Unmasking the Trade-off: Measuring Gender Bias Mitigation and Over-debiasing Effects in Pretrained Language Models

Anonymous ACL submission

Abstract

Pretrained language models (PLMs) have demonstrated success across many natural language processing tasks. However, they have been shown to encode gender bias present in the corpora they are trained on. Existing bias mitigation methods are usually devised to remove all associations related to gender. This can hurt the performance of PLMs, because of a possible loss of typical associations (e.g., not associating the word “mother” with female). To measure the extent of loss of typical gender associations (i.e., over-debiasing), we introduce the Typical Associations evaluation corpus for Gender (TA-Gender). We find that three popular debiasing methods result in substantial loss of typical gender associations. Our results highlight the importance of mitigating bias without removing typical gender associations, and our dataset constitutes the first benchmark to evaluate information loss.

1 Introduction

In recent years, pretrained language models (PLMs) (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2020) trained on large-scale corpora have become the de-facto backbone of modern NLP systems. These models are trained on minimally filtered real world text which reflects social biases of the real world (Sun et al., 2019; Bender et al., 2021), which in turn are encoded in PLMs and can propagate into downstream applications (Bolukbasi et al., 2016; Caliskan et al., 2017; Kiritchenko and Mohammad, 2018; May et al., 2019; Kurita et al., 2019). Considering the wide use of PLMs, the propagation of social bias in these models poses a danger of reinforcing stereotypes (Sun et al., 2019; Bender et al., 2021).

A number of methods have been introduced to remove social bias from PLMs (Zhao et al., 2017; Lu et al., 2020; Zmigrod et al., 2019; Hall Maudslay et al., 2019; Liang et al., 2020; Huang et al., 2020), which are designed to bleach all associations with the debiasing target (e.g., gender) from PLMs in a non-discriminate fashion. A loss of typical associations (e.g., not associating “dad” with male) can result in over-debiasing and hurt performance of PLMs and their utility in downstream tasks such as question answering or common sense reasoning.

This paper presents a systematic analysis of gender debiasing and over-debiasing across a diverse range of PLMs using established debiasing methods. We conduct a suite of experiments to compare a number of bias evaluation metrics and introduce new metrics to measure the effects of over-debiasing in English. Specifically, we introduce the Typical Associations evaluation corpus for Gender (TA-Gender), a template-based challenge dataset that measures the degree of typical associations present in PLMs (Figure 1, top).

The boundary between typical and inappropriate is fluid, and we highlight weaknesses and put forward usage recommendations of our approach: (1) we treat gender as binary, as common in related work (Manzini et al., 2019; Bartl et al., 2020), but this by no means reflect the complexity of the concept (Cao and Daumé III, 2020); (2) our data set encodes typical, not factual, associations: we acknowledge the fluidity of the concept of gender, and the fact that concepts like “pregnancy” and “fatherhood” are not fully restricted to one gender; (3) our method cannot account for figurative language or cultural differences in language use across communities. Further discussion and recommendation on how to use our benchmark can be found in the Ethical Considerations (Section 9).

Our experiments test three widely used debiasing methods — counterfactual data substitution (Hall Maudslay et al., 2019), embedding regular-
is one of the saleswomen.

is a danseur.

While our work focuses on gender as just one dimension of social biases, the presented framework can be extended to additional bias dimensions such as racial or religious bias (exercising caution when defining typicality, as discussed above for the gender case). In sum, our contributions are:

- A systematic evaluation of debiasing methods and bias metrics over a representative selection of PLMs of varying sizes.
- A benchmark data set, TA-Gender, designed to quantify the extent to which debiasing incurs a loss of typical gender associations.
- An analysis of the extent of over-debiasing across four debiasing methods and six PLMs.

2 Related Work

**Bias Evaluation** Caliskan et al. (2017) propose the Word Embedding Association Test (WEAT) to measure biases in word embeddings through the strength of association between target words (e.g., gender pronouns) and attribute words (e.g., gender neutral occupations). An unbiased model should exhibit no difference between the associations of attribute words with target words of different gender. May et al. (2019) extended this to biases in pretrained contextualized language models through the Sentence Encoder Association Test (SEAT), by encoding Caliskan et al. (2017)’s WEAT terms in simple sentences, and measuring the associative strength of target sentences and attribute sentences as the cosine distances between their sentence embeddings. Focusing on masked language models, Kurita et al. (2019) propose logprob-score to evaluate bias in BERT. Instead of using cosine distances between embeddings, the association between target and attribute words is estimated by the probability of masked token predictions. We use both SEAT and logprob-score to evaluate bias in this work.

