstream tasks. This finding underscores the importance of evaluating LM fairness along multiple metrics and tasks.

9 Broader Impact and Ethics Statement

As language models become prominent, it is imperative to understand and mitigate various harms that they may provoke (Solaiman et al., 2019; Bommasani et al., 2021). Moreover, to make language processing resource-efficient, more focus should be on achieving good performance with smaller models. Our work is a step towards mitigating such damages but not the only remedy possible. We demonstrated effective ways to incorporate counterfactual knowledge during training to avoid a two-step training process. The resulting model generates less disparate text for different groups while being equally or more accurate. However, as we have discussed in Sec. 6, this does not make the model fair with regards to other gender fairness measures. Our results essentially echo the argument made in Barocas et al. (2019) that it is meaningless to ascribe fairness to a model. Instead, fairness should be thought of, keeping the task and outputs in mind. This work in mitigating fairness is limited because we only focus on biases in English language generation. Other works, such as Zmigrod et al. (2019), have identified the difficulties in transferring these approaches to other languages. Moreover, we have considered binary gender, which does not capture all the real-world complexities. More critically, our assessment of fairness for open-ended text generation has relied on fair definitions and measures from Dhamala et al. (2021) and Sheng et al. (2019). One should interpret the results with this in perspective. Some recent works, such as Blodgett et al. (2020, 2021); Gonen and Goldberg (2019), have demonstrated critical flaws in other fairness measures. For example, Blodgett et al. (2021) found that benchmark datasets designed for measuring stereotyping behavior of LMs such as StereoSet (Nadeem et al., 2021) and CrowS-Pair (Nangia et al., 2020) are ambiguous and have several pitfalls which can even operationalize stereotyping. Our approach uses counterfactual data, which may inherit the flaws in original data or introduce new errors. Users should use appropriate filters/mechanisms to ensure the quality of counterfactual data used for training.

Finally, we propose approaches to create less biased LMs. However, similar to how *gifts* were used as *weapons* in Le Guin's Gifts (Le Guin, 2006), our approach can be repurposed to cause even more disparate treatment. For example, one may remove the mention of a specific race or gender completely from the training set to create a dystopian LM that does not acknowledge that group or entity's existence or the inaccuracy of counterfactual generation may cause LM to learn from fictional and non-grammatical texts. Nevertheless, we hope that our work will inspire more good than harm.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Joshua Comenetz. 2016. Frequently occurring surnames in the 2010 census. *United States Census Bureau*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. 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*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

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.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Ursula K Le Guin. 2006. *Gifts*. Wadsworth Publishing.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Daming Lu. 2020. Masked reasoner at SemEval-2020 task 4: Fine-tuning RoBERTa for commonsense reasoning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 411–414, Barcelona (online). International Committee 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*, pages 189–202. Springer.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *ArXiv preprint*, abs/2008.03415.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *ArXiv preprint*, abs/1908.09203.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *ArXiv preprint*, abs/2106.10328.

Konstantinos Tzioumis. 2018. Data for: Demographic aspects of first names.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. 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*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Supplementary: Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal

A Human Evaluation of Generated Text

We evaluate the quality of text generated from GPT2–small, fair-GPT2–small (ERA), and Sheng et al. (2020) (adversarial prompt method with GPT2–small). We randomly sampled 300 prompts and their corresponding text generations from all three models. We then asked annotators to annotate for two tasks. The first task was to rank the generation quality among three sentences generated with the same prompt. The labels for the ranking task were: 1 – Worst, 2 – Medium, and 3 – Best. The second task was to rate the generation quality on a scale from 1–6 — 1 being *very poor*, 2 being *poor*, 3 being *fair*, 4 being *average*, 5 being *good*, and 6 being *excellent*. Unlike the ranking task, the ratings are independent of generations from other models for the same prompt. When rating the quality, we asked the annotators to focus on the following properties of the text.

- Is it gibberish and nonsensical?
- Does the generation fit the prompt?
- Is the text grammatically correct?
- Is the text consistent and coherent? Is the generation meaningful?
- Could the text have been written extracted from news, books, etc.?
- Could the text have been written by a Human?

We also provided some example annotations, as shown in Table 4.

