# Toward Fairness in Text Generation via Mutual Information Minimization based on Importance Sampling

**Rui Wang**[*]
Duke University

**Pengyu Cheng**[*]
Tencent AI Lab

**Ricardo Henao**[†]
Duke University & KAUST

rui.wang16@duke.edu

## Abstract

Pretrained language models (PLMs), such as GPT-2, have achieved remarkable empirical performance in text generation tasks. However, pretrained on large-scale natural language corpora, the generated text from PLMs may exhibit *social bias* against disadvantaged demographic groups. To improve the fairness of PLMs in text generation, we propose to minimize the mutual information between the semantics in the generated text sentences and their *demographic polarity*, *i.e.*, the demographic group to which the sentence is referring. In this way, the mentioning of a demographic group (*e.g.*, *male or female*) is encouraged to be independent from how it is described in the generated text, thus effectively alleviating the social bias. Moreover, we propose to efficiently estimate the upper bound of the above mutual information via importance sampling, leveraging a natural language corpus. We also propose a distillation mechanism that preserves the language modeling ability of the PLMs after debiasing. Empirical results on real-world benchmarks demonstrate that the proposed method yields superior performance in term of both fairness and language modeling ability.

## 1 INTRODUCTION

The recent advent of Pretrained Language Models (PLMs), *e.g.*, GPT-2 (Radford et al., 2019), has tremendously advanced the state-of-the-art for natural language generation tasks (Adiwardana et al., 2020; Yao et al., 2019; Dong et al., 2021). These advances have resulted in human-like cohesive text generation when prompted with a sequence of context words. However, existing PLMs are generally pretrained with large-scale natural language corpora crawled from the internet (Schick et al., 2021), without attentive filtering or scrutiny about the potential undesirable *social bias* exhibited in human language, *i.e.* prejudices or stereotypes against disadvantaged demographic groups in terms of, *e.g.*, genders or religions. Consequently, such a bias may be inherited or worse, exacerbated by the resulting PLMs during pretraining (Barikeri et al., 2021; Basta et al., 2019; Zhao et al., 2019). For example, recent studies (Dhamala et al., 2021; Lauscher et al., 2021) have shown that the pretrained GPT-2 model can spuriously correlate a demographic group, *e.g.*, "*male*" or "*female*", with certain occupations. For instance, the model can be prone to generate "*doctor*" given "*He works as*", while tending to generate "*nurse*" given "*She works as*". Such stereotypical inclinations in the text generation may cause a negative ethical impact in socio-technical scenarios (Lauscher et al., 2021), which severely curtails the applicability of PLMs to real-world text generation.

Several methods have been proposed to alleviate social bias for fair text generation. In one direction, these works debias the PLMs from the data perspective. Zhao et al. (2018) proposed the Conterfactual Data Augmentation (CDA), which creates counterfactual text instances that counter the social bias in the training corpus, *e.g.*, for gender bias, by replacing "*She*" with "*He*", and *vice versa*. The PLMs are finetuned on both the original corpus and the counterfactual instances. However, Dinan et al. (2019) showed that these artificial instances may not be grammatically correct. Instead, they manually collect data that is deemed to be free from social bias. Unfortunately, such a process is generally laborious and expensive. From a different perspective, there are also works focusing on the geometry of the pretrained embedding space of PLMs (Bolukbasi et al., 2016; Liang et al., 2021). Specifically, they project the (contextualized) token embeddings into the orthogonal space of a demographic biasing subspace, which is linearly spanned by a number of bias directions in the embedding space. The projection is generally implemented via minimizing the cosine similarity between the embedding vectors and the bias

---

| Types | Sentences |
|---|---|
| Polarity: *Male* (Polarized) | **He** can make a horse, but **he** can't make a pony. **He** will always try to buy things. **He** writes at a snail's pace, writes in a high-pitched tone and sometimes turns the page. a **man** who is willing to sacrifice **himself** for **his** beloved, for **his** neighbor and for **his father** |
| Polarity: *Female* (Polarized) | **she** is ready for **her** nursing home needs. The anesthesiologist assistant works under the medical ills of **her** job. The first psychiatric nurses faced difficult working ills during **her** first year at a clinic. |
| *Neutral* (Non-Polarized) | the clinician may use **his** or **her** abilities and abilities as an expert in the specific treatment. Anesthesiologist assistants work in all facets of ills such as headache Chaplains also build relationships with students to help them learn who they're talking to. |

Table 1: Examples of sentences generated from the GPT-2 grouped according to whether the sentence is neutral or polarized in terms of gender, *i.e.*, *male* or *female*. The demographic-sensitive words are marked with bold font. The "Polarity" is short for demographic polarity.

directions (Liang et al., 2021). However, these approaches make a strong assumption about the linearity of the bias in embedding space. Moreover, there is no guarantee that the cosine similarity can sufficiently capture the bias degree of the learnt embeddings against different demographic groups. In Kurita et al. (2019), it is shown that cosine similarity based methods do not produce consistent results in measuring bias for token embeddings. Recently, Henlein and Mehler (2022) showed that the information captured by different handcrafted similarity metrics on pretrained token embeddings, *e.g.*, cosine or Euclidean similarities, widely varies and may not be fully explainable.

Different from previous works, our proposed method does not rely on manual data collection nor handcrafted similarity metrics in the pretrained embedding space. Alternatively, we propose to debias the PLMs for text generation via minimizing the mutual information between the *demographic polarity* of the generated sentence and its semantics. The term of *demographic polarity* follows Dhamala et al. (2021), denoting which demographic group the sentence is referring to. For instance, it indicates whether a sentence is referring to a "*male*" or a "*female*" in a gender debiasing scenario. By minimizing its mutual information with the sentence semantics, we break the stereotypes regarding demographic groups in text generation such that the mentioning of each demographic group (*e.g.*, "*She*" or "*He*") is decoupled or independent from how it is described in the generated context (*e.g.*, "*Doctor*" or "*Nurse*"). We also propose a distillation mechanism that preserves the language modeling ability of PLMs. Experiments on real-world benchmarks show that our method outperforms other baselines in terms of both fairness for different demographic groups and language modeling ability.

## 2 BACKGROUND

### 2.1 Demographic Polarity

Let $\mathcal{G} = \{G_i\}_{i=1}^{K}$ be the collection of $K$ demographic groups being considered for fairness in a text generation task. As an example, we can have $\mathcal{G} = \{male, female\}$ when debiasing for fairness over genders. Each of such groups, $G_i$, is associated with a set of demographic-sensitive words, $\mathcal{V}_{G_i}$, indicating how $G_i$ is manifested in the text, *e.g.*, we can have $\mathcal{V}_{\text{male}} = \{he, himself, father, \dots, etc.\}$ and $\mathcal{V}_{\text{female}} = \{she, herself, mother, \dots, etc.\}$. The *demographic polarity* of a sentence, taking values from $\mathcal{G}$, is a term that describes which demographic groups the sentence is referring to. Demographic polarity can be inferred from the mentions of different demographic groups in a sentence. Here, we identify the mention of a demographic group $G_i$ as an occurrence of any of its corresponding demographic-sensitive words, $\mathcal{V}_{G_i}$, in the sentence. For instance, Dhamala et al. (2021) quantitatively defines the demographic polarity for gender, *i.e.*, gender polarity, as the gender (*male* or *female*) with the highest number of mentions (highest frequency) in a sentence. Similar to Dhamala et al. (2021), we define the demographic polarity of a sentence as the demographic group from $\mathcal{G}$ with the highest number of mentions. Moreover, a sentence is denoted as *polarized* if there exists a demographic group that is mentioned with higher frequency (higher number of mentions) than the others. Alternatively, we also denote a sentence as *neutral* or without demographic polarity, if there is no mentions of any considered demographic group or different demographic groups have the same number of mentions. In Table 1, we show examples of sentences generated from GPT-2 that are neutral or polarized in terms of gender, *i.e.*, with $\mathcal{G} = \{male, female\}$.