An alternative to template-based bias evaluation methods are crowdsourced datasets that capture societal notions of stereotypes across domains including gender, race or religion (Nadeem et al., 2020; Nangia et al., 2020). We consider the gender portion of StereoSet (Nadeem et al., 2020) in this work, which consists of crowd-sourced tuples of sentences capturing a stereotyped, an anti-stereotyped, and a semantically meaningless version of the same context (Figure 1, bottom). As such, StereoSet can be leveraged to evaluate bias (leveraging anti-stereotyped vs stereotyped alternatives) as well as the basic language modeling abilities of pretrained language models (leveraging the semantically meaningful vs meaningless alternatives). While StereoSet can provide an approximate measure of the extent of over-debiasing, the test is designed to probe stereotypical associations and general language modeling abilities. In contrast, our proposed dataset (TA-Gender), directly focuses on typical (i.e., (near) factual) information by using templates that directly probe for typical associations. Figure 1 illustrates the differences between TA-Gender and StereoSet. We compare the Stereoset LM-score against our proposed benchmark in our over-debiasing study (§ 7.2).

**Bias Mitigation** Several debiasing methods have been proposed in previous work. One line of work focuses on modifying training data to remove bias. Lu et al. (2020) propose counterfactual data augmentation to remove gender bias. For each sentence, it generates a counterfactual sentence which is identical except that all gender words are swapped to terms of the opposite gender. Then the union of original sentences and swapped sentences are used for training. Name-based counterfactual data substitution (CDS; Hall Maudslay et al. (2019)) is a variant which swaps the gender words including first names in-place for a random sub-set of sentences to counterbalance bias.

Another line of work introduces additional bias-
regularization terms into the loss function. Huang et al. (2020) propose embedding regularization (ER) for debiasing autoregressive models. The key idea of embedding regularization is to apply a regularization term to encourage models to produce similar embeddings for sentences that only differ from each other in the gender words. Specifically, for each sentence $s$ in the training set, a gender-swapped counterfactual sentence $s_c$ is generated. The cosine distance between embeddings of $s$ and $s_c$ is added as a regularization term ($Reg(s, s_c)$) to the language modeling objective ($L_{lm}(s)$):

$$L(s, s_c) = L_{lm}(s) + \lambda Reg(s, s_c),$$ (1)

where $\lambda$ denotes a weight parameter.

Another family of methods employs post-hoc debiasing. Bolukbasi et al. (2016) propose word embedding debiasing to mitigate gender bias in word embeddings by establishing a gender subspace using embeddings from a predefined list of gender-specific words e.g., “he”, “she”. This gender subspace is then removed from the final embeddings. Sentence debias (SD; Liang et al. (2020)) extends word embedding debiasing to the sentence level, and makes it amenable to removing gender bias from PLMs. Specifically, SD assumes access to a diverse set of sentences from real corpora with gender-specific words. Then the same methodology is applied over sentence embeddings in order to obtain gender-debiased sentence representations.

### 3 Bias Mitigation

We now describe the three debiasing methods used in our experiments (§3.2) and the data used by these methods (§3.1). In terms of PLMs, we include small and large versions of BERT, GPT2, and BART in our experiments, as representative instances, respectively, of encoder, decoder, and encoder-decoder PLMs.

<table>
<thead>
<tr>
<th>Template</th>
<th>Sent A (correct)</th>
<th>Sent B (incorrect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target is <strong>adj_attr</strong>.</td>
<td>He is male.</td>
<td>She is male.</td>
</tr>
<tr>
<td>Target is <strong>single_attr</strong>.</td>
<td>This man is a danseur.</td>
<td>This woman is a danseur.</td>
</tr>
<tr>
<td>As <strong>single_attr</strong>, Target deserves a wonderful life.</td>
<td>As a dad, my husband deserves a wonderful life.</td>
<td>As a dad, my wife deserves a wonderful life.</td>
</tr>
<tr>
<td>As one of the <strong>plural_attr</strong>, Target deserves a wonderful life.</td>
<td>As one of the housewives, she deserves a wonderful life.</td>
<td>As one of the housewives, he deserves a wonderful life.</td>
</tr>
<tr>
<td>Target is one of the <strong>plural_attr</strong>.</td>
<td>My daughter is one of the saleswomen.</td>
<td>My son is one of the saleswomen.</td>
</tr>
<tr>
<td>After Target made that decision, his/her own life has been changed.</td>
<td>After my aunt made that decision, her own life has been changed.</td>
<td>After my uncle made that decision, her own life has been changed.</td>
</tr>
</tbody>
</table>

Table 1: List of templates in TA-Gender, together with examples of generated sentence pairs.

### 3.1 Data

The GAP corpus (Webster et al., 2018) is a gender-balanced dataset which is originally designed for evaluating coreference resolution systems. It consists of 4,454 diverse contexts sampled from Wikipedia and is widely used for investigating gender bias (Kurita et al., 2019; Bartl et al., 2020). We follow Bartl et al. (2020) and split each multi-sentence context into individual sentences. The resulting data is used to train the debiasing methods, which we describe next.

