

# Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts

Yue Guo<sup>1</sup>, Yi Yang<sup>1</sup>, Ahmed Abbasi<sup>2</sup>

<sup>1</sup> The Hong Kong University of Science and Technology

<sup>2</sup> University of Notre Dame

yguoar@connect.ust.hk imyiyang@ust.hk aabbasi@nd.edu

## Abstract

Human-like biases and undesired social stereotypes exist in large pretrained language models. Given the wide adoption of these models in real-world applications, mitigating such biases has become an emerging and important task. In this paper, we propose an automatic method to mitigate the biases in pretrained language models. Different from previous debiasing work that uses external corpora to fine-tune the pretrained models, we instead directly probe the biases encoded in pretrained models through prompts. Specifically, we propose a variant of the beam search method to automatically search for *biased prompts* such that the cloze-style completions are the most different with respect to different demographic groups. Given the identified biased prompts, we then propose a distribution alignment loss to mitigate the biases. Experiment results on standard datasets and metrics show that our proposed **Auto-Debias** approach can significantly reduce biases, including gender and racial bias, in pretrained language models such as BERT, RoBERTa and ALBERT. Moreover, the improvement in fairness does not decrease the language models' understanding abilities, as shown using the GLUE benchmark.

## 1 Introduction

Pretrained language models (PLMs), such as masked language models (MLMs), have achieved remarkable success in many natural language processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Brown et al.). Unfortunately, pretrained language models, which are trained on large human-written corpora, also inherit human-like biases and undesired social stereotypes (Caliskan et al., 2017; Bolukbasi et al., 2016; Blodgett et al., 2020). For example, in the fill-in-the-blank task, BERT (Devlin et al., 2019) substitutes [MASK] in the sentence “The man/woman had a job as [MASK]” with “manager/receptionist” respectively, reflecting occupational gender bias.

The human-like biases and stereotypes encoded in PLMs are worrisome as they can be propagated or even amplified in downstream NLP tasks such as sentiment classification (Kiritchenko and Mohammad, 2018), co-reference resolution (Zhao et al., 2019; Rudinger et al., 2018), clinical text classification (Zhang et al., 2020) and psychometric analysis (Abbasi et al., 2021; Ahmad et al., 2020).

However, although it is important to mitigate biases in PLMs, debiasing masked language models such as BERT is still challenging, because the biases encoded in the contextualized models are hard to identify. To address this challenge, previous efforts seek to use additional corpora to retrieve the contextualized embeddings or locate the biases and then debias accordingly. For example, Liang et al. (2020); Kaneko and Bollegala (2021); Garimella et al. (2021) use external corpora to locate sentences containing the demographic-specific words (e.g., man and women) or stereotype words (e.g., manager and receptionist) and then use different debiasing losses to mitigate the biases.

Using external corpora to debias PLMs heavily relies on the quality of the corpora. Empirical results show that different corpora have various effects on the debiasing results: some external corpora do mitigate the bias, while others introduce new biases to the PLMs (Garimella et al., 2021; Liang et al., 2020). This is because the corpora used for debiasing may not have enough coverage of the biases encoded in the PLMs. Nevertheless, our understanding of how to quantitatively assess the level of biases in a corpus remains limited (Blodgett et al., 2020).

Mitigating biases in PLMs without external corpora is an open research gap. Recent work in language model prompting shows that through cloze-style prompts, one can probe and analyze the knowledge (Petroni et al., 2019), biases (May et al., 2019) or toxic content (Ousidhoum et al., 2021) in PLMs. Motivated by this, instead of refer-

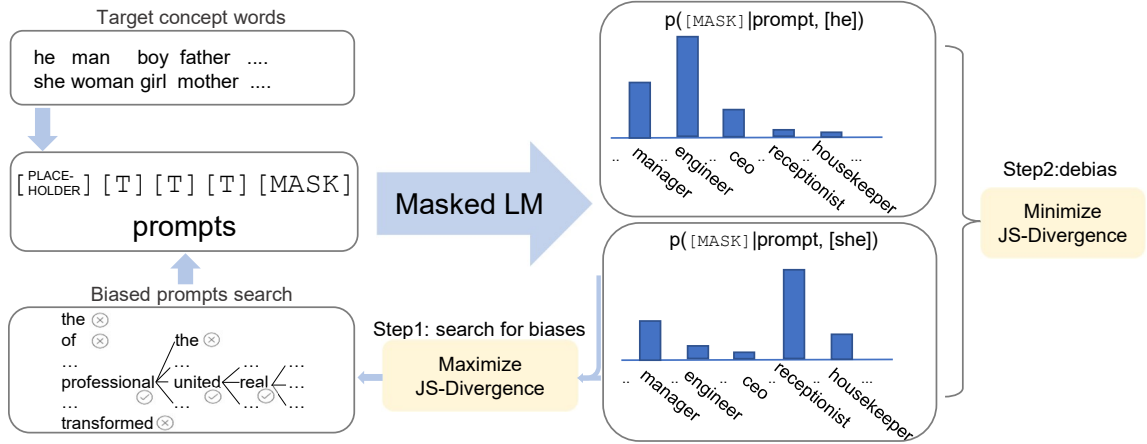


Figure 1: The Auto-Debias framework. In the first stage, our approach searches for the *biased prompts* such that the cloze-style completions (i.e., masked token prediction) have the highest disagreement in generating stereotype words. In the second stage, the language model is fine-tuned by minimizing the disagreement between the distributions of the cloze-style completions.

