MABEL: Attenuating Gender Bias using Textual Entailment Data

Jacqueline He* **Mengzhou Xia** **Christiane Fellbaum** **Danqi Chen**
Department of Computer Science, Princeton University
jacquelinehe00@gmail.com
{mengzhou, fellbaum, danqic}@cs.princeton.edu

Abstract

Pre-trained language models encode undesirable social biases, which are further exacerbated in downstream use. To this end, we propose MABEL (a Method for Attenuating Gender Bias using Entailment Labels), an intermediate pre-training approach for mitigating gender bias in contextualized representations. Key to our approach is the use of a contrastive learning objective on counterfactually augmented, gender-balanced entailment pairs from natural language inference (NLI) datasets. We also introduce an alignment regularizer that pulls identical entailment pairs along opposite gender directions closer. We extensively evaluate our approach on intrinsic and extrinsic metrics, and show that MABEL outperforms previous task-agnostic debiasing approaches in terms of fairness. It also preserves task performance after fine-tuning on downstream tasks. Together, these findings demonstrate the suitability of NLI data as an effective means of bias mitigation, as opposed to only using unlabeled sentences in the literature. Finally, we identify that existing approaches often use evaluation settings that are insufficient or inconsistent. We make an effort to reproduce and compare previous methods, and call for unifying the evaluation settings across gender debiasing methods for better future comparison.[1]

1 Introduction

Pre-trained language models have reshaped the landscape of modern natural language processing (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019). As these powerful networks are optimized to learn statistical properties from large training corpora imbued with significant social biases (e.g., gender, racial), they produce encoded representations that inherit undesirable associations as a byproduct (Zhao et al., 2019; Webster et al., 2020; Nadeem et al., 2021). More concerningly, models trained on these representations can not only propagate but also amplify discriminatory judgments in downstream applications (Kurita et al., 2019).

A multitude of recent efforts have focused on alleviating biases in language models. These can be classed into two categories (Table 1): 1) *task-specific* approaches perform bias mitigation during downstream fine-tuning, and require data to be annotated for sensitive attributes; 2) *task-agnostic* approaches directly improve pre-trained representations, most commonly either by removing discriminative biases through projection (Dev et al., 2020; Liang et al., 2020; Kaneko and Bollegala, 2021), or by performing intermediate pre-training on gender-balanced data (Webster et al., 2020; Cheng et al., 2021; Lauscher et al., 2021; Guo et al., 2022), resulting in a new encoder that transfers fairness effects downstream via standard fine-tuning.

In this work, we present MABEL, a novel and lightweight method for attenuating gender bias. MABEL is task-agnostic and can be framed as an intermediate pre-training approach with a contrastive learning framework. Our approach hinges on the use of entailment pairs from supervised natural language inference datasets (Bowman et al., 2015; Williams et al., 2018). We augment the training data by swapping gender words in both premise and hypothesis sentences and model them using a contrastive objective. We also propose an alignment regularizer, which minimizes the distance between the entailment pair and its augmented one. MABEL optionally incorporates a masked language modeling objective, so that it can be used for token-level downstream tasks.

To the best of our knowledge, MABEL is the first to exploit supervised sentence pairs for learning fairer contextualized representations. Supervised contrastive learning via entailment pairs is known to learn a more uniformly distributed rep-

[*] This work was done before JH graduated from Princeton University.

[1] Our code is publicly available at https://github.com/princeton-nlp/MABEL.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9681–9702
December 7-11, 2022 ©2022 Association for Computational Linguistics

resentation space, wherein similarity measures between sentences better correspond to their semantic meanings (Gao et al., 2021). Meanwhile, our proposed alignment loss, which pulls identical sentences along contrasting gender directions closer, is well-suited to learning a fairer semantic space.

We systematically evaluate MABEL on a comprehensive suite of intrinsic and extrinsic measures spanning language modeling, text classification, NLI, and coreference resolution. MABEL performs well against existing gender debiasing efforts in terms of both fairness and downstream task performance, and it also preserves language understanding on the GLUE benchmark (Wang et al., 2019). Altogether, these results demonstrate the effectiveness of harnessing NLI data for bias attenuation, and underscore MABEL's potential as a general-purpose fairer encoder.

Lastly, we identify two major issues in existing gender bias mitigation literature. First, many previous approaches solely quantify bias through the Sentence Encoding Association Test (SEAT) (May et al., 2019), a metric that compares the geometric relations between sentence representations. Despite scoring well on SEAT, many debiasing methods do not show the same fairness gains across other evaluation settings. Second, previous approaches evaluate on extrinsic benchmarks in an inconsistent manner. For a fairer comparison, we either reproduce or summarize the performance of many recent methodologies on major evaluation tasks. We believe that unifying the evaluation settings lays the groundwork for more meaningful methodological comparisons in future research.