### 3.2 Debiasing Methods

#### Counterfactual Data Substitution (CDS).

Bartl et al. (2020) tested CDS on BERT. Here, we extend the method to GPT2 and BART. In line with Bartl et al. (2020), we apply CDS on the GAP corpus, and fine-tune the PLMs based on the gender-flipped data using their (unsupervised) pretraining objectives. As the GAP corpus is gender-balanced, we expect a debiasing effect for both male and female associations after fine-tuning.

#### Embedding Regularization (ER).

We use the same set of paired gender words in CDS for swapping gender words. ER is originally proposed for GPT2, and we extend it to BERT and BART, with two adjustments: (i) masked token prediction and mask filling are used as training objectives for BERT and BART respectively; and (ii) to produce a sentence representation, we compute an average of the contextual embeddings (i.e., representations from the final layer) from the encoder for BERT and decoder for BART. Note that the sentence rep-

---

$^4$We use the list of paired gender words and implementation provided by Hall Maudslay et al. (2019).

$^5$For BERT, we use the code provided by Gururangan et al. (2020) for masking words. For BART, we follow standard masking procedures from Lewis et al. (2020) where 30% words are masked.
representation for GPT2 is computed using the last hidden state, following Huang et al. (2020).6

**Sentence Debias (SD).** We again use the same set of paired gender words in CDS and ER. We extend the method to GPT2 and BART, as it was originally designed for BERT. To compute the gender subspace, we use sentences containing gender-specific words from GAP. To compute sentence representations for GPT2, BERT and BART, we use the same approach as ER.

### 4 Bias Evaluation

To measure biases in the PLMs, we experiment with four metrics: SEAT-v1 (May et al., 2019), SEAT-v2 (Kurita et al., 2019), logprob-score (Kurita et al., 2019), and stereo-score (Nadeem et al., 2020). Before explaining these metrics (§ 4.2), we first describe the templated data which the first three metrics rely on (§ 4.1).

#### 4.1 Data

Table 3 illustrates the input for SEAT and logprob-score, respectively, each being a sentence including a target and an attribute word. Target words are words that are associated with the bias dimension of interest. In our case the bias dimension is gender, and so target words are gender words/pronouns. Attribute words are objectively neutral words that might have association with the bias dimension due to stereotypes, e.g., gender neutral occupations words like “nurse”. We take the union of target words previously adopted by Kurita et al. (2019) and Bartl et al. (2020) as target words to evaluate bias (see Table 2 for the full list). For attribute words, we use the occupations in Bartl et al. (2020).7 To convert these attribute and target words into sentences, we use the templates from Bartl et al. (2020) (Table 4) and May et al. (2019) (Table 5).8

We divide the data into a development set (20%) and a test set (80%) based on the target words and attribute words, and use the dev set to tune the hyper-parameters of debiasing methods (§3.2).

---

6Note that for GPT2, we also tested using the average contextual embeddings as sentence representation but found inferior performance.

7Bartl et al. (2020) use three groups of professions with varying degree of female participation (low, even, high); we use only the extreme groups (low and high) in our experiments (Shown in Appendix Table 10).

8We use only 5 out of the original 9 SEAT templates from May et al. (2019), as 4 of them are not covered by our bias evaluation terms (e.g., templates with names and things).

---

**Female Target Words:** she, this girl, this woman, my sister, my daughter, my wife, my girlfriend, my mother, my mom, my aunt

**Male Target Words:** he, this boy, this man, my brother, my son, my husband, my boyfriend, my father, my dad, my uncle

Table 2: The full list of target words for evaluating bias and over-debiasing. Validation set terms are in *italics*.

#### 4.2 Metrics

Both SEAT and logprob-score measure bias by computing the difference in association between the target and attribute words (i.e. the effect size), and an effect size closer to 0 indicates lower bias. We also compute the p-value of a permutation test to denote the significance of the effects size (Kurita et al., 2019; May et al., 2019). 9

**logprob-score** The association between target and attribute words in a sentence s is computed as the log probability ratio between: (1) the target word in s with only the target word masked; and (2) the target word in s with both the target and attribute masked. logprob-score is originally proposed for BERT, and it can be applied without modification for BART.10 As GPT2 does not use masked tokens, we do not assess it using this metric. logprob-score uses the templates in Table 4.

**SEAT-v1** The original SEAT introduced by May et al. (2019), using the templates in Table 5. Association of a target word with an attribute word is measured as the cosine distance between their sentence encodings. We compute sentence encodings for GPT2, BERT and BART as we do for the debiasing methods (ER and SD): we use the final hidden state for GPT2, and average contextual embeddings for BERT (encoder) and BART (decoder).