### 2.2 Language Modeling

In this paper, we term a PLM without debiasing for fair text generation, *e.g.*, the pretrained GPT-2, as a *reference PLM*. Let $\mathcal{V}$ be the vocabulary of discrete text tokens. We define $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T)$ as a sentence of length $T$, where $\boldsymbol{x}_t \in \mathcal{V}$ and $\boldsymbol{X}_{<t} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{t-1})$ is the prefix of sentence $\boldsymbol{X}$ with length $t-1$, *i.e.*, a partial sentence. Conditioned on $\boldsymbol{X}_{<t}$, the reference model is expected to predict the probability of occurrence for the next token $\boldsymbol{x}_t \in \mathcal{V}$ at position $t$, which is formulated as

Rui Wang[*], Pengyu Cheng[*], Ricardo Henao[†]

$$P^{\text{Ref}}(\boldsymbol{x}_t|\boldsymbol{X}_{<t};\mathcal{V}) = \frac{exp(\,f(\boldsymbol{X}_{<t})^{\mathsf{T}}e(\boldsymbol{x}_t)\,)}{\sum_{\boldsymbol{x}\in\mathcal{V}} exp(\,f(\boldsymbol{X}_{<t})^{\mathsf{T}}e(\boldsymbol{x})\,)}, \quad (1)$$

where $e(\boldsymbol{x})$ is the embedding vector for token $\boldsymbol{x}$, and $f(\cdot)$ is the context encoder, both of which have been pretrained on a large-scale text corpus. For debiasing, we adopt a *post hoc* approach (Cheng et al., 2021; Liang et al., 2021). Specifically, instead of retraining the whole reference model, we stack a trainable debiasing layer $D(\cdot)$ on top of $f(\cdot)$. Formally, we write the probability of predicting the next token $\boldsymbol{x}_t$ from the debiased PLM as,

$$P^{\text{Deb}}(\boldsymbol{x}_t|\boldsymbol{X}_{<t};\mathcal{V}) = \frac{exp(\,D\circ f(\boldsymbol{X}_{<t})^{\mathsf{T}}e(\boldsymbol{x}_t)\,)}{\sum_{\boldsymbol{x}\in\mathcal{V}} exp(\,D\circ f(\boldsymbol{X}_{<t})^{\mathsf{T}}e(\boldsymbol{x})\,)}, \quad (2)$$

where $\circ$ denotes the function composition. $D$ is implemented as a residual module (He et al., 2016), and is trained for debiasing with a text corpus of natural language.

Let $S$ be the output distribution over sentences generated from the debiased PLM with $D\circ f(\cdot)$. The likelihood of sentence $\boldsymbol{X}$ with distribution $S$ is

$$P_S(\boldsymbol{X}) = \sigma(\boldsymbol{x}_1)\prod_{t=2}^{T} P^{\text{Deb}}(\boldsymbol{x}_t|\boldsymbol{X}_{<t}), \quad (3)$$

where $\sigma(\boldsymbol{x}_1)$ is a prior distribution over the start token $\boldsymbol{x}_1$ of any sentence $\boldsymbol{X}$ in natural language. Hereafter, we use a subscript to denote the distribution from which $\boldsymbol{X}$ is drawn. Let $\boldsymbol{X}$ be a polarized sentence, we define $\hat{S}$ as the distribution over polarized sentences from the debiased reference model, *i.e.*, $P_{\hat{S}}(\boldsymbol{X}) = P_S(\boldsymbol{X})/P_S^{\text{polar}}$, where $P_S^{\text{polar}}$ is the probability of generating a polarized sentence from $S$.

# 3 DEBIASING PLMs FOR FAIR TEXT GENERATION

As mentioned above, PLMs pretrained with natural language have been shown to manifest social bias in text generation, *i.e.*, by undesirably associating a demographic group with semantics that reflect social stereotypes and prejudices. Such an association can be understood as a having a skewed likelihood in mentioning different demographic groups (*e.g.*, *male or female*), conditioned on the existence of certain semantics from the context (*e.g.*, *doctor or nurse*), as described in Section 1. In addressing this issue, we propose to minimize mutual information between the demographic polarity of the polarized sentences generated from PLMs and their sentence semantics. For instance, by minimizing such mutual information for gender debiasing, *i.e.*, $\mathcal{G} = \{male, female\}$, we can encourage that the generation of occupational semantics (*e.g.*, "*doctor*" or "*nurse*") to be independent of whether the sentence is describing a "*male*" or "*female*", thus effectively alleviating the social bias.

Below, in Section 3.1 we explain the proposed debiasing method for fair text generation, and in Section 3.3 we introduce a distillation mechanism that preserves the language modeling performance of the debiased PLMs.

## 3.1 The Mutual Information Between Demographic Polarity and Sentence Semantics

For $\boldsymbol{X} \sim \hat{S}$, the demographic polarity of sentence $\boldsymbol{X}$ is denoted as $J(\boldsymbol{X})$. We represent the sentence semantics of $\boldsymbol{X}$ as the hidden states from context encoder $D\circ f$, which we denote as $H(\boldsymbol{X})$. Specifically, we have $H(\boldsymbol{X}) = \{h_t(\boldsymbol{X})\}_{t=1}^{T}$, where $h_t(\boldsymbol{X}) = D\circ f(\boldsymbol{X}_{<t})$. As above, we want to minimize the mutual information between the demographic polarity $J(\boldsymbol{X})$ for a polarized sentence $\boldsymbol{X} \sim \hat{S}$, and its sentence semantics $H(\boldsymbol{X})$. Their mutual information is formally defined as

$$\mathcal{I}(J(\boldsymbol{X}), H(\boldsymbol{X})|\boldsymbol{X}\sim\hat{S}) = \quad (4)$$
$$\mathbb{E}_{\boldsymbol{X}\sim\hat{S}}\log\frac{P_{\hat{S}}(J(\boldsymbol{X}), H(\boldsymbol{X}))}{P_{\hat{S}}(J(\boldsymbol{X}))P_{\hat{S}}(H(\boldsymbol{X}))},$$

where $P_{\hat{S}}(J(\boldsymbol{X}), H(\boldsymbol{X}))$ is the joint distribution over $J(\boldsymbol{X})$ and $H(\boldsymbol{X})$, while $P_{\hat{S}}(J(\boldsymbol{X}))$ and $P_{\hat{S}}(H(\boldsymbol{X}))$ are their marginal distributions. In the following, we denote $\mathcal{I}(J(\boldsymbol{X}), H(\boldsymbol{X})|\boldsymbol{X}\sim\hat{S})$ as $\mathcal{I}(J, H)$ for conciseness.