ring to any external corpus, we directly use cloze-style prompts to probe and identify the biases in PLMs. But what are the biases in a PLM? Our idea is motivated by the assumption that a fair NLP system should produce scores that are independent to the choice of identities mentioned in the text (Prabhakaran et al., 2019). In our context, we propose automatically searching for “discriminative” prompts such that the cloze-style completions have the highest disagreement in generating stereotype words (e.g., manager/receptionist) with respect to demographic words (e.g., man/woman). The automatic *biased prompt* search also minimizes human effort.

After we obtain the biased prompts, we probe the biased content with such prompts and then correct the model bias. We propose an equalizing loss to align the distributions between the [MASK] tokens predictions, conditioned on the corresponding demographic words. In other words, while the automatically crafted biased prompts maximize the disagreement between the predicted [MASK] token distributions, the equalizing loss minimizes such disagreement. Combining the automatic prompts generation and the distribution alignment fine-tuning, our novel method, **Auto-Debias** can debias the PLMs without using any external corpus. Auto-Debias is illustrated in Figure 1.

In the experiments, we evaluate the performance of Auto-Debias in mitigating gender and racial biases in three popular masked language models: BERT, ALBERT, and RoBERTa. Moreover, to alleviate the concern that model debias-

ing may worsen a model’s performance on natural language understanding (NLU) tasks (Meade et al., 2021), we also evaluate the debiased models on GLUE tasks. The results show that our proposed Auto-Debias approach can effectively mitigate the biases while maintaining the capability of language models. We have released the Auto-Debias implementation, debiased models, and evaluation scripts at <https://github.com/Irenehere/Auto-Debias>.

## 2 Related Works

As NLP models are prevalent in real-world applications, a burgeoning body of literature has investigated human-like biases in NLP models. Bias in NLP systems can stem from training data (Dixon et al., 2018), pre-trained word embeddings or can be amplified by the machine learning models. Most existing work focuses on the bias in pre-trained word embeddings due to their universal nature (Dawkins, 2021). Prior work has found that traditional static word embeddings contain human-like biases and stereotypes (Caliskan et al., 2017; Bolukbasi et al., 2016; Garg et al., 2018; Manzini et al., 2019; Gonen and Goldberg, 2019). Debiasing strategies to mitigate static word embeddings have been proposed accordingly (Bolukbasi et al., 2016; Zhao et al., 2018; Kaneko and Bollegala, 2019; Ravfogel et al., 2020).

Contextualized embeddings such as BERT have been replacing the traditional static word embeddings. Researchers have also reported similar human-like biases and stereotypes in contextual

embedding PLMs (May et al., 2019; Kurita et al., 2019; Tan and Celis, 2019; Hutchinson et al., 2020; Guo and Caliskan, 2021; Wolfe and Caliskan, 2021) or in the text generation tasks (Schick et al., 2021; Sheng et al., 2019). Compared to static word embeddings, mitigating the biases in contextualized PLMs is more challenging since the representation of a word usually depends on the word’s context. Garimella et al. (2021) propose to augment the pretraining corpus with demographic-balanced sentences. Liang et al. (2020); Cheng et al. (2021) suggest removing the demographic-direction from sentence representations in a post-hoc fashion. However, augmenting the pretraining corpus is costly and post-hoc debiasing does not mitigate the intrinsic biases encoded in PLMs. Therefore, recent work has proposed to fine-tune the PLMs to mitigate biases by designing different debiasing objectives (Kaneko and Bollegala, 2021; Garimella et al., 2021; Lauscher et al., 2021). They rely on external corpora, and the debiasing results based on these external corpora vary significantly (Garimella et al., 2021). Moreover, Garimella et al. (2021) find that existing debiasing methods are generally ineffective: first, they do not generalize well beyond gender bias; second, they tend to worsen a model’s language modeling ability and its performance on NLU tasks. In this work, we propose a debiasing method that does not necessitate referring to any external corpus. Our debiased models are evaluated on both gender and racial biases, and we also evaluate their performance on NLU tasks.

### 3 Auto-Debias: Probing and Debiasing using Prompts

We propose Auto-Debias, a debiasing technique for masked language models that does not entail referencing external corpora. Auto-Debias contains two stages: First, we automatically craft the *biased prompts*, such that the cloze-style completions have the highest disagreement in generating stereotype words with respect to demographic groups. Second, after we obtain the biased prompts, we debias the language model by a distribution alignment loss, with the motivation that the prompt completion results should be independent to the choice of different demographic-specific words.