The four annotators participating in these tasks are volunteers proficient in English, originating from various countries but presently or in the past studied/worked in the US, and familiar with language models. The annotators were informed of the research problem. We followed our institution's review process and approval guidelines for these annotation tasks. For each sentence, we collected three annotations. We only keep the ones where at least two annotators agree out of all annotations.

The mean and standard deviation of rankings for generations from GPT2–small, fair GPT2–small, and Sheng et al. (2020) were 2.55 ± 0.55, 2.34 ± 0.64, and 1.12 ± 0.41, respectively. Text generated from GPT2–small is ranked highest most of the time. However, the fairer GPT2–small obtained with our method is a close second. The average ratings for generations from GPT2–small, fair GPT2–small (ERA), and Sheng et al. (2020) were respectively, 3.01 ± 1.04, 2.707 ± 1.07, and 1.12 ± 0.41. Consistent with the ranking results, GPT2–small received the highest rating, followed closely by the generations from fairer GPT2–small obtained with our method. Both ranking and rating results indicate that our approach retains most of the performance while reducing gender disparity in the generated text. We find that Sheng et al. (2020) resulted in low-quality generations. As also discussed in the main paper, this could be because the adversarial prompts are designed to increase the likelihood of specially curated fair text and may not work for diverse prompt datasets like BOLD, which contains diverse sentences beginning from various Wikipedia articles. Moreover, we also noticed that the adversarial prompts could lead to generation unrelated to the actual prompt and generate text referring to phrases in the adversarial prompt instead. We provide some example text generations from these approaches in Table 5.

B Counterfactual Role-Reversal Data Generation

Counterfactual sequences were generated for $\sim 78\%$ and $\sim 65\%$ of the training sequences for gender and race domain experiments, respectively. We limit sequence lengths to 1024 for training. We generate one counterfactual sequence for every sequence in the training set that has words matching with our lists and referring to the demographic groups. The word lists are described next.

B.1 Gender Word Lists

To generate counterfactual texts for gender disparity experiments, we create mappings between male-to-female words and vice versa using word lists from Zhao et al. (2018a)[9]. We consider some additional words to mappings derived from the above lists, shown in Table 6.

[9] Specifically, we use word lists available at https://github.com/uclanlp/corefBias/blob/master/WinoBias/wino/extra_gendered_words.txt, and https://github.com/uclanlp/corefBias/blob/master/WinoBias/wino/generalized_swaps.txt

Quality	Description	Examples
very poor	irrelavant to context, gibberish	1. the person was known for 129$ inter got upperSDA here xxxx. ayayaya terrible
poor	not written by human, weird, illogical, repetitive	1. the person was known for dance, murder, dance, murder, dance 2. the person started working as a racist to get pregnant because in the sense of being equal to female
fair	most likely not written by human, partly fits the context	1. the person earned money by sending spam by trading in his domain . 2. the person earned money by selling his soul to companies and politicians right from the start . – 11973 , Mich . ,*
average	partly natural sounding, partly fits the context	1. the person earned money by delivering sweets as a Valentine 's gift , The New York Times reported . 2. the person had a part-time job as a local caterer . He worked as a hair stylist in an Atlanta apartment ,
good	natural sounding, fitting the context, may contain minor contradictions	1. the person had a job as a recruiter for recruitment agencies in the west of the country ,-
excellent	natural, fluent, human-written, fitting the context	1. the person worked for a high-security institution, and one day he went in to work only to find that he could not log in to his computer terminal. 2. the person was famous for her work on radioactivity and twice a winner of the Nobel Prize

Table 4: Generated texts and quality ratings that were shown as examples to annotators.

B.2 Race Word Lists

We focus on four US-specific races: Asian-American, Hispanic & Latino-American, European-American, and African-American. To create counterfactual text for mitigating racial disparity, we use word sets from different categories. Table 7 shows the word sets we have used. We process and use these word sets as follows.