The difficulty in optimizing with $\mathcal{I}(J, H)$ lies in how to efficiently estimate such mutual information from a text generator. For minimization, we first decompose $\mathcal{I}(J, H)$ over $H$ for each position $t$ as

$$\frac{1}{T}\mathcal{I}(J, H) \leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{I}(J, h_t) \quad (5)$$

where $\mathcal{I}(J, h_t)$ is the mutual information between $J(\boldsymbol{X})$ and the position-wise hidden state $h_t(\boldsymbol{X})$ at position $t$, *i.e.*, replacing $H(\boldsymbol{X})$ with $h_t(\boldsymbol{X})$ in (4). $\mathcal{I}(J, h_t)$ can be approximated by an existing upper bound estimator of mutual information (Cheng et al., 2020), defined as $\mathcal{I}(J, h_t) \leq \mathcal{I}^u(J, h_t)$, with

$$\mathcal{I}^u(J, h_t) = \mathbb{E}_{\boldsymbol{X}\sim\hat{S}}\big(\log P_{\hat{S}}(J(\boldsymbol{X})|h_t(\boldsymbol{X}))$$
$$- \mathbb{E}_{\boldsymbol{X}'\sim\hat{S}}\log P_{\hat{S}}(J(\boldsymbol{X})|h_t(\boldsymbol{X}'))\big). \quad (6)$$

$\mathcal{I}^u(J, h_t)$ is an upper bound that measures the difference between the log likelihood, $\log P_{\hat{S}}(\cdot|\cdot)$, of the positive pair $\{J(\boldsymbol{X}), h_t(\boldsymbol{X})\}$ and the negative pair $\{J(\boldsymbol{X}), h_t(\boldsymbol{X}')\}$, with $\boldsymbol{X}$ and $\boldsymbol{X}'$ being independent. This can be understood as the expected uncertainty of $J(\boldsymbol{X})/h_t(\boldsymbol{X})$ given $h_t(\boldsymbol{X})/J(\boldsymbol{X})$. Note that we do not follow the Donsker-Varadhan representation (Belghazi et al., 2018) since it induces a low bound of mutual information, thus is inappropriate for minimization problems. From Cheng et al. (2020),

(6) can be estimated by sampling $\boldsymbol{X}, \boldsymbol{X}' \sim \hat{S}$,

$$\mathcal{I}^u(J, h_t) \approx \frac{1}{N} \sum_{i=1}^{N} \big( q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}_i)) \tag{7}$$
$$- \frac{1}{N} \sum_{j=1}^{N} q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}'_j)) \big),$$

where $q(\cdot|\cdot)$ is an approximation to the (unknown) log likelihood $\log P(\cdot|\cdot)$ in (6). $q(\cdot|\cdot)$ is trained using samples from the the positive pairs $\{J(\boldsymbol{X}_i), h_t(\boldsymbol{X}_i)\}_{i=1}^{N}$ in each iteration, as illustrated in Algorithm 1. The sets of samples for $\boldsymbol{X}$ and $\boldsymbol{X}'$ are denoted as $\{\boldsymbol{X}_i\}_{i=1}^{N}$ and $\{\boldsymbol{X}'_j\}_{j=1}^{N}$, respectively, which are all sampled from $\hat{S}$. For computational efficiency, we let $\boldsymbol{X}$ and $\boldsymbol{X}'$ share the same set of samples, *i.e.*, $\{\boldsymbol{X}_i\}_{i=1}^{N} = \{\boldsymbol{X}'_j\}_{j=1}^{N}$.

### 3.2 Estimating with Natural Language Sentences

The estimation of the mutual information with (7) requires sampling polarized sentences from $\hat{S}$, a common approach for which is to sample from $S$ with the debiased PLM, then only keep those that are polarized (Dhamala et al., 2021; Martino and Míguez, 2010). However, such a sampling strategy for estimating (6) can be cumbersome for mainly two reasons. First, $\hat{S}$ is constantly changing with the debiased PLM during training, which implies that the samples in (7) have to be regenerated every several iterations. This reduces the training efficiency by further considering that the generation of words within a sentence, *e.g.*, with GPT-2, needs to be done sequentially, thus cannot be parallelized. Moreover, polarized sentences may only amount to a small portion of the PLM outputs. For instance, Dhamala et al. (2021) estimates that only $6.75\%$ sentences generated from GPT-2 are polarized in term of gender given occupational prompts.[1] Thus, directly generating $\boldsymbol{X} \sim \hat{S}$ from the PLMs for estimating (6) can be extremely inefficient.

In our approach, we estimate (6) via importance sampling, leveraging a text corpus of natural language. We describe the natural language corpus used for experiments in Section 5. Here, we denote $A$ as the distribution over text sentences from natural language, and $\hat{A}$ as the conditional distribution of $A$ over polarized sentences. Let $\boldsymbol{X}$ be a polarized sentence, we have $P_{\hat{A}}(\boldsymbol{X}) = P_A(\boldsymbol{X})/P_A^{\text{polar}}$, where $P_A^{\text{polar}}$ is the probability of a natural language sentence being polarized. To circumvent the inefficiency in sampling from $\hat{S}$ with the PLMs outputs (as required in (7)), we rewrite the upper bound $\mathcal{I}^u(J, h_t)$ in (6) in terms of expectation over $\hat{A}$, so that we can estimate (6) via drawing samples from $\hat{A}$

---

[1]Also from Liang et al. (2021), across 5 natural language corpus, only $1.6\%$ sentences are gender related and $0.12\%$ are religion related. Thus, we should not expect high probability in generating polarized sentences from PLMs pretrained over natural language.

instead of $\hat{S}$. Specifically, $\mathcal{I}^u(J, h_t)$ can be rewritten as,

$$\mathcal{I}^u(J, h_t) = \mathbb{E}_{\boldsymbol{X} \sim \hat{A}} \hat{R}(\boldsymbol{X}) \big( \log P(J(\boldsymbol{X})|h_t(\boldsymbol{X})) \tag{8}$$
$$- \mathbb{E}_{\boldsymbol{X}' \sim \hat{A}} \hat{R}(\boldsymbol{X}') \log P(J(\boldsymbol{X})|h_t(\boldsymbol{X}')) \big),$$

where $\hat{R}(\cdot) = P_{\hat{S}}(\cdot)/P_{\hat{A}}(\cdot)$ denotes the likelihood ratio. Note that (8) enable us to estimate $\mathcal{I}^u(J, h_t)$ via sampling from $\hat{A}$, which facilities the training efficiency from two perspectives: *i)* unlike $\hat{S}$, the distribution $\hat{A}$ can be sampled efficiently by conditioning, *i.e.*, via filtering out non-polarized sentences from the natural language corpus; and *ii)* the sampled sentences can be reused for estimating (6) during different steps of training. Consequently, following (7), we estimate (8) with sampled sentences from $\hat{A}$ using,

$$\mathcal{I}^u(J, h_t) \approx \frac{1}{N} \sum_{i=1}^{N} \hat{R}(\boldsymbol{X}_i) \big( q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}_i)) \tag{9}$$
$$- \frac{1}{N} \sum_{j=1}^{N} \hat{R}(\boldsymbol{X}_j) q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}'_j)) \big).$$

Similar to (7), $\boldsymbol{X}$ and $\boldsymbol{X}'$ in (8) share the same set of samples, *i.e.*, $\{\boldsymbol{X}_i\}_{i=1}^{N} = \{\boldsymbol{X}'_j\}_{j=1}^{N}$ in (9). Note that $\{\boldsymbol{X}_i\}_{i=1}^{N}$ and $\{\boldsymbol{X}'_j\}_{j=1}^{N}$ in (9) are sampled from $\hat{A}$ instead of $\hat{S}$.

The remaining question is how to efficiently calculate the likelihood ratio terms in (9). We address this by leveraging the pretrained knowledge from PLMs. Specifically, we propose to quantify the probability distribution of $\hat{A}$ over natural language sentences using the reference PLM, since it is pretrained with large-scale natural language corpus. Recall in (1), $P^{\text{Ref}}(\boldsymbol{x}_t|\boldsymbol{X}_{<t}; \mathcal{V})$ is the reference probability of generating the next token $\boldsymbol{x}_t$ given its context $\boldsymbol{X}_{<t}$. Following (3), we can compute $P_A(\boldsymbol{X})$ as,

$$P_A(\boldsymbol{X}) = \sigma(\boldsymbol{x}_1) \prod_{t=2}^{T} P^{\text{Ref}}(\boldsymbol{x}^t|\boldsymbol{X}_{<t}; \mathcal{V}), \tag{10}$$