#### 3.1 Task Formulation

Let  $\mathcal{M}$  be a Masked Language Model (MLM), and  $\mathcal{V}$  be its vocabulary. The language model

pre-trained with human-generated corpus contains social bias towards certain demographic groups. To mitigate the bias, we have two types of words: *target concepts* which are the paired tokens related to demographic groups (e.g., he/she, man/woman), and *attribute words* which are the stereotype tokens with respect to the target concepts (e.g., manager, receptionist). We denote the target concepts as a set of m-tuples of words  $\mathcal{C} = \{(c_1^{(1)}, c_2^{(1)}, \dots, c_m^{(1)}), (c_1^{(2)}, c_2^{(2)}, \dots, c_m^{(2)}), \dots\}$ . For example, in the two-gender debiasing task, the target concepts are  $\{(he, she), (man, woman), \dots\}$ . In the three-religion debiasing task, the target concepts are  $\{(judaism, christianity, islam), (jew, christian, muslim), \dots\}$ . We omit the superscript of  $\mathcal{C}$  if without ambiguity. We denote the set of attribute words as  $\mathcal{W}$ .

An MLM can be probed by cloze-style prompts. Formally, a prompt  $x_{\text{prompt}} \in \mathcal{V}^*$  is a sequence of words with one masked token [MASK] and one placeholder token. We use  $x_{\text{prompt}}(c)$  to denote the prompt with which the placeholder is filled with a target concept  $c$ . For example, given  $x_{\text{prompt}} = \text{“[placeholder] has a job as [MASK]”}$ , we can fill in the placeholder with the target concept "she" and obtain

$$x_{\text{prompt}}(\text{she}) = \text{she has a job as [MASK]}.$$

Given a prompt and a target concept  $x_{\text{prompt}}(c)$  as the input of  $\mathcal{M}$ , we can obtain the predicted [MASK] token probability as

$$\begin{aligned} p(\text{[MASK]} = v | \mathcal{M}, x_{\text{prompt}}(c)) \\ = \frac{\exp(\mathcal{M}_{\text{[MASK]}}(v | x_{\text{prompt}}(c)))}{\sum_{v' \in \mathcal{V}} \exp(\mathcal{M}_{\text{[MASK]}}(v' | x_{\text{prompt}}(c)))} \end{aligned} \quad (1)$$

where  $v \in \mathcal{V}$ . Prior literature has used this [MASK] token completion task to assess MLM bias (May et al., 2019). To mitigate the bias in an  $\mathcal{M}$ , we hope that the output distribution predicting a [MASK] should be conditionally independent on the choice of any target concept in the m-tuple  $(c_1, c_2, \dots, c_m)$ . Therefore, for different  $c_i \in (c_1, c_2, \dots, c_m)$ , our goal to debias  $\mathcal{M}$  is to make the conditional distributions  $p(\text{[MASK]} = v | \mathcal{M}, x_{\text{prompt}}(c_i))$  as similar as possible.

#### 3.2 Finding Biased Prompts

The first stage of our approach is to generate prompts that can effectively probe the bias from  $\mathcal{M}$ , so that we can remove such bias in the second stage. One straightforward way to design such

---

**Algorithm 1:** Biased Prompt Search

---

**input** : Language model  $\mathcal{M}$ , candidate vocabulary  $\mathcal{V}'$ , target words  $\mathcal{C}$ , stereotype words  $\mathcal{W}$ , prompt length  $PL$ , beam width  $K$ .  
**output** : Generated Biased Prompts  $\mathcal{P}$

- 1  $\mathcal{P} \leftarrow \{\}$ ;
- 2 Candidate prompts  $\mathcal{P}_{can} \leftarrow \mathcal{V}'$ ;
- 3 **for**  $l \leftarrow 1$  **to**  $PL$  **do**
- 4      $\mathcal{P}_{gen} \leftarrow \text{top-}K_{x \in \mathcal{P}_{can}} \{JSD(p([\text{MASK}] | x_{\text{prompt}}(c_i), \mathcal{M}), i \in \{1, 2, \dots, m\})\}$ ;
- 5     // where  $x_{\text{prompt}}(c_i) = c_i \oplus x \oplus [\text{MASK}]$  and we only consider the probability of the attribute words  $\mathcal{W}$  in the  $[\text{MASK}]$  position
- 6      $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}_{gen}$ ;
- 7      $\mathcal{P}_{can} \leftarrow \{x \oplus v | \forall x \in \mathcal{P}_{gen}, \forall v \in \mathcal{V}'\}$
- 8 **end**

---

prompts is by manual generation. For example, “A [placeholder] has a job as [MASK]” is such biased prompts as it generates different mask token probabilities conditioned on the placeholder word being man or woman. However, handcrafting such biased prompts at scale is costly and the models are highly sensitive to the crafted prompts.