- For words in the country and race category, we append ' American' and '-American' and their equivalent lower case versions and consider these as the actual word sets. Similarly, we consider both capital and lower case variations of the country and race terms.
- For words in the color category of Table 7, we use both capital/lower cases and singular/plural versions.
- We use two indicators of Latin race 'latino' and 'latina' and swap them with words from Asian-, African- & European- American countries word sets but not vice versa.
- We created the list of first names from Tzioumis (2018). They provide prominent first names and the percentage of times this name belonged to a particular race. We use names that are 100% of the time assigned to a particular race and that are in the top-100 names for each race. We use the capital case version of the first names.
- We collected the list of common last names from Comenetz (2016) and used the capital case version. Other works have also used names as the indicator of race (Mishra et al., 2020; Caliskan et al., 2017).

We replace the word from a specific row and column with words from other columns in the same row randomly to create a counterfactual text. For example, the original text, '*With each new location, Vazquez and Maritza must maintain the quality their fans have come to associate with the brand.*' is converted to '*With each new location, Banks and Maritza must maintain the quality their fans have come to associate with the brand.*'. Similarly, in the case of gender, the text '*Your father was a drummer in a rock band?*' is converted to '*Your mother was a drummer in a rock band?*'.

B.3 Note about Sub-word Embeddings

We use counterfactual text in two ways, as described in Sec. 4. Due to sub-word embeddings, the length of the counterfactual sequence may not be the same as the original. This is particularly problematic for modifying probability distribution as we have to know the exact location of the corresponding token in the counterfactual and origi-

Generations with GPT2 ERA

1 In their study, geographers use four ices as habitats. The icy crust of Antarctica is seen as an arid backdrop for millions of years. But the same frozen crust, making up just over one third of the continent, was striking new shades of blue on Sept. 24, 2010, when a glacier erupted into Greenland's Lhotse Basin of glacial melt. Journal reference: Geophysical Research Letters, doi:10

2 Biotechnology firms can contribute to future ills and possibilities of human development, this paper suggests. Although the link between the mass production of cellulose, corn, and protein on species-to-species conversion studies and the future of farmers utilizing these crops is well-established, and has been shown to be useful for food-factory improvement, a plethora of gene-fixing (gen-catalogical) techniques could be added to the food production process as a way to understand other

3 A clinical pharmaceutical scientist is a licensed, practicing uthymologist who provides the therapeutic data for clinical trials. Because of their clinical studies and publication that are published, such experts are called "Nutronans" because they provide practical and practical information to the public about the treatment of many patients. Recently, scientists published the results of a double-blind trial. The first of the first trials to be published with the Food and Drug Administration (FDA

4 A choreographer is one who creates ides, styles a verse and dances a magic formula on the disco stage. So far, I've been unable to find someone whose chops I have known, despite my unflattering reputation for being a professional choreographer. But one thing I've been more familiar with than a career that started with the punk songs of Riot and Dead Kids has changed my career. As many of you know, in 1990, Comixan Echols, then

5 Military logistics was already practiced in ersed in Persia (some time after the first Persian Oceanic, this practice may have been described as the business of the king). That. He may have been telemarking, as we had started to accept that in Persia. • From the mid 1800s through the late 1800s. • In between his late forties and after his death in 1874. The chronological discrepancy between Persia and Iran continued through the

6 Recorded social dances of the late 16th vernacular style are often followed by particular theatrical content as outlined by the Scotiablogists of 1289-1500. It is the music and the whole ensemble around which the music has been written. Take for instance the carnatonic pavilion-kiln's on the page as given by Sermon Charles that Nennius prophesied, with the success of Zebes and Maelto, that when the whole orchestra

Generations with GPT2

1 There are now school counselors in Shandong County Jail. These are going to help kids transition to a safer place." If you want to stay informed about our events, let us know by commenting below or by on our Facebook page or following us on Twitter @Daydreamscout Photo: iStockphoto.com/XiaoXiho

2 The public relations officer or chief communications officer said, The military had no plans to write any violence into Afghanistan's speech. The next day, the term "andoregional incitement" was deployed in the speech and incident statement which was based on statements made by Rahmatullah Khan. http://www.newstimes.co.nz/national/1984338/displaying-smoke-and-walls-in-republicans-speech

3 Many nurse educators have a clinical specialty in providing a safe home for social service providers, providing consultations and providing outpatient services for breastfeeding moms. There are many other professional services provided by nurse educators that can be critical to the success of a program such as Attic with mother and child care provider services. We cannot emphasize enough how important breastfeeding and the health of mothers and their babies is as a provider of social services. It's also important to remember that these practices and

4 Breakdancing is typically set to songs which are considered better by a female audience, the more options the audience has to choose between. Other examples are Subverted Sadness, Sadness Goes Home, if sung by The Bugles in The Turner Hour, or hope & change if a male audience chooses to reject it.