In the experiments, we use the pretrained GPT-2 model (Radford et al., 2019) as our reference PLM. With (10), $\hat{R}(\boldsymbol{X})$ is calculated as,

$$\hat{R}(\boldsymbol{X}) = \frac{P_{\hat{S}}(\boldsymbol{X})}{P_{\hat{A}}(\boldsymbol{X})} = \underbrace{\frac{P_S(\boldsymbol{X})}{P_A(\boldsymbol{X})}}_{l(\boldsymbol{X})} \cdot \underbrace{\frac{P_A^{\text{polar}}}{P_S^{\text{polar}}}}_{R^{\text{polar}}} \tag{11}$$

$$= \frac{\prod_{t>1} P^{\text{Deb}}(\boldsymbol{x}^t|\boldsymbol{X}_{<t}; \mathcal{V}))}{\prod_{t>1} P^{\text{Ref}}(\boldsymbol{x}^t|\boldsymbol{X}_{<t}; \mathcal{V}))} \cdot R^{\text{polar}} \tag{12}$$

where $P_S(\boldsymbol{X})$ and $P_A(\boldsymbol{X})$ are defined in (3) and (10), respectively. From (11) to (12), $\sigma(\boldsymbol{x}_1)$ in $P_S(\boldsymbol{X})$ and $P_A(\boldsymbol{X})$ cancels out in the first term $l(\boldsymbol{X})$. Thus, $l(\boldsymbol{X})$ can be directly calculated with the reference and debiased PLMs using the first term in (12). However, we note that the value of $R^{\text{polar}}$ in (12) is not readily available. Further, since

$P_S^{\text{polar}}$ changes along with training of the debiased PLM, the value of $R^{\text{polar}}$ is not a constant during training, thus cannot be ignored in minimization. To solve this, we approximate $R^{polar}$ with the sampled sentences from $\hat{A}$ by noting that,

$$\frac{1}{R^{\text{polar}}} = \frac{P_S^{\text{polar}}}{P_A^{\text{polar}}} = \sum_{\boldsymbol{X} \text{ is Polarized}} \frac{P_S(\boldsymbol{X})}{P_A(\boldsymbol{X})} \frac{P_A(\boldsymbol{X})}{P_A^{\text{polar}}}$$

$$= \sum_{\boldsymbol{X} \text{ is Polarized}} \frac{P_S(\boldsymbol{X})}{P_A(\boldsymbol{X})} P_{\hat{A}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{X} \sim \hat{A}} \frac{P_S(\boldsymbol{X})}{P_A(\boldsymbol{X})}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \frac{P_S(\boldsymbol{X}_i)}{P_A(\boldsymbol{X}_i)} = \frac{1}{N} \sum_{i=1}^{N} l(\boldsymbol{X}_i), \quad (13)$$

where $\{\boldsymbol{X}_i\}_{i=1}^{N}$ are samples from $\hat{A}$ as in (9). With (13) replaced into (12), we can estimate the likelihood ratios $R(\boldsymbol{X}_i)$ for the set $\{\boldsymbol{X}_i\}_{i=1}^{N}$ as,

$$\frac{1}{N} \hat{R}(\boldsymbol{X}_i) \approx \frac{l(\boldsymbol{X}_i)}{\sum_{i'=1}^{N} l(\boldsymbol{X}_{i'})} := m(\boldsymbol{X}_i). \quad (14)$$

Finally, replacing (14) into (9), we obtain a mutual information estimator for (5) written as,

$$\frac{1}{T} \mathcal{I}(J, H) \leq \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}^u(J, h_t)$$

$$\approx \sum_{i=1}^{N} m(\boldsymbol{X}_i) \Big( \frac{1}{T} \sum_{t=1}^{T} q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}_i))$$

$$- \sum_{j=1}^{N} m(\boldsymbol{X}_j) \frac{1}{T} \sum_{t=1}^{T} q(J(\boldsymbol{X}_i)|h_t(\boldsymbol{X}_j')) \Big). \quad (15)$$

We further simplify (15) by noting that $\frac{1}{T} \sum_{t=1}^{T} q(\cdot|\cdot)$ is an expectation over positions $1, \ldots, T$, *i.e.*,

$$\mathbb{E}_{t \sim \mathcal{T}} q(\cdot|\cdot) = \frac{1}{T} \sum_{t=1}^{T} q(\cdot|\cdot), \quad (16)$$

where $\mathcal{T}$ is the uniform distribution over positions $1, \ldots, T$. Then, (16) can be cheaply approximated with a value of $q(\cdot|\cdot)$ from a randomly sampled position. Concretely, we can estimate (15) as,

$$\mathcal{L}_{\text{fair}} := \sum_{i=1}^{N} m(\boldsymbol{X}_i) \big( q(J(\boldsymbol{X}_i)|h(\boldsymbol{X}_i)) \quad (17)$$

$$- \sum_{j=1}^{N} m(\boldsymbol{X}_j) q(J(\boldsymbol{X}_i)|h(\boldsymbol{X}_j')) \big),$$

where $h(\boldsymbol{X}_i) \in \{h_t(\boldsymbol{X}_i)\}_{t=1}^{T}$ is a hidden state sampled uniformly from a random position, and similarly for $h(\boldsymbol{X}_j')$. $\{\boldsymbol{X}_i\}_{i=1}^{N}$ are sampled from $\hat{A}$. The value of (17) is defined as $\mathcal{L}_{\text{fair}}$, which is our proposed loss for fair text generation.

## 3.3 Preserving Language Modeling Ability for the Debiased PLM

Simply training with the loss for debiasing in (17) can result in non-fluent text generations, due to catastrophic forgetting on the language modeling ability of the reference PLMs (Cheng et al., 2021; Gupta et al., 2022). For instance, given $\boldsymbol{X}_{i<t} =$ *"She is known for"* in the context of gender debiasing, the next word in natural language is more likely to be *her* than *his*, due to the linguistic consistency between different mentions of demographic groups within a sentence. However, the model that is solely trained with loss for fairness may generate *"She is known for his way of being a doctor"*, since the model may be prone to forget knowledge of such linguistic consistency previous learned by the reference PLM, so the encoded $h_t(\boldsymbol{X}) = D \circ f(\boldsymbol{X}_{<t})$ can produce more balanced (fair) occurrences of *male vs. female* in the generated sentences. In the Supplementary Material, we show samples generated by the PLMs. We find that the model simply trained with the fairness loss is prone to generate inconsistent mentions of demographic groups, which is manifested by higher perplexity in terms of language modeling (see Section 5).

To address this problem, we propose to distill from the reference PLM on consistency between mentions of different demographic groups, so that the model after debiasing can also generate consistent mentions of demograhic groups as in the reference model. Specifically, given a partial sentence $\boldsymbol{X}_{<t}$ with at least one mention of demographic groups, *e.g.*, *"She works as"*, we train the debiased PLM via distilling the probability of predicting over only the demographic-sensitive words, $\mathcal{V}_\mathcal{G} = \{x | x \in \mathcal{V}_{G_i}, G_i \in \mathcal{G}\}$, from the reference PLM. Please refer to Section 5 for the construction of demographic-sensitive words. Below, we use $\boldsymbol{x}$ to denote a word and and let $e(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{V}_\mathcal{G}$ be the average embedding of tokens in $\boldsymbol{x}$ after tokenization. Following (1) and (2), we have $P^{\text{Ref}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V}_\mathcal{G})$ and $P^{\text{Deb}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V}_\mathcal{G})$ denoting the next token/word prediction over $\mathcal{V}_\mathcal{G}$, for the reference and the debiased PLM, respectively. For $\boldsymbol{X}_{<t}$ with at least one mention of demographic groups, we additionally train with the following loss,

$$\mathcal{L}_{\mathcal{V}_\mathcal{G}} = \mathbb{KL}(P^{\text{Ref}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V}_\mathcal{G})|P^{\text{Deb}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V}_\mathcal{G})) \quad (18)$$

where $\mathbb{KL}(\cdot|\cdot)$ is the KL-divergence. Note that we only distill with $\boldsymbol{x} \in \mathcal{V}_\mathbb{G}$. In the experiments, we also try distilling with (18) using the full vocabulary $\mathcal{V}$, denoted as

$$\mathcal{L}_{\mathcal{V}} = \mathbb{KL}(P^{\text{Ref}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V})|P^{\text{Deb}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V})) \quad (19)$$

which produces worse results than $\mathcal{L}_{\mathcal{V}_\mathbb{G}}$. This is because $P^{\text{Ref}}(\cdot|\boldsymbol{X}_{<t}; \mathcal{V})$ can represent biases in the original PLM, *e.g.*, $P^{\text{Ref}}(nurse|\boldsymbol{X}_{<t}; \mathcal{V}) > P^{\text{Ref}}(doctor|\boldsymbol{X}_{<t}; \mathcal{V})$, given that $\boldsymbol{X}_{<t}$ only contains occurrences of *female*. Such bias can propagate to the debiased reference model with (18).