To address the problem, we propose *biased prompt search*, as described in Algorithm 1, a variant of the beam search algorithm, to search for the most discriminative, or in other words, the most biased prompts with respect to different demographic groups. Our motivation is to search for the prompts that have the highest disagreement in generating attribute words  $\mathcal{W}$  in the  $[\text{MASK}]$  position. We use Jensen–Shannon divergence (JSD), which is a symmetric and smooth Kullback–Leibler divergence (KLD), to measure the agreement between distributions. In the case of the two-gender debiasing (male/female) task, JSD measures the agreement between the two distributions.

The JSD among distributions  $p_1, p_2, \dots, p_m$  is defined as

$$JSD(p_1, p_2, \dots, p_m) = \frac{1}{m} \sum_i KLD(p_i || \frac{p_1 + p_2 + \dots + p_m}{m}), \quad (2)$$

where the Kullback–Leibler divergence (KLD) between two distributions  $p_i, p_j$  is computed as  $KLD(p_i || p_j) = \sum_{v \in \mathcal{V}} p_i(v) \log(\frac{p_i(v)}{p_j(v)})$ .

Algorithm 1 describes our algorithm for searching biased prompts. The algorithm finds the sequence of tokens  $x$  from the search space to craft prompts, which is firstly the candidate vocabulary space<sup>1</sup>, and then, after the first iteration, the con-

catenation of searched sequences and candidate vocabulary. Specifically, during each iteration, for each candidate  $x$  in the search space, we construct the prompt as  $x_{\text{prompt}}(c_i) = c_i \oplus x \oplus [\text{MASK}]$ , where  $\oplus$  is the string concatenation, for  $c_i$  in an m-tuple  $(c_1, c_2, \dots, c_m)$ . Given the prompt  $x_{\text{prompt}}(c_i)$ ,  $\mathcal{M}$  predicts the  $[\text{MASK}]$  token distribution over attribute words  $\mathcal{W}$  (e.g. manager, receptionist,...):  $p([\text{MASK}] = v | \mathcal{M}, x_{\text{prompt}}(c_i)), v \in \mathcal{W}$ .

Next, we compute the JSD score between  $p([\text{MASK}] = v | \mathcal{M}, x_{\text{prompt}}(c_i))$  for each  $c_i \in (c_1, c_2, \dots, c_m)$ , and select the prompts with high scores — indicating large disagreement between the  $[\text{MASK}]$  predictions for the given target concepts. The algorithm finds the top  $K$  prompts  $x_{\text{prompt}}$  from the search space in each iteration step, and the procedure repeats until the prompt length reaching the pre-defined threshold. We merge all the generated prompts as the final biased prompts set  $\mathcal{P}$ .

### 3.3 Fine-tuning MLM with Prompts

After we obtain the biased prompts, we fine-tune  $\mathcal{M}$  to correct the biases. Specifically, given an m-tuple of target words  $(c_1, c_2, \dots, c_m)$  and a biased prompt  $x_{\text{prompt}}$ , we expect  $\mathcal{M}$  to be unbiased in the sense that  $p([\text{MASK}] = v | \mathcal{M}, x_{\text{prompt}}(c_i)) = p([\text{MASK}] = v | \mathcal{M}, x_{\text{prompt}}(c_j))$  for any  $c_i, c_j \in (c_1, c_2, \dots, c_m)$ . This equalizing objective is motivated by the assumption that a fair NLP system should produce scores that are independent to the choice of the target concepts in our context, men-

---

contains punctuations, word pieces and meaningless words. Therefore, instead of using the vocabulary  $\mathcal{V}$ , we use the 5,000 highest frequency words in Wikipedia as the search space. <https://github.com/IlyaSemenov/wikipedia-word-frequency>

<sup>1</sup>We could use the entire  $\mathcal{V}$  as the search space, but it

tioned in the text (Prabhakaran et al., 2019).

Therefore, given a prompt  $x_{\text{prompt}}$ , our equalizing loss aims to minimize the disagreement between the predicted [MASK] token distributions. Specifically, it is defined as the Jensen-Shannon divergence (JSD) between the predicted [MASK] token distributions:

$$\text{loss}(x_{\text{prompt}}) = \sum_k \text{JSD}(p_{c_1}^{(k)}, p_{c_2}^{(k)}, \dots, p_{c_m}^{(k)}) \quad (3)$$

where  $p_{c_i}^{(k)} = p([\text{MASK}] = v | \mathcal{M}, x_{\text{prompt}}(c_i^{(k)}))$ , for  $v$  in a certain stereotyped word list. And the total loss is the average over all the prompts in the prompt set  $\mathcal{P}$ .

**Discussion:** Another perspective for Auto-Debias is that the debiasing method resembles adversarial training (Goodfellow et al., 2014; Papernot et al., 2017). In the first step, Auto-Debias searches for the biased prompts by maximizing disagreement between the masked language model (MLM) completions. In the second step, Auto-Debias leverages the biased prompts to fine-tune the MLM, by minimizing disagreement between the MLM completions. Taken together, Auto-Debias corrects the biases encoded in the MLM without relying on any external corpus. Overcoming the need to manually specify biased prompts would also make the entire debiasing pipeline more objective.