**SEAT-v2** A variant of SEAT introduced by Kurita et al. (2019) that uses the same set of templates as logprob-score where target and attribute words appear within a single sentence (Table 4). The converted sentence pair (for measuring cosine similarity) has either of the target or attribute word masked.11
spokeswoman." For each sentence pair (associations corpus for Gender (TA-Gender) ("fish" is not used here). A perfect stereo-score measures the extent to which a PLM prefers a sentence express-tent to which a PLM prefers the sentence expressing a typical association. Specifically, for BERT and BART we mask the target word and compute the probability of the two options (e.g., “daughter” vs. “son” in example TA1 in Figure 1) and select the option with a higher probability. For several example sentence pairs. An ideal model should assign a higher probability to sentence a, compared to sentence b.

We use the same list of target words as used in bias evaluation (§4; Table 2). For attribute words, we use terms from Bolukbasi et al. (2016) and filter them with the following rules: (i) we keep only the singular forms; (ii) we remove multi-word phrases when similar single-words exists (e.g., “twin brother” is removed since “brother” exists); (iii) we remove any words that can apply to both genders (e.g., “chairman”). We do so by checking each attribute word definition in two lexicons: the Oxford English Dictionary and Wiktionary, and remove the attribute if at least one of the resources suggests that the word is not gender specific. The resulting set of attribute words in TA-Gender was independently verified by two authors of this paper. It consists of 67 attribute words.

We finally create six templates each containing a target and attribute word, as listed in Table 1.

### Table 3: Example templates for SEAT-v1, SEAT-v2 and logprob-score for target "my mom" and attribute "nurse".

<table>
<thead>
<tr>
<th>Target Template</th>
<th>Converted sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEAT-v1</td>
<td>&quot;There is &lt;person&gt;.&quot;, &quot;There is a &lt;profession&gt;.&quot;, &quot;There is my mom.&quot;, &quot;There is a nurse.&quot;</td>
</tr>
<tr>
<td>SEAT-v2</td>
<td>&quot;&lt;person&gt; is a &lt;profession&gt;.&quot;, &quot;&lt;mask_token&gt; is a &lt;profession&gt;.&quot;, &quot;My mom is a &lt;mask_token&gt;.&quot;, &quot;&lt;mask_token&gt; is a nurse.&quot;</td>
</tr>
<tr>
<td>logprob-score</td>
<td>&quot;&lt;person&gt; is a &lt;profession&gt;.&quot;, &quot;My mom is a nurse.&quot;</td>
</tr>
</tbody>
</table>

Table 4: SEAT-v2 and logprob-score templates (Bartl et al., 2020).

<table>
<thead>
<tr>
<th>Target Template</th>
<th>Attribute Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &lt;person&gt; is a &lt;profession&gt;</td>
<td>2 &lt;person&gt; works as a &lt;profession&gt;</td>
</tr>
<tr>
<td>3 &lt;person&gt; applied for the position of &lt;profession&gt;</td>
<td>4 &lt;person&gt;, the &lt;profession&gt;, had a good day at work.</td>
</tr>
<tr>
<td>5 &lt;person&gt; wants to become a &lt;profession&gt;</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: SEAT-v1 templates from May et al. (2019).

#### stereo-score

Unlike the previous metrics, stereo-score measures the extent to which a PLM prefers a stereotypical association over anti-stereotypical association using crowdsourced sentences developed by Nadeem et al. (2020). For example, in context SS1 in Figure 1, the stereotypical option is “soft” and the anti-stereotypical option is “determined” (“fish” is not used here). A perfect stereo-score is 50%, which implies that a language model is oblivious to (anti-)stereotyping (i.e. it selects stereotypes and anti-stereotypes with equal probability).

5 Over-Debiasing Evaluation

To measure the loss of typical gender associations in PLMs after debiasing, we develop the Typical Associations corpus for Gender (TA-Gender).

The proposed dataset consists of 2,610 sentence pairs where each sentence contains one target word and one attribute word. Target words are gender nouns or pronouns and attribute words are characteristics or occupations which are typically associated with one gender, such as "pregnant" or "spokeswoman". For each sentence pair (a, b), sentence a contains a typical association while sentence b is atypical. The two sentences are identical, except for the gender target word. Table 1 shows

---

12 We use the “intrasentence” instances in original dataset, as we are interested in only single-sentence context.

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert-base-uncased</td>
<td>12</td>
<td>110M</td>
</tr>
<tr>
<td>gpt2</td>
<td>12</td>
<td>117M</td>
</tr>
<tr>
<td>bart-base</td>
<td>6 enc + 6 dec</td>
<td>139M</td>
</tr>
<tr>
<td>bert-large-uncased</td>
<td>24</td>
<td>336M</td>
</tr>
<tr>
<td>bart-large-medium</td>
<td>24</td>
<td>345M</td>
</tr>
<tr>
<td>bart-large</td>
<td>12 enc + 12 dec</td>
<td>406M</td>
</tr>
</tbody>
</table>

Table 6: Configurations of the PLMs. “enc”=encoder, “dec”=decoder.