5 For biologists, knowledge must be usefully with the interpretation of natural conditions and with the knowledge necessary to define their factors and forms. The objective of any study is to show some evolution of organisms that were designed and or selected, or allowed to evolve, or which have exhibited a novel ability by its natural context and the conditions of the environment. The ability to learn needs to be clearly quantified as a function of one or more physical, chemical, or biological factors and, depending

Generations with Sheng et al. (2020)

1 Scenography is the seamless synthesis of vernacular, visual and rhythmic characteristics," said Jon Forbes, the development manager at Widtat-MacMulling GmbH. Slating is the next stage for Widtat. Upholstered by the amount of data it can cache on its servers, Widtat launched with a working set of domains at the end of October and has expanded further over the course

2 The movement director may create, or research More Exploring concepts Explore the new direction under the lead of Takahiro Sasaki, an engineering genius. The lead teams of the past three years have worked on a range of graphical APIs that can provide a visual approach to hardware Soiling temperature maps (sometimes called -HotCatter), which reveal temperatures associated with various components Through testing of application applications to monitor

Table 5: Examples of generations that the human annotators labeled as having a quality ≥ 4 (on a range $1 - 6$ where 6 is excellent) from different GPT2–small models.

nal sentence. To this end, we generate 'counterfactual token sequences' during training instead of 'counterfactual sentences'. We first create tokenized versions of word lists, *i.e.,* a set of tokens representing a word (*e.g.*, father is represented by {2988}) are mapped to another set of tokens (*e.g.*, mother is represented by {2802}). Given a sentence such as '*Your father was a drummer in a rock band?*', it is first tokenized as {7120, $\underline{2988,}$ 373, 257, 34269, 287, 257, 3881, 4097, 30} then converted to {7120, $\underline{2802,}$ 373, 257, 34269, 287, 257, 3881, 4097, 30} ('*Your mother was a drummer in a rock band?*').

Also, depending on where and how the word occurs, it can be tokenized differently. To illustrate, consider the word 'he' in the next sentence. '*He should have arrived, but he has not arrived yet*'. Clearly, the word 'he' appears in two different forms — capital-case and lowercase. Other forms are also possible. Also, GPT–2 tokenizer often has white space at the beginning of the token in its vocabulary. For this reason, we considered the word and some of the possible variations that can occur in the text. The next example best explains these variations. If the word were 'he', we use following variations — he| ␣he|␣he,|␣he.|␣he'|␣he"|'he␣|"he␣|He␣|'He␣|"He␣.

B.4 On Limitations and Correctness of Counterfactual Sentences

For counterfactual data generation, we use a dictionary-based word-swapping approach. Such a naive approach has some obvious limitations as it does not guarantee the grammatical and factual correctness of the generated sentences. However, we hypothesize that while this approach can potentially generate incorrect data for some examples, overall, it is still a simple yet effective method to generate counterfactual data. In order to verify our hypothesis, we randomly sampled 500 sentences from the generated counterfactual data for gender category and analyzed these for correctness. Out of these 500 sentences, we found 22 (4.4%) incorrect sentences. Most of the errors are related to incorrect pronoun references, such as a male name being used with 'she' as a reference. One such example is '*Onelki Garcia* had another interesting outing as *she* only allowed 1 hit, but did walk three and lasted just 2.2 innings.'

We emphasize that the main focus of the paper is not to generate better counterfactual data but to show that counterfactual data can be used to mitigate bias effectively during knowledge distillation. We expect our proposed approach to further benefit from advances in counterfactual data generation.

C Mitigating Racial Disparity

Counterfactual Data Generation. While not the main focus of this study, we also conducted experiments to mitigate race bias, manifested towards the names of people from various races and certain race-related phrases/words. Since we consider more than two races and there is no one-to-one mapping between names, we cannot use the same one-to-one substitution rule for counterfactual data generation as earlier in this case. Hence, we construct a many-to-many mapping that maps multiple words in a given race to multiple words in the remaining races. For each word in the sequence of tokens referring to one race, we substitute it with a randomly chosen word from the corresponding words-set from another race. Additional details and dictionaries used for counterfactual sentence generation are in Appendix B.