Recent research has adopted the adversarial training idea to remove biases from sensitive features, representations and classification models (Zhang et al., 2018; Elazar and Goldberg, 2018; Beutel et al., 2017; Han et al., 2021). Our work differs from this line of research in two ways. First, our work aims to mitigate biases in the PLMs. Second, the crafted biased prompts are not adversarial examples.

## 4 Debiasing Performance

We evaluate the performance of Auto-Debias in mitigating biases in masked language models.

**Debiasing strategy benchmarks.** We consider the following debiasing benchmarks. Based on which stage the debiasing technique applies to, the benchmarks can be grouped into three categories.

- *Pretraining:* **CDA** is a data augmentation method that creates a gender-balanced dataset for language model pretraining (Zmigrod et al., 2019). **Dropout** is a debiasing method by increasing the dropout parameters in the PLMs (Webster et al., 2020);

- *Post-hoc:* **Sent-Debias** is a post-processing debias work that removing the estimated gender-direction from the sentence representations (Liang et al., 2020). **FairFil** uses a contrastive learning approach to correct the biases in the sentence representations (Cheng et al., 2021);
- *Fine-tuning:* **Context-Debias** proposes to debias PLM by a loss function that encourages the stereotype words and gender-specific words to be orthogonal (Kaneko and Bollegala, 2021). **DebiasBERT** proposes to use the equalizing loss to equalize the associations of gender-specific words (Garimella et al., 2021). Both works essentially fine-tune the parameters in PLMs.

Our proposed Auto-Debias approach belongs to the fine-tuning category. It does not require any external corpus compared to the previous fine-tuning debiasing approaches.

**Pretrained Models.** In the experiments, we consider three popular masked language models: BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019) and RoBERTa (Liu et al., 2019). We implement BERT, ALBERT, and RoBERTa using the Huggingface Transformers library (Wolf et al., 2020).

**Bias Word List.** Debiasing approaches leverage existing hand-curated target concepts and stereotype word lists to identify and mitigate biases in the PLMs. Those word lists are often developed based on concepts or methods from psychology or other social science literature, to reflect cultural and cognitive biases. In our experiments, we aim to mitigate gender or racial biases. Following prior debiasing approaches, we obtain the gender concept/stereotype word lists used in (Kaneko and Bollegala, 2021)<sup>2</sup> and racial concept/stereotype word lists used in (Manzini et al., 2019)<sup>3</sup>.

**Evaluating Biases: SEAT.** Sentence Embedding Association Test (SEAT) (May et al., 2019) is a common metric used to assess the biases in the PLM embeddings. It extends the standard static word embedding association test (WEAT) (Caliskan et al., 2017) to contextualized word embeddings. SEAT leverages simple templates such as “This is a[n] <word>” to obtain individual

<sup>2</sup><https://github.com/kanekomasa/hiro/context-debias/>

<sup>3</sup><https://github.com/TManzini/DebiasMulticlassWordEmbedding/>

	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8b	avg.
BERT	0.48	0.11	0.25	0.25	0.40	0.64	0.35
+CDA(Zmigrod et al., 2019)	0.46	-0.19	-0.20	0.40	0.12	<b>-0.11</b>	0.25
+Dropout(Webster et al., 2020)	0.38	0.38	0.31	0.40	0.48	0.58	0.42
+Sent-Debias(Liang et al., 2020)	-0.10	-0.44	0.19	0.19	-0.08	0.54	0.26
+Context-Debias(Kaneko and Bollegala, 2021)	1.13	-	0.34	-	0.12	-	0.53
+FairFil(Cheng et al., 2021)	0.18	0.08	<b>0.12</b>	<b>0.08</b>	0.20	0.24	0.15
+Auto-Debias (Our approach)	<b>0.09</b>	<b>0.03</b>	0.23	0.28	<b>0.06</b>	0.16	<b>0.14</b>
ALBERT	0.36	0.18	0.50	0.09	0.33	0.25	0.28
+CDA(Zmigrod et al., 2019)	-0.24	<b>-0.02</b>	0.26	0.31	-0.49	0.47	0.30
+Dropout(Webster et al., 2020)	-0.31	0.09	0.53	<b>-0.01</b>	0.32	<b>0.14</b>	0.24
+Context-Debias(Kaneko and Bollegala, 2021)	0.18	-	<b>-0.05</b>	-	-0.77	-	0.33
+Auto-Debias (Our approach)	<b>0.07</b>	0.15	0.21	0.23	<b>0.16</b>	0.23	<b>0.18</b>
RoBERTa	1.61	0.72	-0.14	0.70	<b>0.31</b>	0.52	0.67
+Context-Debias(Kaneko and Bollegala, 2021)	1.27	-	0.86	-	1.14	-	1.09
+Auto-Debias (Our approach)	<b>0.16</b>	<b>0.02</b>	<b>0.06</b>	<b>0.11</b>	0.42	<b>0.40</b>	<b>0.20</b>

Table 1: Gender debiasing results of SEAT on BERT, ALBERT and RoBERTa. Absolute values closer to 0 are better. Auto-Debias achieves better debiasing performance. The results of Sent-Debias, Context-Debias, FairFil are from the original papers. CDA, Dropout are reproduced from the released model (Webster et al., 2020). "-" means the value is not reported in the original paper.