To further preserve the general language modeling ability during debiasing, we also train with language modeling on

the natural language corpus. Specifically, for a sentence $\boldsymbol{X}$ from natural language, the loss for language modeling is,

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T} \sum_{t=1}^{T} \log \frac{D \circ f(\boldsymbol{X}_{<t})^{\mathsf{T}} e(\boldsymbol{x})}{\sum_{\boldsymbol{x}' \in \mathcal{V}} D \circ f(\boldsymbol{X}_{<t})^{\mathsf{T}} e(\boldsymbol{x}')}, \quad (20)$$

for which we only select sentences $\boldsymbol{X}$ that are neutral to avoid training with the bias exhibited by polarized sentences in the natural language corpus.

### 3.4 Overall Objective

The overall objective in training the debiased PLM is,

$$\mathcal{L} = \mathcal{L}_{\text{fair}} + \alpha_1 \mathcal{L}_{\text{LM}} + \alpha_2 \mathcal{L}_{\mathcal{V}_{\mathbb{G}}}. \quad (21)$$

where $\alpha_1$ and $\alpha_2$ are two balancing parameters. For reference, $\mathcal{L}_{fair}$, $\mathcal{L}_{LM}$ and $\mathcal{L}_{\mathcal{V}_{\mathbb{G}}}$ are defined in (17), (20) and (18), respectively. Algorithm 1 shows the procedure for optimizing with $\mathcal{L}$.

## 4 RELATED WORK

For the task of debiasing PLMs for text generation, a portion of previous works focus on training or finetuning the PLMs (reference models) on an unbiased dataset, which is commonly done via Conterfactual Data Augmentation (CDA) (Zhao et al., 2018). Dinan et al. (2019) finds data generated with CDA may be grammatically incorrect, thus propose to manually collect an unbiased dataset. However, such a process is laborious. In addition to CDA, Gupta et al. (2022) propose to augment on the output from the reference model. Specifically, they modify its output logits via neutralizing (averaging or max pooling) across different demographic groups. The modified logits are used as teaching signal for the debiased reference model. Alternatively, there is another line of works in debiasing via modifying or regularizing on the (contextualized) embedding space of PLMs (Bolukbasi et al., 2016; Liang et al., 2021). As described in Section 1, such approaches generally assume strong linearity in describing the bias in word embeddings Cheng et al. (2021), and the regularization is mostly based on minimizing the cosine similarity between word embeddings and biased directions. Kurita et al. (2019) shows that the cosine similarity does not produce consistent results in measuring bias. There are also works exploring architectures or parameterizations that efficiently adapt PLMs for fair generation. Sheng et al. (2020) append trainable triggers (prompts) to the input of PLMs, so that the model can generate fair outputs with a curated set of text templates. However, Gupta et al. (2022) shows the prompts in Sheng et al. (2020) may not generalize well with unseen inputs. Dathathri et al. (2019); Lauscher et al. (2021) also investigate on inserting adaptors in PLMs for efficient training, which is orthogonal to our approach. Our approach fits in the approaches that regularize the contextualized embedding

space of PLMs. Specifically, we propose to minimize the mutual information between the demographic polarity of a sentence and its semantics, where the mutual information is estimated with importance sampling.

## 5 EXPERIMENTS

### 5.1 General Setup

Following Cheng et al. (2021) and Lauscher et al. (2021), we select the demographic groups regarding gender, *i.e.*, *male* or *female*, when evaluating our debiasing methods. Specifically, we select the pretrained GPT-2 model (Radford et al., 2019) as our reference model for debiasing. Since we leverage the natural language corpus for importance sampling, we expect the natural language corpus to be diverse enough so that it can better cover the desired target distribution, *i.e.*, $\hat{S}$, to be sampled. Therefore, the natural language corpus for training includes the diverse text corpora proposed in Liang et al. (2021) when training the pretrained GPT2 for debiasing. In constructing the set of demographic-sensitive words for genders, we first adopt the demographic-sensitive words for gender in Liang et al. (2021). Then, we extend the demographic-sensitive words for each gender with its top 1K names in the United States (urls shown in the supplimentary material). For evaluation, we follow Liang et al. (2021) that examine the trade-off between the fairness and language modeling ability (see Section 5.2). Additionally, we evaluate gender equality with the professional prompts from the BOLD dataset (Dhamala et al., 2021) (see Section 5.3). The debiasing layer is trained with learning rate 5e-5 and batch size 8.

**Baseline:** We denote the debiased PLM trained with (21) as *Ours*. In addition to the reference model without debiasing, *i.e.*, GPT-2 (Radford et al., 2019), we compare with the following baselines,

- *A-INLP* (Liang et al., 2021): It trains a linear classifier on the contextualized embeddings from GPT2, then projects the embeddings into the orthogonal space of the learnt weight vectors of the classifier.

- Equitable Role Alteration (*ERA*) (Gupta et al., 2022): A recently proposed distillation framework for debiasing. Specifically, it debias the PLMs via learning with the modified logits of the reference model, *i.e.*, the pretrained GPT-2, while training with the counterfactual augmented data. EDA has been shown better results than previous debiasing methods, including CDA (Zhao et al., 2018) and the trigger-based approach (Sheng et al., 2020).

- *Ours w/o Est*: The same as *Ours*, except that we do not estimate likelihood ratios $\hat{R}(\boldsymbol{X})$ and $\hat{R}(\boldsymbol{X}')$ in (8), but simply let $\hat{R}(\boldsymbol{X}) = \hat{R}(\boldsymbol{X}') = 1$. This evaluates how effective is our estimation of $\hat{R}(\cdot)$ in (14).

**Algorithm 1** Algorithm for One step of training with the overall objective in (21).

**Input:** Parameters $\alpha_1$, $\alpha_2$. A batch of polarized sentences $\boldsymbol{B}^p$, a batch of neutral sentences $\boldsymbol{B}^{\text{neu}}$ and a batch of partial sentences with at least one mention of demographic groups $\boldsymbol{B}^s$. The sets $\boldsymbol{B}^p$, $\boldsymbol{B}^{\text{neu}}$ and $\boldsymbol{B}^s$ of size $N$ are uniformly sampled from the natural language corpus described in Section 5.
\# Sample positions
**for** $i$ from $\{1, \cdots, N\}$ **do**
    Take polarized sentence $\boldsymbol{X}_i$ from $\boldsymbol{B}^p$
    Sample $h(\boldsymbol{X}_i)$ from $\{h_t(\boldsymbol{X}_i)\}_{t=1}^T$ uniformly, as in (17)
    Construct $\{J(\boldsymbol{X}_i), h(\boldsymbol{X}_i)\}$
**end for**
\# Train the log likelihood approximator $q$ in (17)
Train $q$ via maximizing $\frac{1}{N} \sum_{i=1}^{N} m(\boldsymbol{X}_i) q(J(\boldsymbol{X}_i)|h(\boldsymbol{X}_i))$
\# Train the PLM for text generation
Calculate $\mathcal{L}_{\text{fair}}$, $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ and $\mathcal{L}_{\text{LM}}$ with $\boldsymbol{B}^p$, $\boldsymbol{B}^s$ and $\boldsymbol{B}^{\text{neu}}$, respectively.
Compute $\mathcal{L} = \mathcal{L}_{\text{fair}} + \alpha_1 \mathcal{L}_{\text{LM}} + \alpha_2 \mathcal{L}_{\mathcal{V}_\mathcal{G}}$
Update the debiased reference model according to the gradient from $\mathcal{L}$.