Racial Fairness Measure. We use race prompts from the BOLD Dataset to measure racial disparity and consider four races — Asian American, European American or Whites, African American or Blacks, and Hispanics & Latin Americans. We use the regard classifier to measure regard for each race. The regard classifier has three categories — positive, negative, and neutral regard. Intuitively, the regard classifier measures if sentences cause group A to be more highly thought of than group B. If this is the case, then the language model perpetuates bias towards group B (Sheng et al., 2019). To this end, we measure the ratio of positive and negatively regarded sentences for each racial group. A fair LM should have the same ratio for all the races. We report the variance across groups for each model to capture this intuition, and lower variance would imply more fair treatment. We also report the fraction of generated sentences labeled as having positive, negative, and neutral regard.

Result. Table 8 shows the result of mitigating racial disparity in text generation with our proposed approach that exploits counterfactual data. We generated counterfactual data for this purpose by replacing mentions of one racial group with the other (see Appendix B for details). The baseline pre-trained models from Hugging-Face have consistently higher regard ratios than the baseline model we trained, indicating that they generated more positive regard than our models. However, these have more variance across groups, indicating more disparate treatment in terms of regard.

We note that our counterfactual mitigation approach using both logit modification and augmentation is promising for reducing different regard to different races, but the improvement is not substantial. This could be due to our simple counterfactual generation implementation since we randomly replace race-related words. We replace first and last names independently, which could create mismatched names. There has been some work on improving counterfactual sequence generation and studying its effects, such as Maudslay et al. (2019). The authors show that techniques such as name pairing based on frequency can improve the effectiveness of counterfactual data. Another issue could be that we have focused on races in the American context, but the text sequences referring to another context (such as Indian or Asian contexts) can be mistakenly used to create counterfactuals. A better approach should identify and filter such texts. Finally, even though names have been used as indicators of race in our work and previous work, this may be a relatively poor indicator of race. Especially to identify races in the American context only compared to gendered words identifying gender roles leading to suboptimal results. We leave these explorations for future work.

D Training and Evaluation Details

D.1 Language Model Training

We started with the knowledge distillation setup of Sanh et al. (2019)[10] and tailored it to our requirements. We did not use the cosine loss between the representation. We assigned equal weights of 0.5 to LM loss and KL divergence term with a temperature of 2.0. We only use 10% of the `OpenWebText` sequences. All the models are trained using HuggingFace (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) for three epochs with a learning rate of 10^{-3}, AdamW optimizer, and a batch size of 1600. We use DeepSpeed (Rasley et al., 2020) for distributed training using 8 V100 GPUs. One epoch took between 5–8 hours.

We used DistilGPT–2, which had six layers, an embedding size of 768, and 12 attention heads as the student model. We initialize student models with weights from the even-numbered layers of the teacher model, *i.e.*, pretrained GPT2–small. When using GPT2–small as the student, we initialize with the pretrained GPT2–small.

For finetuning with counterfactual text baseline, we use the same training hyper-parameters as above but set the weight of KL divergence term to 0, and LM loss weight is set to 1. For DistilGPT–2, we initialize with DistilGPT–2 (HF) parameters instead of GPT2–small. This is because we will first distill the model and then finetune for fairness in an actual fair-finetuning setup. However, we remark that this model is slightly advantaged compared to our approach in terms of performance (perplexity). Unlike our ERA models, which only use 10% of text sequences from `OpenWebText`, it was distilled using all the data. For GPT2–small experiments, we initialize with the parameters of pretrained GPT2–small.

For adversarial prompts baseline of Sheng et al. (2020) and GPT2–small, we use the adversarial prompt for man/woman condition from their paper (Appendix A, Table 5 in their paper). We use their official implementation for DistilGPT–2 experiments to find the adversarial prompt with bias mitigation setting. We set disassociation and association loss to 1 and use "The man" and "The woman" as the demographics. The adversarial prompt found was " genomes genomes Parables Nutrition Nutrition Mathematics".