	Stereo	Anti-stereo	Overall
BERT	55.06	62.14	57.63
+Auto-Debias	<b>52.64</b>	<b>58.44</b>	<b>54.92</b>
ALBERT	<b>54.72</b>	60.19	56.87
+Auto-Debias	43.58	<b>54.47</b>	<b>47.86</b>
RoBERTa	62.89	42.72	54.96
+Auto-Debias	<b>53.53</b>	<b>44.08</b>	<b>49.77</b>

Table 2: Gender debiasing performance on CrowS-Pairs. An ideally debiased model should achieve a score of 50%. Auto-Debias mitigates the overall bias on all three models.

word’s context-independent embeddings, which allows measuring the association between two demographic-specific words (e.g., man and woman) and stereotypes words (e.g., career and family). An ideally unbiased model should exhibit no difference between the demographic-specific words and their similarity to the stereotype words. We report the effect size in the SEAT evaluation. Effect size with an absolute value closer to 0 indicates lower biases. In the experiment, following prior work (Liang et al., 2020; Kaneko and Bollegala, 2021), we use SEAT 6, 6b, 7, 7b, 8, and 8b for measuring gender bias. Also, we use SEAT 3, 3b, 4, 5, and 5b for measuring racial bias. The SEAT test details, including the bias types and demographic/stereotype word associations, are presented in Appendix A.

**Experiment Setting.** In our prompt searching algorithm 1, we set the maximum biased prompt length  $PL$  as five and beam search width  $K$  as 100. In total, we automatically generate 500 biased

prompts for debiasing each model. In the gender debias experiments, we use BERT-base-uncased, RoBERTa-base, and ALBERT-large-v2. In the racial debiasing experiments, we use BERT-base-uncased and ALBERT-base-v2. We use different ALBERT models in the two experiments to allow a fair comparison with existing benchmarks. We do not debias RoBERTa-base in the race experiment because it has a pretty fair score in the SEAT metric. All Auto-Debias models are trained for 1 epoch with AdamW (Loshchilov and Hutter, 2019) optimizer and  $1e^{-5}$  learning rate. All models are trained on a single instance of NVIDIA RTX 3090 GPU card. For gender and race experiments, we run Auto-Debias separately on each base model five times and report the average score for the evaluation metrics<sup>4</sup>.

#### 4.1 Mitigating gender bias

**SEAT.** We report gender debiasing results in Table 1, leading to several findings. First, our proposed Auto-Debias approach can meaningfully mitigate gender bias on the three tested masked language models BERT, ALBERT, and RoBERTa, in terms of the SEAT metric performance. For example, the average SEAT score of the original BERT, ALBERT, and RoBERTa is 0.35, 0.28, and 0.67, respectively. Auto-Debias can substantially reduce the score to 0.14, 0.18, and 0.20. Second, Auto-Debias is more effective in mitigating gender biases compared to the existing state-of-the-art bench-

<sup>4</sup>The SEAT score is based on the average of *absolute* value.

Prompt Length	Generated Prompts
1	substitute, premier, united, became, liberal, major, acting, professional, technical, against, political
2	united domestic, substitute foreign, acting field, eventual united, professional domestic, athletic and
3	professional domestic real, bulgarian domestic assisted, former united free, united former inside
4	eventual united reading and, former united choice for, professional domestic central victoria
5	united former feature right and, former united choice for new, eventual united reading and

Table 3: Examples of prompts generated by Biased Prompt Search (BERT model, for gender).

	SEAT-3	SEAT-3b	SEAT-4
BERT	<b>-0.10</b>	0.37	0.21
+Auto-Debias	0.25	<b>0.19</b>	<b>0.12</b>
ALBERT	0.60	0.29	0.53
+Auto-Debias	<b>0.10</b>	<b>0.12</b>	<b>0.19</b>
	SEAT-5	SEAT-5b	avg.
BERT	0.16	0.34	0.23
+Auto-Debias	<b>0.15</b>	<b>0.17</b>	<b>0.18</b>
ALBERT	0.40	0.46	0.46
+Auto-Debias	<b>0.26</b>	<b>0.19</b>	<b>0.17</b>

Table 4: Mitigating racial biases in BERT and ALBERT. RoBERTa is excluded because it barely exhibits racial bias in terms of the SEAT metric.

marks. BERT is the most studied model in prior work, so we include the state-of-the-art debiasing numbers reported in existing benchmark papers. We can see that Auto-Debias achieves the lowest average SEAT score in all three pretrained model experiments. For example, in SEAT-6 and SEAT-6b, where we examine the association between male/female names/terms and career/family terms, Auto-Debias achieves SEAT scores that are close to 0, indicating the debiased model can almost eliminate the gender bias in the career/family direction. Third, we observe that Auto-Debias, while achieving the lowest average SEAT score, is also relatively stable on SEAT score across different tasks. Conversely, benchmark debiasing approaches have high variance across tasks, which is consistent with recent empirical findings (Meade et al., 2021). This indicates that Auto-Debias is a more stable and generalizable in terms of its debiasing performance.