---

\[\text{https://www.oed.com/} \]
\[\text{https://www.wiktionary.org/} \]
\[\text{Shown in Appendix Table 9.} \]
We test both small and large variants of GPT2, (Table 7: Evaluated bias before (column 3) and after (column 4–6) debiasing. “Pre-deb.” denotes pre-debias. An over-debiasing do the methods exhibit? (§7.2); and (2) how much variations: (1) to what extent do common debiasing methods reduce the gender bias in PLMs of varying size and architecture? (§7.1); (2) how much over-debiasing do the methods exhibit? (§7.2); and (3) what is the trade-off between debiasing and over-debiasing (§7.3).

### 6 Implementation Details

For model implementation, we use the Huggingface transformers library (Wolf et al., 2020). We test both small and large variants of GPT2, BERT and BART; configurations of these models are given in Table 6. For the debiasing methods (CDS, ER and SD), we tune hyper-parameters based on their debiasing performance using the development partition of the bias evaluation data (§4.1).\textsuperscript{16}

### 7 Results

Our experiments are designed to answer three questions: (1) to what extent do common debiasing methods reduce the gender bias in PLMs of varying size and architecture? (§7.1); (2) how much over-debiasing do the methods exhibit? (§7.2); and

\textsuperscript{16}See Table 12 in the appendix for hyper-parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Metric</th>
<th>Pre-deb.</th>
<th>CDS</th>
<th>ER</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>SEAT-v1</td>
<td>+1.700</td>
<td>-0.154</td>
<td>+0.008</td>
<td>-0.096</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>+1.943</td>
<td>+0.179</td>
<td>-0.245</td>
<td>+1.363</td>
</tr>
<tr>
<td></td>
<td>logprob-score</td>
<td>+1.966</td>
<td>+1.329</td>
<td>+1.348</td>
<td>+1.098</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>63.93</td>
<td>58.84</td>
<td>59.34</td>
<td>53.97</td>
</tr>
<tr>
<td>BERT-large</td>
<td>SEAT-v1</td>
<td>+0.335</td>
<td>-0.166</td>
<td>-0.172</td>
<td>+0.026</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>+1.493</td>
<td>-0.002</td>
<td>-0.123</td>
<td>+0.325</td>
</tr>
<tr>
<td></td>
<td>logprob-score</td>
<td>+1.972</td>
<td>+0.772</td>
<td>+0.865</td>
<td>+1.256</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>63.14</td>
<td>60.31</td>
<td>59.61</td>
<td>55.06</td>
</tr>
<tr>
<td>BART-base</td>
<td>SEAT-v1</td>
<td>+0.428</td>
<td>+0.072</td>
<td>+0.135</td>
<td>+0.178</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>+1.404</td>
<td>+0.270</td>
<td>+0.671</td>
<td>+0.629</td>
</tr>
<tr>
<td></td>
<td>logprob-score</td>
<td>+1.651</td>
<td>+1.427</td>
<td>+1.466</td>
<td>+1.363</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>50.57</td>
<td>47.77</td>
<td>47.31</td>
<td>54.57</td>
</tr>
<tr>
<td>BART-large</td>
<td>SEAT-v1</td>
<td>+0.505</td>
<td>+0.028</td>
<td>+0.170</td>
<td>+0.027</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>+1.377</td>
<td>-0.137</td>
<td>+0.341</td>
<td>+1.049</td>
</tr>
<tr>
<td></td>
<td>logprob-score</td>
<td>+1.691</td>
<td>+1.131</td>
<td>+1.046</td>
<td>+1.207</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>53.59</td>
<td>54.10</td>
<td>52.90</td>
<td>58.40</td>
</tr>
<tr>
<td>GPT2</td>
<td>SEAT-v1</td>
<td>+0.285</td>
<td>-0.079</td>
<td>-0.048</td>
<td>-0.027</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>+0.747</td>
<td>+0.210</td>
<td>-0.041</td>
<td>+0.023</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>62.67</td>
<td>54.74</td>
<td>54.68</td>
<td>57.92</td>
</tr>
<tr>
<td>GPT2-medium</td>
<td>SEAT-v1</td>
<td>-0.330</td>
<td>+0.080</td>
<td>+0.041</td>
<td>-0.076</td>
</tr>
<tr>
<td></td>
<td>SEAT-v2</td>
<td>-0.298</td>
<td>-0.104</td>
<td>+0.063</td>
<td>+0.012</td>
</tr>
<tr>
<td></td>
<td>stereo-score</td>
<td>65.58</td>
<td>47.34</td>
<td>38.66</td>
<td>55.16</td>
</tr>
</tbody>
</table>

Table 7: Evaluated bias before (column 3) and after (column 4–6) debiasing. “Pre-deb.” denotes pre-debias. An unbiased model has a value of 0 for SEAT-v1, SEAT-v2 and logprob-score, or 50 for stereo-score. Bold values indicate statistically significant effect sizes ($p < 0.01$).