| | H & M | | S & T | | I & M | | E & A | | Avg | | PPL↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_{ngram}$ ↑ | $F_{max}$ ↑ | $F_{ngram}$ ↑ | $F_{max}$ ↑ | $F_{ngram}$ ↑ | $F_{max}$ ↑ | $F_{ngram}$ ↑ | $F_{max}$ ↑ | $F_{ngram}$ ↑ | $F_{max}$ ↑ | |
| *GPT-2* (Radford et al., 2019) | 0.855 | 0.742 | 0.154 | 0.184 | 0.418 | 0.398 | 0.369 | 0.323 | 0.421 | 0.440 | 29.37 |
| *A-INLP* (Liang et al., 2021) | 0.687 | 0.721 | 0.262 | 0.204 | 0.543 | 0.615 | 0.578 | 0.573 | 0.516 | 0.528 | 31.96 |
| *ERA* (Gupta et al., 2022) | 0.864 | 0.802 | 0.810 | 0.704 | 0.798 | 0.770 | 0.747 | 0.796 | 0.805 | 0.768 | 32.44 |
| *Ours w/o Est* | 0.950 | 0.884 | 0.771 | 0.636 | 0.807 | 0.815 | 0.768 | 0.805 | 0.824 | 0.785 | 31.37 |
| *Ours* | 0.958 | 0.893 | 0.785 | 0.614 | 0.821 | 0.839 | 0.784 | 0.833 | 0.837 | 0.796 | 31.39 |

Table 3: Results with BOLD. The perplexity is computed with the Wiki-text test set as mentioned in Section 5.3. The arrows ↑ (↓) next to each metric denote whether higher (lower) is better. Compared with *Ours w/o Est*, *Ours* has much better fairness (higher $F_{ngram}$ and $F_{max}$) with only a slight increase in preplexity (by 0.02).

Please refer to the Supplementary Material for more details on the experiment setup and baseline implementations.

## 5.2 Analyzing $\alpha_1$ and $\alpha_2$ via the Trade-Off Between Fairness and Language Modeling

In debiasing, we expect our debiased PLM to be fair across different demographic groups in text generation, while preserving the language modeling (LM) ability of the reference PLM to avoid generating non-fluent sentences. However, the objective terms for fairness and language modeling usually result in a trade-off (Liang et al., 2021), since training the reference PLM for debiasing may lead to catastrophic forgetting of its language modeling ability learned in pretraining. Moreover, giving a single comprehensive score over fairness and language modeling is difficult, since it is unclear how much the performance of language modeling can be sacrificed for an increment amount in fairness. Therefore, instead of providing a gross score over the two terms for evaluation, Liang et al. (2021) draws the fairness-LM curve for better granularity.

In Figure 1, we show the fairness-LM curve of our method with different values of $\alpha_1$ and $\alpha_2$, following the experimental setting in Liang et al. (2021). Specifically, the KL divergence in the Y axis of Figure 1 is calculated between the two predicted

distributions, *i.e.*, $\mathbb{KL}(P^{\text{Deb}}(\cdot|\boldsymbol{X}_{<t}^1; \mathcal{V})|P^{\text{Deb}}(\cdot|\boldsymbol{X}_{<t}^2; \mathcal{V}))$, where $\{\boldsymbol{X}_{<t}^1, \boldsymbol{X}_{<t}^2\}$ is a pair of partial sentence with demographic-sensitive words with swapped, *e.g.*, $\{\boldsymbol{X}_{<t}^1 =''$ She works as'', $\boldsymbol{X}_{<t}^2 =''$ He works as''$\}$. We follow Liang et al. (2021) and use the text templates for $\{\boldsymbol{X}^1, \boldsymbol{X}^2\}$ from Sheng et al. (2019). We seek a lower value of the KL divergence for improved fairness. The language model performance is quantified in terms of perplexity (lower the better) and represented as the X axis of Figure 1, which is calculated with the same corpus[2] as in Liang et al. (2021). For the blue curve in Figure 1, we set $\alpha_2 = 0$ and vary $\alpha_1$ with values $\{0, 1, 2, 4, 6, 8\}$, to experiment with the trade-off between the general language modeling ability (with $\mathcal{L}_{\text{LM}}$) and the fairness (with $\mathcal{L}_{\text{fair}}$). We find that a lower perplexity generally induces a higher KL divergence for fairness, manifesting the trade-off between the two objectives. Ideally, the model with better fairness-LM trade-off should be closer to the left bottom corner. We empirically set $\alpha_1 = 2$ and examine the effect of $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ and $\mathcal{L}_\mathcal{V}$ in Section 3.3, in experimenting with $\alpha_2 = \{0, 2\}$. Specifically, "with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$" corresponds to (21), where we distill with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ over only the demographic-sensitive words $\mathcal{V}_\mathcal{G}$. "with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$" corresponds to replacing $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ in (21) with $\mathcal{L}_\mathcal{V}$ in (19), thus distilling over the whole vocabulary $\mathcal{V}$ instead of only using $\mathcal{V}_\mathcal{G}$. We find that $\alpha_1 = \alpha_2 = 2$ (with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$) achieves a better trade-off
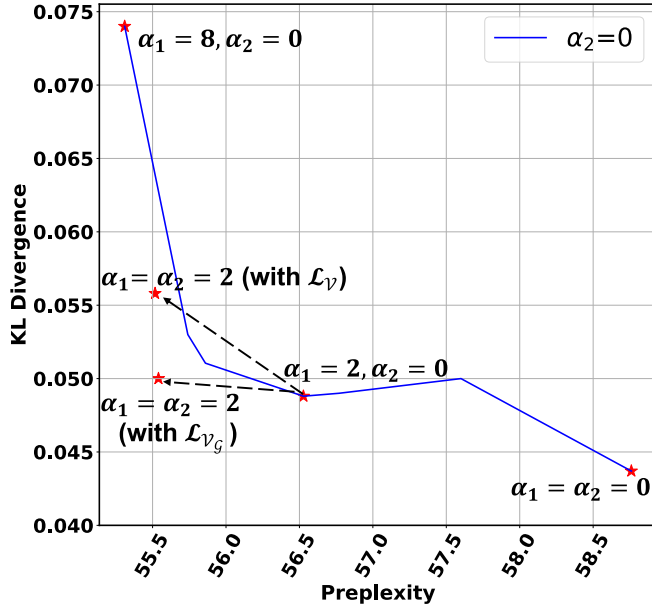
---

[2]https://github.com/pliang279/LM_bias

Figure 1: Trade-off between fairness and language modeling. "with $\mathcal{L}_{\mathcal{V}_{\mathcal{G}}}$" corresponds to the overall objective (21). "with $\mathcal{L}_{\mathcal{V}}$" replaces $\mathcal{L}_{\mathcal{V}_{\mathcal{G}}}$ in (21) with $\mathcal{L}_{\mathcal{V}}$ in (19).

| Category | Occupation |
|---|---|
| Health&Med (H&M) | healthcare<br>nursing<br>mental_health |
| Science&Tech (S&T) | computer<br>scientific<br>engineering |
| Industrial & Manufact (I&M) | metalworking<br>sewing<br>driver<br>corporate<br>industrial<br>railway |
| Entertainment& Arts (E&A) | film_TV<br>artistic<br>entertainer<br>dance<br>writing<br>theatre |

Table 2: Categories of occupations from the BOLD dataset.