**CrowS-Pairs.** In addition to the word association test, we also evaluate debiasing performance using the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) (Nangia et al., 2020). This dataset contains a set of sentence pairs that are intended to be minimally distant, semantically speaking, except that one sentence in each pair is considered to be more indicative of stereotyping than the other. The CrowS-Pairs benchmark metric measures the percentage of sentence pairs in which the

language model assigns a higher likelihood to the sentence deemed to be more stereotyping. An ideal model is expected to achieve a score of 50%.

Table 2 shows the debiasing performance on CrowS-Pairs (gender subset) for BERT, ALBERT, and RoBERTa. The original model’s stereotype scores are also presented in the table for direct reference. Note that a score closer to 50 is preferred, as it implies that the model assigns equal probability to male and female sentences. In the BERT and RoBERTa models, Auto-Debias reduces the language models’ bias and assigns more equal likelihood to the sentences in both gender groups. Interestingly, in ALBERT, for the sentences in the dataset that demonstrate stereotypes (*Stereo*), Auto-Debias even over-corrects the stereotypes: it slightly prefers the historically disadvantaged groups. Overall, Auto-Debias can reduce the biases in all three models.

**Biased prompts.** We present some examples of the generated biased prompts in Table 3. Although the biased prompts from Auto-Debias are not grammatical, which is expected in the case of automatically generated prompts (Shin et al., 2020; Zhong et al., 2021), they do contain stereotype related tokens such as professional, political, and liberal. Also, the automated biased generation can minimize human effort and may scale well.

## 4.2 Mitigating racial bias

Mitigating non-gender biases is a challenging task in debiasing research. Meade et al. (2021) empirically show that some of the debiasing techniques considered in our benchmarks generalize poorly in racial debiasing. One of the challenges could be the ambiguity of words (white, black) in different contexts. Therefore, the counterfactual data-augmentation approach or the fine-tuning approach relying on external corpora may be less effective.

In this experiment, we evaluate Auto-Debias’s performance in mitigating racial biases in the PLMs and evaluate the performance using SEAT 3, 3b, 4, 5, and 5b tests. Table 4 reports the SEAT score on

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI
BERT	0.53	0.92	0.88	0.87	0.90	0.84/0.85	0.92	0.58	0.55
+Auto-Debias	0.52	0.92	0.89	0.88	0.91	0.84/0.85	0.91	0.60	0.56
ALBERT	0.59	0.92	0.91	0.91	0.91	0.88/0.87	0.92	0.74	0.55
+Auto-Debias	0.58	0.94	0.91	0.90	0.91	0.87/0.87	0.92	0.75	0.47
RoBERTa	0.52	0.94	0.89	0.88	0.91	0.88/0.87	0.93	0.61	0.56
+Auto-Debias	0.46	0.94	0.89	0.87	0.91	0.88/0.87	0.93	0.61	0.56

Table 5: GLUE test results on the original and the gender-debiased PLMs. Auto-Debias can mitigate the bias while also maintaining the language modeling capability.

the original and debiased BERT and ALBERT. The RoBERTa model is excluded because it barely exhibits racial biases in the SEAT test with an average score of 0.05. We do not include other debiasing benchmarks in Table 4 because most benchmark papers do not focus on racial debiasing. Thus, we focus on comparing the Auto-Debias performance against the original models.

We can see from Table 4 that Auto-Debias can meaningfully mitigate the racial biases in terms of the SEAT metric. Note that the racial SEAT test examines any association difference between European-American/African American names/terms and the stereotype words (pleasant vs. unpleasant). For example, on BERT, Auto-Debias considerably mitigates the racial bias in 4 out of 5 SEAT sub-tests, and the overall score is reduced from 0.23 to 0.18. On ALBERT, Auto-Debias also significantly mitigates the bias in all subsets.

## 5 Does Auto-Debias affect downstream NLP tasks?

Meade et al. (2021) find that the previous debiasing techniques often come at a price of worsened performance in downstream NLP tasks, which implies that prior work might over-debias. Our work instead directly probes the bias encoded in PLM, alleviating the concern of over-debias. In this section, we evaluate the gender debiased BERT/ALBERT/RoBERTa on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019), to examine the capabilities of the language models. The results are reported in Table 5. The racial-debiased PLM models achieve similar GLUE scores.