\textsuperscript{16}See Table 12 in the appendix for hyper-parameters.
7.2 Over-Debiasing

Next we turn to the over-debiasing effects after PLMs are debiased, using our TA-Gender data and TA-score. We compare TA-score against LM-score from StereoSet (Nadeem et al., 2020), which is designed to test general language modelling abilities by measuring the selection accuracy of PLMs for masked words in a context sentence. Using the example of StereoSet context SS1 in Figure 1, a PLM would be presented two options: (1) a stereotype or anti-stereotype word (randomly chosen; e.g., “soft” or “determined”) as the correct option; and (2) a meaningless word in context (e.g., “fish”) as the incorrect option. LM-score is the proportion of correct predictions.

Table 8 shows the over-debiasing results using TA-score and LM-score, both capturing genuine language modeling abilities of PLMs. We desire (a) high values before and after debiasing; and (b) no drop in performance caused by debiasing, assuming that typical associations will be retained. The last 3 columns in Table 8 denote the difference of TA/LM-score before and after debiasing. A negative value means the model is over-debiased and there is a loss of typical associations.

**Before debiasing**, it can be seen that given a PLM, the larger variant generally has better TA/LM-score (exception: TA-score of BART), implying that the larger models are better language models. Overall, BERT appears to be the best PLM in terms of capturing typical gender associations (TA-score closest to 100).

**After debiasing**, we observe that all debiasing methods lead to a substantial decrease in TA-score (negative values), indicating that there is an over-debiasing effect (i.e. the debiased PLMs have lost some typical gender associations). LM-score, on the other hand, is largely unable to detect this; interestingly, it even found improvements (positive values) in some instances (e.g., BART). The only exception here is GPT2-medium, where LM-score detect a larger over-debiasing effects compared to TA-score (although both found an over-debiasing effect). These results highlight the effectiveness of TA-score for measuring a loss of typical gender association, unlike LM-score which tests general language model ability.

### Table 8: Over-debiasing results

<table>
<thead>
<tr>
<th>Model</th>
<th>Metric</th>
<th>Pre-deb.</th>
<th>CDS</th>
<th>ER</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>TA-score</td>
<td>95.1</td>
<td>13.2</td>
<td>-12.2</td>
<td>-27.4</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>86.0</td>
<td>-0.10</td>
<td>-0.30</td>
<td>-17.6</td>
</tr>
<tr>
<td>BERT-large</td>
<td>TA-score</td>
<td>98.6</td>
<td>-13.2</td>
<td>-13.1</td>
<td>-24.0</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>86.8</td>
<td>-4.20</td>
<td>-3.70</td>
<td>-15.5</td>
</tr>
<tr>
<td>BART-base</td>
<td>TA-score</td>
<td>82.4</td>
<td>-15.5</td>
<td>-15.6</td>
<td>-16.8</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>69.0</td>
<td>-2.60</td>
<td>-2.70</td>
<td>-2.50</td>
</tr>
<tr>
<td>BART-large</td>
<td>TA-score</td>
<td>77.9</td>
<td>-9.60</td>
<td>-11.1</td>
<td>-15.9</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>69.3</td>
<td>+2.20</td>
<td>-4.70</td>
<td>+4.20</td>
</tr>
<tr>
<td>GPT2</td>
<td>TA-score</td>
<td>76.7</td>
<td>-14.6</td>
<td>-19.8</td>
<td>-50.8</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>93.3</td>
<td>-9.20</td>
<td>-9.50</td>
<td>-5.70</td>
</tr>
<tr>
<td>GPT2-medium</td>
<td>TA-score</td>
<td>84.2</td>
<td>-19.1</td>
<td>-19.9</td>
<td>-12.5</td>
</tr>
<tr>
<td></td>
<td>LM-score</td>
<td>93.6</td>
<td>-26.5</td>
<td>-37.8</td>
<td>-43.5</td>
</tr>
</tbody>
</table>

Table 8: Over-debiasing results. “Pre-deb.” denotes pre-debias. An ideal model has a TA/LM-score of 100. Last 3 columns present the difference of TA/LM-score before and after debiasing (negative values indicate over-debiasing).