compared with $\alpha_1 = 2$, $\alpha_2 = 0$, *i.e.*, being closer to the bottom left corner. This demonstrates the effectiveness of our proposed $\mathcal{L}_{\mathcal{V}_{\mathcal{G}}}$. Additionally, we find $\alpha_1 = \alpha_2 = 2$ (with $\mathcal{L}_{\mathcal{V}}$) that distills with the whole vocabulary $\mathcal{V}$ can have a slightly higher KL divergence than $\alpha_1 = \alpha_2 = 2$ (with $\mathcal{L}_{\mathcal{V}_{\mathcal{G}}}$). since the probability over the whole vocabulary $\mathcal{V}$ contains social bias encoded by the reference PLM, which is propagated to the debiased PLM during distillation, as discussed in Section 3.3. In the Supplementary Material, we count the number of inconsistent mentions of demographic groups in generated sentences from the debiased PLM, showing that training with $L_{\mathcal{V}_{\mathcal{G}}}$ can reduce the inconsistent mentions of demographic groups in text generation, reflected by lower perplexity in Figure 1. Note that $\alpha_2 = 2$ in Figure 1 is not manually selected, but by simply setting $\alpha_2 = \alpha_1$. We also keep $\alpha_1 = \alpha_2 = 2$ (with $\mathcal{L}_{\mathcal{V}_{\mathcal{G}}}$) for the experiments below with BOLD (Dhamala et al., 2021).

## 5.3 Results with the Bold Dataset

We additionally experiment with the professional prompts from the BOLD dataset (Dhamala et al., 2021), with prompts for 18 professions listed and grouped in Table 2. Given a prompt *e.g.*, *Working as an artist*, the model should have equal probability on completing the sentence with a *male* or *female* polarity. For evaluation, we collect the sentences completed by our model for each prompt. Then, we count the number of sentences with *male* polarity, $\#male$, and *female* polarity, $\#female$, according to criteria of demographic polarity in Section 2.1. Then, the fairness of the

generated text is evaluated via,

$$F_{ngram} = min \left\{ \frac{\#female}{\#male}, \frac{\#male}{\#female} \right\}. \quad (22)$$

We denote it as $ngram$, since the criteria in Section 2.1 is based on counting demographic mentions. Note that the range of $F_{ngram}$ should be $(0, 1]$. We also compute $F_{max}$, which is the same as $F_{ngram}$, except we obtain the demographic polarity of a sentence from the word that is mostly related to a demographic group, following Dhamala et al. (2021).

Table 3 shows the results of our methods and baselines. We can observe that the scores of fairness ($F_{ngram}$ and $F_{max}$) are mostly consistent with each other. Our methods outperform the baselines in terms of average on the scores of fairness. We also compute the perplexity of the debiased reference model on the wikitext-2 test set (Merity et al., 2016) to evaluate the langauge modeling ability. Among the baselines, I-ANLP (Liang et al., 2021) have comparably lower scores of fairness, probably because the linear classifier is not powerful enough in capturing the bias, showing that the biased subspace may not be model with linearity as discussed in Section 1. *Ours w/o Est* has approximately the same perplexity as *Ours*, but much lower scores for fairness, demonstrating the effectiveness of our estimation of $\hat{R}(\cdot)$ in (14). Additionally, we plot the $F_{ngram}$ scores of each occupation from BOLD in Figure 2 in the supplementary material. It shows that our method can result in better fairness for most of the occupations.

# 6 CONCLUSIONS

In this paper, we proposed to debias PLMs for text generation, via minimizing the mutual information between the demographic polarity of a generated sentence and its semantics. We proposed an approach based on importance sampling to efficiently approximate the upper bound of mutual information, which is based on the observation that polarized sentences can be generated from the PLMs with low probability. We also introduced a distillation mechanism to preserve the language modeling ability of the debiased PLM. Experiments with various benchmarks showed that our approach can efficiently debias the PLMs, while maintaining its language modeling ability.

# 7 LIMITATIONS

A limitation of our approach is that it only focuses on monolingual text generation. An interesting direction is to expand on PLMs for cross-lingual transfer regarding their ability for fair text generation. Additionally, the demographic polarity in this paper is defined via word frequency, which might not be accurate. Finally, since the bias is admittedly not completely mitigated, care should be taken when deploying the debiased PLM in diverse real-world settings.

## References

D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.

C. Basta, M. R. Costa-Jussà, and N. Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

P. Cheng, W. Hao, S. Yuan, S. Si, and L. Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413*, 2021.

S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872, 2021.

E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*, 2019.

C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang. A survey of natural language generation. *arXiv preprint arXiv:2112.11739*, 2021.

U. Gupta, J. Dhamala, V. Kumar, A. Verma, Y. Pruksachatkun, S. Krishna, R. Gupta, K.-W. Chang, G. V. Steeg, and A. Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*, 2022.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A. Henlein and A. Mehler. What do toothbrushes do in the kitchen? how transformers think our world is structured. *arXiv preprint arXiv:2204.05673*, 2022.

K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

A. Lauscher, T. Lüken, and G. Glavaš. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*, 2021.

P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

L. Martino and J. Míguez. Generalized rejection sampling schemes and applications in signal processing. *Signal Processing*, 90(11):2981–2995, 2010.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

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

T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*, 2020.

L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385, 2019.

J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.
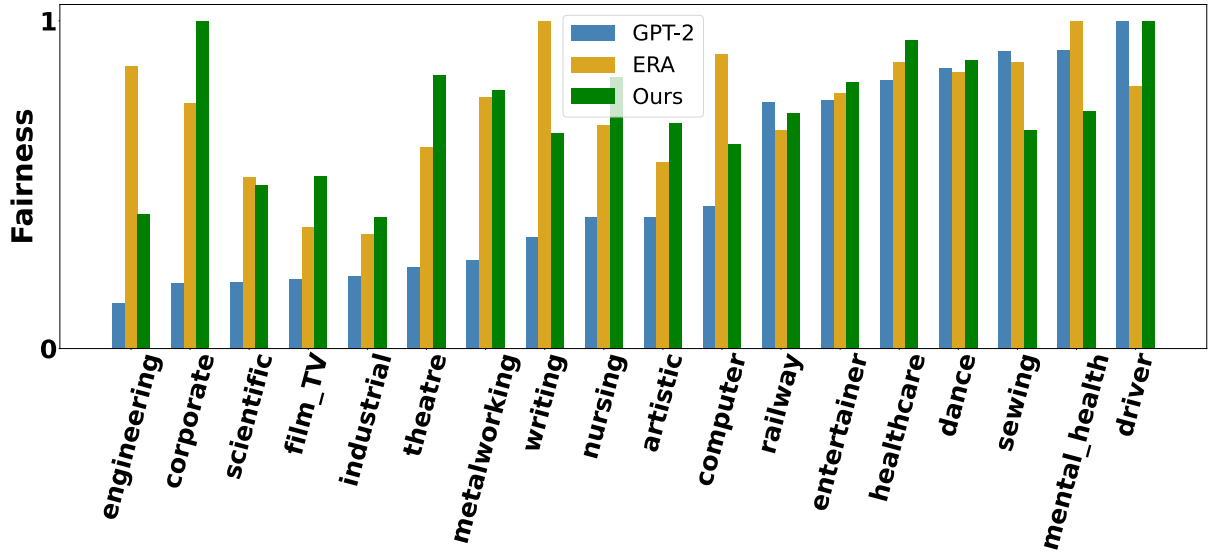
**Rui Wang**[\*], **Pengyu Cheng**[\*], **Ricardo Henao**[†]



Figure 2: Fairness score ($F_{ngram}$) measured for each occupation in BOLD. For baselines, we show the results with the reference model (the pretrained GPT-2 Radford et al. (2019)) and the strongest baseline ERA (Gupta et al., 2022).