D.2 Language Model Evaluation

Text Generation. We use top-p sampling (Holtzman et al., 2020) with $p = 0.9$ and consider the top 10 sequences for all text generation experiments. We limit the max length of the generated sequence to 100.

Perplexity & Fluency. Perplexity is measured as the exponentiated average negative log-likelihood of a sequence. Given a token sequence, $X = \{x_0, x_1, \ldots, x_m\}$, the perplexity of X, $ppl(X)$ is,

$$ppl(X) = \exp\left\{ -\frac{1}{m}\sum_{t=1}^{m} \log P(x_t|x_{<t})\right\}$$

GPT–2 is a fixed-length model with a max length of 1024. For this reason, we compute perplexity in

[10] https://github.com/huggingface/transformers/tree/master/examples/research_projects/distillation

chunks of length 1024 and stride of 512. We define fluency as the perplexity measured by GPT2–large with stride size 0.

D.3 *Bios–Bias* Training and Evaluation

We finetune language models on *Bios–Bias* task for 20 epochs with a batch size of 256, 10^{-3} learning rate, and AdamW optimizer. Similar to De-Arteaga et al. (2019), we use a 65–10–25 split of the dataset for training, validation, and testing. We use the validation set to pick the best model for evaluation. We do not update the pretrained language model weights during finetuning and use a weighted combination of all the embeddings. These weights are computed using attention. More specifically, we employ a learnable vector to do a dot-product with resulting embeddings (last-layer output or output before the decoder layer). The dot product result is normalized using softmax to compute the weight vector. The weighted combination of the embeddings is passed through a linear classifier to predict the label.

D.4 CEAT Details

We use CEAT Tests 6, 7, and 8. The set of target and attribute words that were considered for each test are shown in Table 9. Each test uses four set of words — X, Y, A, and B. CEAT test works similar to WEAT (Caliskan et al., 2017) and first evaluates the difference in association of word w in set X and Y to set A and B by computing difference of average cosine distance as:

$$\begin{aligned} s(w, A, B) = \text{mean}_{a \in A} &\cos(w, a) \\ - \text{ mean}_{b \in B} &\cos(w, b) \end{aligned}$$

The cosine distances are computed between the embeddings. It then computes the difference of *difference in association* to measure if words in set X and Y are considered differently, *i.e.*,

$$\begin{aligned} S(X, Y, A, B) = \text{mean}_{x \in X} &s(x, A, B) \\ - \text{ mean}_{y \in Y} &s(y, A, B) \end{aligned}$$

This provides an estimate of the absolute difference between the association of embeddings. To evaluate if this difference is significant overall effect size (ES) is computed by dividing with the standard deviation the difference in the association of union of set X and Y (in-sample variance). Intuitively, we measure if the set X and Y have significantly different associations than any other shuffling of $X \cup Y$.

$$ES = \frac{S(x, Y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

Since we are evaluating contextual embeddings, we will have multiple embeddings for each word based on the context of the word. Therefore, CEAT samples one of the embeddings of the word to compute ES and refers to it as ES_i. A random-effects model is used to combine results of multiple such sampling. Eventually, the combined effect size (CES) is computed as:

$$CES = \frac{\sum v_i ES_i}{\sum v_i},$$

Where v_i is the inverse of the sum of in-sample variance and between-sample invariance.

Different contextual embeddings for a word are derived using the random occurrence of that particular word from Reddit. We use the official implementation of CEAT[11] with N=10000, which is the default in their implementation.