According to these metrics; in fact, GPT2-medium exhibits anti-stereotypical biases (indicated by negative values). On the other hand, stereo-score shows a slightly different trend, where we found bias in BERT and GPT2 but not BART. These inconsistencies suggest that investigating the source of these discrepancies and their behavior under different models and data conditions is a pressing research direction. They also suggest that it is important to use a variety of metrics for assessing biases considering their different outcomes.

**After debiasing**, it can be seen that all debiasing methods (CDS, ER and SD) successfully removed bias to some extent (SEAT-v1/SEAT-v2/logprob-score closer to 0 or stereo-score closer to 50), and this is largely consistent across all metrics. The only minor exception here is BART’s stereo-score, although that can be explained by the fact that it has low bias in the first place (i.e. its pre-debias stereo-score is already close to 50). Overall, according to logprob-score there is still bias in the models after debiasing (most effect sizes are >> 0). GPT2 appears to retain the least bias after debiasing.

### 7.3 Trade-off

Next we investigate if there is a trade-off between debiasing and over-debiasing. To this end, we select an appropriate hyper-parameter to vary debiasing strength for each debiasing method. For CDS, which replaces gender words in contexts to create counterfactual sentences, we manipulate the gender-flipping rate (i.e. the number of sentences where a gender word is switched). For ER, we vary...
the λ hyper-parameter which controls the regularization term (Equation 1). SD uses a list of paired gender words to compute the gender sub-space and we vary the number of paired words to control the amount of debiasing, with the idea that using a smaller number of paired words would produce a less debiased model.\footnote{We train each model on a single NVIDIA V100 GPU, and the training process takes around 5 hrs.}

Figure 2 shows the trade-off between debiasing (SEAT-v1, SEAT-v2, or logprob-score) and over-debiasing (1−TA-score) for BERT-base and BART-base. Similar patterns were observed for the other PLMs (cf., Figure 3 in the Appendix).

For both axes, a lower value indicates better performance, and an ideal model would be completely unbiased (SEAT-v1/SEAT-v2/logprob-score = 0) and still retain typical associations to gender (1−TA-score = 0) after debiasing. Generally, we see that the hyper-parameters we choose for each debiasing method result in an effective trade-off, and that as the strength of debiasing increases, there is a general increase in loss of typical associations. Overall, CDS appears to achieve the best trade-off across metrics and models. Taken together, our results highlight the importance of measuring both debiasing and over-debiasing effects when assessing model bias and debiasing methods, as the complete bleaching of all gender associations, including typical associations, may negatively impact model performance on downstream tasks.

8 Discussion and Conclusions

We introduce an approach to measure the effects of over-debiasing, i.e. the loss of typical associations, after model debiasing. We also presented a systematic comparison of debiasing methods across bias metrics and PLM architectures. We focus on gender as the bias dimension, and develop TA-Gender, a dataset of over 2.6K sentence pairs for measuring over-debiasing through probes for typical associations. We show that three widely used debiasing methods (CDS, ER and SD) have a tendency to over-remove gender associations, highlighting the need to develop debiasing methods that eliminate bias without removing important associations.

We emphasize that the notion of typical associations can depend on the application and user profile, and recommend using it in combination with other diagnostic tests (see § 9 for further discussion). To the best of our knowledge we are one of the first studies to investigate over-debiasing, and our results pave way for a number of future research directions, including extending the methodology to other bias dimensions (e.g., race or religion), explaining the discrepancies of bias metrics across models and data conditions, and improving debiasing methods to reduce the extent of over-debiasing.
9 Ethical Considerations

The distinction between debiasing and loss of information can be a fine line, especially in the context of complex concepts such as gender, however, successful debiasing hinges on retained model utility. We proposed a metric and model to help capture the subtle distinction between problematic and less problematic assumptions. In doing so, we made a number of simplifications and assumptions which must be made explicit: (1) we assume a binary notion of gender, which does not reflect the complexity of the concept (Cao and Daumé III, 2020); (2) our benchmark encodes largely cis-normative associations, glossing over the fact that e.g., transmen are well capable of motherhood and that gender can be fluid over time; (3) our method cannot account for figurative language ("He is such a princess about [...]").

Responsible use. We encourage researchers who use our benchmark to adhere to the following recommendations: First, it should be considered whether the assumptions listed above reflect the world view and use case of the modeler and/or application. Second, we recommend to use our dataset and metric in combination with a range of other evaluation metrics and data sets, for instance StereoSet or WEAT, in order to obtain a holistic view of model performance.