Auto-Debias performs on par with the base models on most natural language understanding tasks. There is only one exception: CoLA dataset. CoLA evaluates linguistic acceptability, judging whether a sentence is grammatically correct. Our method

adjusts the distribution of words using prompts, which may affect the grammatical knowledge contained in PLMs. But overall speaking, Auto-Debias does not adversely affect the downstream performance. Taking the results together, we see that Auto-Debias can alleviate the bias concerns while also maintaining language modeling capability.

## 6 Discussion

Prompts have been an effective tool in probing the internal knowledge relations of language models (Petroni et al., 2019), and they can also reflect the stereotypes encompassed in PLMs (Ousidhoum et al., 2021; Sheng et al., 2019). Ideally, when prompted with different demographic targets and potential stereotype words, a fair language model’s generated predictions should be equally likely. Our method shows that, from the other direction, imposing fairness constraints on the prompting results can effectively promote the fairness of a language model.

We also observe a trade-off between efficiency and equity: tuning with more training steps, more prompts and more target words leads to a fairer model (which can even make the SEAT score very close to 0), however, it comes at the price of harming the language modeling ability. Over-tuning may harm the internal language patterns. It is important to strike a balance between efficiency and equity with appropriate fine-tuning.

Also, in order not to break the desirable connections between targets and attributes, carefully selecting the target words and stereotyped attribute words is crucial. However, acquiring such word lists is difficult and depends on the downstream applications. Some prior work establishes word lists based on theories, concepts, and methods from psychology and other social science literature (Kaneke and Bollegala, 2021; Manzini et al., 2019). However, such stereotyped word lists are usually lim-

ited, are often contextualized, and offer limited coverage. Moreover, word lists about other protected groups, such as the groups related to education, literacy, or income, or even intersectional biases (Abbasi et al., 2021), are still missing. One promising method to acquire such word lists is to probe related words from a pre-trained language model, for example, “the man/woman has a job as [MASK]” yields job titles that reflect the stereotypes. We leave such probing-based stereotype word-list generation as an important and open future direction.

## 7 Conclusion

In this work, we propose Auto-Debias, a framework and method for automatically mitigating the biases and stereotypes encoded in PLMs. Compared to previous efforts that rely on external corpora to obtain context-dependent word embeddings, our approach automatically searches for biased prompts in the PLMs. Therefore, our approach is effective, efficient, and is perhaps also more objective than prior methods that rely heavily on manually crafted lists of stereotype words. Experimental results on standard benchmarks show that Auto-Debias reduces gender and race biases more effectively than prior efforts. Moreover, the debiased models also maintain good language modeling capability. Bias in NLP systems can stem from different aspects such as training data, pretrained embeddings, or through amplification when fine-tuning the machine learning models. We believe this work contributes to the emerging literature that sheds light on practical and effective debiasing techniques.

## Acknowledgement

This work was funded in part through U.S. NSF grant IIS-2039915 and an Oracle for Research grant entitled “NLP for the Greater Good.”

## References

- Ahmed Abbasi, David Dobolyi, John P. Lalor, Richard G. Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of EMNLP*, pages 3748–3758.
- Faizan Ahmad, Ahmed Abbasi, Jingjing Li, David G. Dobolyi, Richard G. Netemeyer, Gari D. Clifford, and Hsinchun Chen. 2020. A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems (TOIS)*, 38(1):1–29.
- Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. 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, ACL 2020, Online, July 5–10, 2020*, pages 5454–5476. 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and Pranav et al. Shyam. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *Proceedings of ICLR*.
- Hillary Dawkins. 2021. Marked attribute bias in natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4214–4226.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of EMNLP*, pages 11–21.

## A Appendix: SEAT Test Details

We present more information on the SEAT tests that are used in the experiments, in Table 6.

## B Appendix: Target Word Lists

We provide details about the gender and racial word lists used in the debiasing experiments.

For gender, we use the target concept words and stereotype words listed in (Kaneko and Bollegala, 2021).

For race, we use the target concept words and stereotype words listed in (Manzini et al., 2019), with a slight modification on the target concept words. We present the racial concept word lists below:

**African American:** black, african, black, africa, africa, africa, black people, african people, black people, the africa

**European American:** caucasian, caucasian, white, america, america, europe, caucasian people, caucasian people, white people, the america

Bias type	Test	Demographic-specific words	Stereotype words
Racial	SEAT-3	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-3b	European-American/African American terms	Pleasant vs. Unpleasant
	SEAT-4	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5b	European-American/African American terms	Pleasant vs. Unpleasant
Gender	SEAT-6	Male vs. Female names	Career vs. Family
	SEAT-6b	Male vs. Female terms	Career vs. Family
	SEAT-7	Male vs. Female terms	Math vs. Arts
	SEAT-7b	Male vs. Female names	Math vs. Arts
	SEAT-8	Male vs. Female terms	Science vs. Arts
	SEAT-8b	Male vs. Female names	Science vs. Arts

Table 6: The SEAT test details, extended from (Caliskan et al., 2017).