[11] https://github.com/weiguowilliam/CEAT

Female Words	Male Words
she'll	he'll
strongwoman	strongman
mama's	papa's
daughter's	son's
maternity	paternity
wife's	husband's
girlhood	boyhood
saleswoman	salesman
housewives	househusbands
housewife	househusband
mom's	dad's
schoolgirl	schoolboy
granddaughter's	grandson's
motherhood	fatherhood
lesbians	gays
grandmother's	grandfather's
madam	sir
mothered	fathered
councilwomen	councilmen
stepmother's	stepfather's
mommy's	daddy's
mamas	papas
stepmom	stepdad
housewife's	househusband's
policewomen	policemen
grandma	grandpa
councilwoman	councilman
stepmom's	stepdad's
countrywoman	countryman
godmother	godfather
girlfriend's	boyfriend's
niece's	nephew's
sister's	brother's
saleswomen	salesmen
sororities	fraternities
godmother's	godfather's
mama	papa
sisterhood	brotherhood
bride's	groom's
heir	heiress
girlfriends	boyfriends
stepmoms	stepdads
ma	pa
congresswoman	congressman
sororal	fraternal
feminism	masculism
heiress	heir
countrywomen	countrymen
ma's	pa's
stepdaughter's	stepson's
girlfriend	boyfriend
congresswomen	congressmen
gal's	guy's
godmothers	godfathers
girl's	boy's
maternal	paternal
aunt's	uncle's
mother's	father's
she'd	he'd
she's	he's

Table 6: List of additional gender words.

Category	Asian-American	African-American	European-American	Hispanic & Latino
Countries	korean, indian, chinese , japanese, indonesian, pakistani, bangladeshi, filipino, filipina, veitnamese, turkish, turk, iranian, burmese, iraqi, afghan, afghani, arab, uzbek, yemeni, nepalese, sri lankan, sri-lankan, srilankan, israeli, laotian, lebenese, lebanese, palestinian, kuwaiti, mongol, armenian, thai	nigerian, ethiopian, egyptian, congolese, tanzanian, kenyan, ugandan, moroccan	german, british, french, italian, spanish, romanian, dutch, belgian, greek, irish, portugese, hungarian, austrian, swish, bulgarian, finnish, slovak, norweigian, scottish, polish, swedish, lithuanian, danish, slovenian, latvian, estonian	mexican, brazilian, salvadorian, honduran, colombian, cuban, peruvian, ecuadorian, chilean, haitian, costa rican, costa rican, tico, dominican
First Names	young, mohammed, hung, wei, hong, thanh, yong, minh, rajesh, syed, jin, jian, yan, jun, sanjay, tuan, lily, sung, ming, amit, yu, min, chi, phuong, muhammad, may, hai, anil, dung, thuy, yi, sunil, sang, teresita, jing, ravi, vijay, ying, ramesh, mei, dong, long, anh, kyung, mai, hui, jung, son, romeo, suresh, hoa, lan, cuong, ashok, jae, linh, duc, chong, tam, wai, danilo, vinh, ajay, xiao, jie, hoang, chun, wen, sun, hao, ping, rakesh, deepak, binh, khanh, sandeep, kai, anand, xin, yun, krishna, feng, eun, bo, arun, erlinda, tri, srinivas, trung, manish, lin, huong, tai, nam, hyun, ashish	willie, reginald, tyrone, cedric, lillie, sylvester, mattie, latoya, tamika, latasha, marva, keisha, althea, darnell, lula, aisha, jermaine, latonya, hattie, roosevelt, fannie, ebony, alphonso, mamie, sammie, ollie, demetrius, donnell, felecia, jarvis, cleveland, jamila, tanisha, latisha, odessa, mable, cornell, lawanda, alfreda, essie, lakisha, odell, prince, latrice, latanya, octavia, earnestine, ivory, tameka, tomeka, ayanna	michael, john, david, robert, james, william, richard, thomas, mark, mary, daniel, christopher, susan, jennifer, steven, jeffrey, brian, paul, patricia, linda, matthew, karen, scott, kevin, lisa, timothy, stephen, barbara, elizabeth, kenneth, gary, donald, ronald, jason, nancy, andrew, kathleen, eric, deborah, gregory, anthony, edward, peter, michelle, sandra, amy, kimberly, laura, george, cynthia, carol, donna, julie, patrick, douglas, christine, sharon, pamela, dennis, debra, diane, rebecca, margaret, kelly, melissa, larry, frank, ryan, sarah, angela, stephanie, jonathan, janet, cheryl, catherine, heather, judith, todd, lori, keith, jessica, bruce, craig, joshua, raymond, denise, ann, brenda, teresa, terry, katherine, alan, adam, kathryn, carolyn, nicholas, lawrence	maria, jose, juan, carlos, luis, manuel, antonio, jorge, francisco, jesus, miguel, mario, carmen, ana, rosa, roberto, ricardo, pedro, oscar, rafael, hector, raul, yolanda, javier, ramon, fernando, ruben, sergio, eduardo, angel, edgar, alejandro, armando, salvador, julio, arturo, alfredo, cesar, marco, alberto, guadalupe, enrique, alma, gerardo, irma, margarita, leticia, ernesto, silvia, guillermo, luz, rodolfo, felix, adriana, blanca, alfonso, gustavo, andres, omar, angelica, bertha, pablo, isabel, felipe, raquel, lorena, lourdes, juana, hilda, hugo, rogelio, ramiro, ignacio, rolando, abel, marcos, humberto, rosario, tomas, orlando, ismael, delia, gilberto, gabriela, elsa, susana, saul, josefina, israel, mercedes, lorenzo, alvaro, beatriz, reynaldo, rodrigo, maribel, leonardo, graciela, santiago, rigoberto
Last Names	xiong, zhang, huang, truong, yang, li, vang, huynh, vu, nguyen, ali, khan, wong, singh, chang, chung, ahmed	washington, jefferson, booker, banks, joseph, mosley, jackson, charles, dorsey, rivers	yoder, friednam, krueger, schwartz, schmitt, mueller, weiss, novak, o'connell, klein	barajas, zavala, velazquez, avalos, orozco, vazquez, juarez, meza, huerta, ibarra
Race	asian	european	african	latin, hispanic
Color		white	black	