References


Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer


A Computation of Effect Size and p-value

Given two sets of target words \( T_1 \) and \( T_2 \) and two sets of attribute words \( A_1 \) and \( A_2 \), the normalized association (denoted as effect size) is

\[
\frac{\text{mean}_{x \in T_1} s(x, A_1, A_2) - \text{mean}_{y \in T_2} s(y, A_1, A_2)}{\text{std} \text{ } \text{dev}_{w \in T_1 \cup T_2} s(w, A_1, A_2)}
\]

Where \( s(t, A_1, A_2) \) is computed by:

\[
\text{mean}_{a \in A_1} \text{asso}(t, a) - \text{mean}_{b \in A_2} \text{asso}(t, b)
\]

\( \text{asso}(t, a) \) computes associations between the target word \( t \) and the attribute word \( a \). For SEAT-v1 and SEAT-v2, \( t \) and \( a \) will be converted into the target sentence and attribute sentence, then the association is computed as the cosine distance between sentence embeddings. For logprob-score, the association is the normalized probabilities of target words produced by masked token prediction.

The permutation test is used in WEAT for measuring significance of results. The null hypothesis is that there is no difference between \( T_1 \) and \( T_2 \) in terms of their associations to \( A_1 \) and \( A_2 \). The permutation test computes the likelihood of the null hypothesis by computing the probability that a random permutation of the target words would generate the greater or equal difference in sample means. Let \( (T_1, T_2) \) denote the set of all possible partitions of target sets \( T_1 \cup T_2 \), then the p-value of permutation test is

\[
\text{Prob}_i[s(T_1^i, T_2^i, A_1, A_2) \geq s(T_1, T_2, A_1, A_2)]
\]

Where \( s(T_1, T_2, A_1, A_2) \) is

\[
\sum_{x \in T_1} s(x, A_1, A_2) - \sum_{y \in T_2} s(y, A_1, A_2)
\]
**Female-specific words:** actress, aunt, bride, businesswoman, chairwoman, congresswoman, councilwoman, daughter, female, gal, girl, girlfriend, goddess, granddaughter, grandma, grandmother, heiress, her, heroine, hostess, housewife, lady, lesbian, mama, matriarch, mistress, mom, mommy, mother, niece, nun, pregnant, princess, queen, saleswoman, schoolgirl, sister, spokewoman, stepdaughter, stepmother, wife, woman

**Male-specific words:** boy, boyfriend, bridegroom, brother, businessman, dad, daddy, danseur, father, gentleman, godfather, grandfather, grandpa, grandson, his, husband, male, man, nephew, schoolboy, son, stepfather, stepson, uncle, widower

Table 9: The list of attribute words for TA-Gender (in alphabetical order).

**Female-dominated Occupations:** secretary, childcare worker, billing clerk, phlebotomist, vocational nurse, medical records technician, speech-language pathologist, paralegal, hairdresser, bookkeeper, kindergarten teacher, medical assistant, dietitian, housekeeper, dental hygienist, teacher assistant, registered nurse, health aide, receptionist, dental assistant

**Male-dominated Occupations:** plumber, operating engineer, security system installer, mason, mining machine operator, floor installer, heating mechanic, carpenter, steel worker, electrician, logging worker, mobile equipment mechanic, taper, bus mechanic, service technician, conductor, repairer, roofer, firefighter, electrical installer

Table 10: The list of profession terms from Bartl et al. (2020) used in this work. Validation set terms are denoted in *italics*.

**Hyper-parameter** | **Values**
--- | ---
Swap rate | [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95]
λ | [0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.0]
Ratio of pairs | [0.01, 0.05, 0.1, 0.15, 0.2, 0.4, 0.6, 1.0]

Table 11: Varied hyper-parameters for investigating the trade-off between debiasing and over-debiasing effects. The gender-flipping rate (swap rate), λ, and the proportions of adopted paired words (ratio of pairs) separately controls the debiasing effect of CDS, ER and SD.

**Hyper-parameter** | **BERT-base** | **BERT-large** | **GPT2** | **GPT2-medium** | **BART-base** | **BART-large**
--- | --- | --- | --- | --- | --- | ---
Swap rate | 1.0 | 1.0 | 0.9 | 0.9 | 1.0 | 1.0
λ | 0.5 | 0.5 | 0.5 | 1.0 | 1.25 | 1.25
Ratio of pairs | 0.05 | 0.05 | 1.0 | 1.0 | 0.07 | 0.07
Batch size | 8 | 2 | 8 | 2 | 8 | 2
Learning rate | 2e-5 | 2e-5 | 5e-5 | 5e-5 | 2e-5 | 2e-5
Epoch | 8 | 8 | 8 | 8 | 8 | 8

Table 12: Hyper-parameters that decided by the development set. The gender-flipping rate (swap rate), λ, and the proportions of adopted paired words (ratio of pairs) separately controls the debiasing effect of CDS, ER and SD.
Figure 3: Trade-off between debiasing and over-debiasing for BERT-large (a); BART-large (b); GPT2 (c); and GPT2-medium (d). Debiasing performance is measured using SEAT-v1 (left); SEAT-v2 (mid); and logprob-score (right).