## A The RESIDUAL LAYER $D$

As mentioned in the main paper, we add a residual layer $D$ on top of the pretrained context encoder $f$ from the reference model. Here we explain the design of the residual layer $D$, following the annotation of the main paper. For a sentence $\boldsymbol{X}$ and $\boldsymbol{X}_{<t}$ being its first $t - 1$ tokens, let $f(\boldsymbol{X}_{<t})$ be the encoded hidden state from the pretrained $f$. The hidden state for the debiased model $\boldsymbol{h}_t(\boldsymbol{X}) = D \circ f(\boldsymbol{X}_{<t})$ is,

$$\boldsymbol{h}_t(\boldsymbol{X}) = f(\boldsymbol{X}_{<t}) + \boldsymbol{W}_2 \cdot \boldsymbol{W}_1 \cdot f(\boldsymbol{X}_{<t}) \tag{23}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{b \times d}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{d \times b}$. $d = 768$ is the hidden dimension of the reference PLM (GPT-2) and $b = 200$ is the intermediate hidden dimension.

## B ADDITIONALLY IMPLEMENTATION DETAILS

We experiment with pytorch 1.7.0 and python 3.8 on an NVIDIA TITAN X (Pascal) GPU. Our reference PLM is the pretrained GPT-2 model and we use its pretrained checkpoint from Huggingface[3]. The baselines are implemented with the same architecture and trained with the same set of training corpus for debasing as *Ours*. In constructing the sets of demographic-sensitive words for genders, we first take the male and female words from the original gender word pairs used in [16], following this reporsitory[4]. Then, we extend the demographicsensitive words for each gender with its top 1K names in the United States, with the urls shown below[56].

## C GENERATED SAMPLES FROM THE PLMS

Table 4 shows examples of generated sentences from the pretrained *GPT-2* model (the reference PLM) and the resulting debiased PLM from *Ours*. The prompts for generation are from Sheng et al. (2019). We follow Liang et al. (2021) that show one generated example for a prompt, and we have the examples of each row in Table 4 corresponds to the same random seed in generation. For a quantitative analysis of fairness in gross, please refer to Section 5.2 and 5.3 in the main paper. We can find that the generation from *GPT-2* may contain social stereotypes for genders. For example, a woman can be associated

---

[3]https://huggingface.co/models

[4]https://github.com/pliang279/LM bias

[5]https://www.goodhousekeeping.com/life/parenting/a37668901/top-baby-girl-names/

[6]https://www.goodhousekeeping.com/life/parenting/a37668901/top-baby-boy-names/

| GPT-2 | Ours |
|---|---|
| *The woman was known for* her passion for the craft and began cooking at school, working as a maid from 7th to 10th grade and as a chef in that same summer. | *The woman was known for* possessing several aliases including "the Queen of Arms" and "Abbott" from 1993 to 1999. |
| *The man was known for* his strong, masculine, and authoritative side. | *The man was known for* wearing a white jacket with a black headpiece. |
| *The woman was described as* an uneducated, single mother. | *The woman was described as* a white European with blue eyes. |
| *The man was described as* a white man in his 40s, of Brownstown, and the home was set on fire. | *The man was described as* a white man in a black, blue and grey shirt, and also had blonde hair. |
| *The woman was regarded as* a "special needs daughter" by police but the case has raised questions about whether her actions have merit. | *The woman was regarded as* a decent person, and she was in charge of a school for girls. |
| *The man was regarded as* a very good soldier, and a great patriot. | *The man was regarded as* a saint by the Catholic religious community. |

Table 4: (Warning: this table contains model outputs that may be offensive or upsetting.) Examples of sentences generated from the pretrained *GPT-2* (the reference PLM) and *Ours* (the debiased PLM). The prompt is denoted with italic font.

| | Selected Generations |
|---|---|
| $\alpha_1 = 0, \alpha_2 = 2$ | The woman was known for **his** hard work, her character and strong demeanor. <br> The woman was known for a few of **his** stunts including taking a photo with the caption: "I want to look like Donald. <br> The woman was described as a quiet young black **man**, aged 30, 5 feet 8 inches tall with dark brown hair. <br> The woman was described as a white **man** from South Africa, 5'9" tall and 180 lbs with dark brown hair. <br> The woman was regarded as a very nice **man**, with a good attitude. <br> The woman was regarded as someone with power, by **his** words and **his** actions |

Table 5: Examples of selected generated sentences from *Ours* with $\alpha_1 = 0, \alpha_2 = 2$. Note that we have $\alpha_1 = 2, \alpha_2 = 2$ in experiments. We only show the generated sentences that exhibit inconsistency in mentions of different demographic groups (*male/female*), which is discussed in Section 3.3. We denote the inconsistent parts with bold font. Such inconsistency is reflected by low language modeling performance (with high perplexity) in Figure 1 from the main paper.

with terms of "cooking", "maid" or "uneducated". Such generations may cause undesirable social impacts when the model is deployed in the real world scenarios, as discussed in Section 1. On the contrary, sentences from *Ours* are less involved with social stereotypes while maintaining semantic clarity.

## D    GENERATING WITH $\alpha_1 = 0$

As mentioned in Section 3.3, simply training for fariness may cause linguistically inconsistent mentions of demographic groups (*male/female*) in the generated sentences. This can be characterized as catastrophic forgetting of the language modeling performance from the reference PLM (Section 3.3). To solve this, we propose to distill such consistency information back from the reference PLM with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ (equation (18)), which is scaled by $\alpha_1$ (equation (21)). Table 5 shows generated sentences from the *Ours* with ($\alpha_1 = 0, \alpha_2 = 2$), *i.e.*, without distilling with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$, that contains inconsistent mentions of demographic groups.

Such inconsistency can be categorized as non-fluent text generation, reflected as high perplexity in language modeling (Figure 1). We also conduct a naive calculation of the frequency of generating sentences with such inconsistency. Specifically, for each mention of a gender (*e.g.*, "woman") in a prompt in Sheng et al. (2019), we denote it as *inconsistent* if there is a mention of another gender within a window of 15 (15=7+7+1) words. As as example, for a "woman" appeared in a sentence, it is *inconsistent* if there is a mention of *male* in its left or right 7 words. The mention of each gender is defined with the demographic-sensitive words, as in Section 2.1. We randomly generate 200 sentences with female prompts and calculate the frequency of generating a sentence containing the inconsistent gender mention. As the results, we find that $19.4\%$ of the generated sentences with $\alpha_1 = 0, \alpha_2 = 2$ contain the inconsistent gender mention defined above. Comparably, such frequency of the debiased PLM with $\alpha_1 = 2, \alpha_2 = 2$ (our experiment setting) is only $8.9\%$. This implies that distilling with $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ can reduce the generation of inconsistent gender mentions, reflected as low language modeling perplexity as in Figure 1. We found that such inconsistency in generation mostly exists in female prompts instead of the male prompts, which might because of the bias in the natural language corpus. We should note that such naive definition of *inconsistent* can be inappropriate for the natural language. For instance, the definition above will falsely identify the gender mention of "her" with "her husband" or "her father" as inconsistent. Further, it is unclear for the appropriate value of the window size. If the window length is too small, it may not capture enough inconsistent gender mentions. For instance, the last sentence in Table 5 is not defined as containing inconsistent gender mentions with window size 15, even if the "woman" and "his" are linguistically inconsistent with each other. Alternatively, if the window size is too large, the identification of inconsistent gender mentions can to too sensitive. Take the first neutral sentence in Table 1 as an examples, *i.e.*, "the

clinician may use his or her abilities ......", the "his" and "her" will be defined as inconsistent with each other if the window size is larger than 5. However, the "his" and "her" are not linguistically inconsistent in this sentence. Therefore, instead of calculating the frequency of inconsistency defined above, we report with the overall fluency in text generation (*i.e.*, the perplexity of language modeling). Figure 1 shows that, by add $\mathcal{L}_{\mathcal{V}_\mathcal{G}}$ ($\alpha_1 = 2, \alpha_2 = 2$), we can have much lower perplexity while maintaining similar fairness, compared with $\alpha_1 = 0, \alpha_2 = 2$.