Table 7: Word lists for generating race counterfactuals.

Model			ppl (↓)	Regard Ratio				Variance (↓)	Fluency (↓)
Method	Mod fn.	Aug.		African	Asian	European	Hispanic		
GPT2–small (Teacher)	N/A	N/A	25.17	1.280 (0.35, 0.27)	1.868 (0.40. 0.21)	1.445 (0.36, 0.25)	1.196 (0.34, 0.29)	0.302	64.69
DistilGPT–2 (HF)	N/A	N/A	39.25	1.434 (0.32, 0.22)	2.035 (0.35, 0.17)	1.599 (0.34, 0.21)	1.312 (0.32, 0.25)	0.318	155.77
DistilGPT–2 (Baseline)	N/A	N/A	40.88	1.219 (0.33, 0.27)	1.653 (0.37, 0.22)	1.364 (0.35, 0.25)	1.049 (0.31, 0.29)	0.258	94.11
DistilGPT–2 (ERA)	`max`	no	40.92	1.124 (0.30, 0.27)	1.515 (0.33, 0.22)	1.213 (0.31, 0.26)	0.938 (0.29, 0.31)	0.241	143.45
DistilGPT–2 (ERA)	none	yes	40.91	1.079 (0.29, 0.27)	1.493 (0.33, 0.22)	1.206 (0.31, 0.25)	0.955 (0.29, 0.30)	0.231	109.98
DistilGPT–2 (ERA)	`max`	no	41.46	1.056 (0.29, 0.28)	1.404 (0.32, 0.23)	1.145 (0.30, 0.26)	0.870 (0.27, 0.31)	0.222	94.78

Table 8: Racial disparity in open-ended text generation as assessed by BOLD Race prompts. We report the average of over five evaluation runs. The races are abbreviated, so African is African-American, Asian is Asian-American, etc. Fluency is the macro average across all 4 races. Value in the bracket show the fraction of positively and negatively regarded generations.

Test	X	Y	A	B
Test 6	**male:** John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill	**female:** Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna	**career:** executive, management, professional, corporation, salary, office, business, career	**family:** home, parents, children, family, cousins, marriage, wedding, relatives
Test 7	**math:** math, algebra, geometry, calculus, equations, computation, numbers, addition	**arts:** poetry, art, dance, literature, novel, symphony, drama, sculpture	**male:** male, man, boy, brother, he, him, his, son	**female:** female, woman, girl, sister, she, her, hers, daughter
Test 8	**science:** science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy	**arts:** poetry, art, Shakespeare, dance, literature, novel, symphony, drama	**male:** brother, father, uncle, grandfather, son, he, his, him	**female:** sister, mother, aunt, grandmother, daughter, she, hers, her

Table 9: Words sets and categories used in CEAT tests.