2 Background

2.1 Debiasing Contextualized Representations

Debiasing attempts in NLP can be divided into two categories. In the first category, the model learns to disregard the influence of sensitive attributes in representations during fine-tuning, through projection-based (Ravfogel et al., 2020, 2022), adversarial (Han et al., 2021a,b) or contrastive (Shen et al., 2021; Chi et al., 2022) downstream objectives. This approach is *task-specific* as it requires fine-tuning data that is annotated for the sensitive attribute. The second type, *task-agnostic* training, mitigates bias by leveraging textual information from general corpora. This can involve computing a gender subspace and eliminating it from encoded representations (Dev et al., 2020; Liang et al., 2020; Dev et al., 2021; Kaneko and Bollegala, 2021), or by re-training the encoder with a higher dropout (Webster et al., 2020) or equalizing objectives (Cheng et al., 2021; Guo et al., 2022) to alleviate unwanted gender associations.

We summarize recent efforts of both task-specific and task-agnostic approaches in Table 1. Compared to task-specific approaches that only debias for the task at hand, task-agnostic models produce fair encoded representations that can be used toward a variety of applications. MABEL is task-agnostic, as it produces a general-purpose debiased model. Some recent efforts have broadened the scope of task-specific approaches. For instance, Meade et al. (2022) adapt the task-specific Iterative Nullspace Linear Projection (INLP) (Ravfogel et al., 2020) algorithm to rely on Wikipedia data for language model probing. While non-task-agnostic approaches can potentially be adapted to general-purpose debiasing, we primarily consider other task-agnostic approaches in this work.

2.2 Evaluating Biases in NLP

The recent surge of interest in fairer NLP systems has surfaced a key question: how should bias be quantified? *Intrinsic* metrics directly probe the upstream language model, whether by measuring the geometry of the embedding space (Caliskan et al., 2017; May et al., 2019; Guo and Caliskan, 2021), or through likelihood-scoring (Kurita et al., 2019; Nangia et al., 2020; Nadeem et al., 2021). *Extrinsic* metrics evaluate for fairness by comparing the system's predictions across different populations on a downstream task (De-Arteaga et al., 2019a; Zhao et al., 2019; Dev et al., 2020). Though opaque, intrinsic metrics are fast and cheap to compute, which makes them popular among contemporary works (Meade et al., 2022; Qian et al., 2022). Comparatively, though extrinsic metrics are more interpretable and reflect tangible social harms, they are often time- and compute-intensive, and so tend to be less frequently used.[2]

To date, the most popular bias metric among task-agnostic approaches is the Sentence Encoder Association Test (SEAT) (May et al., 2019), which compares the relative distance between the encoded representations. Recent studies have cast doubt on the predictive power of these intrinsic indicators. SEAT has been found to elicit counter-intuitive re-

[2] As Table 17 in Appendix F indicates, many previous bias mitigation approaches limit evaluation to 1 or 2 metrics.

Method	Proj. based	Con. obj.	Gen. aug.	LM probe	Fine-tune	Intermediate pre-training data
Task-specific approaches						
INLP (Ravfogel et al., 2020)	✓			✓*	✓	Wikipedia*
Con (Shen et al., 2021)		✓			✓	-
DADV (Han et al., 2021b)					✓	-
GATE (Han et al., 2021a)					✓	-
R-LACE (Ravfogel et al., 2022)	✓				✓	-
Task-agnostic approaches						
CDA (Webster et al., 2020)			✓	✓	✓	Wikipedia (1M steps, 36h on 8x 16 TPU)
Dropout (Webster et al., 2020)				✓	✓	Wikipedia (100K steps, 3.5h on 8x 16 TPU)
ADELE (Lauscher et al., 2021)			✓	✓	✓	Wikipedia, BookCorpus (105M sentences)
Bias Projection (Dev et al., 2020)	✓		⋇	✓	✓	Wikisplit (1M sentences)
OSCaR (Dev et al., 2021)			⋇	✓		SNLI$^\sharp$ (190.1K sentences)
Sent-Debias (Liang et al., 2020)	✓		✓	✓	✓	WikiText-2, SST, Reddit, MELD, POM
Context-Debias (Kaneko and Bollegala, 2021)	✓		⋇	✓	✓	News-commentary-v1 (87.66K sentences)
Auto-Debias (Guo et al., 2022)					✓	Bias prompts generated from Wikipedia (500)
FairFil (Cheng et al., 2021)		✓	✓	❦	✓	WikiText-2, SST, Reddit, MELD, POM
★MABEL (ours)		✓	✓	✓	✓	MNLI, SNLI with gender terms (134k sentences)

Table 1: Properties of existing gender debiasing approaches for *contextualized* representations. **Proj. based**: projection-based. **Con. obj.**: based on contrastive objectives. **Gen. aug.**: these approaches use a seed list of gender terms for counterfactual data augmentation. **LM probe** and **Fine-tune** denote that the approach can be used for language model probing or fine-tuning, respectively. $*$: INLP was originally only used for task-specific fine-tuning; Meade et al. (2022) later adapted it for task-agnostic training on Wikipedia for LM probing. ❦: FairFil shows poor LM probing performance in Table 2 as the debiasing filter is not trained with an MLM head. MABEL fixes this issue by jointly training with an MLM objective. ⋇: these works use a single gender pair "he/she" to calculate the gender subspace. $\sharp$: Dev et al. (2021) fine-tunes on SNLI but does not use it for debiasing.

sults from encoders (May et al., 2019) or exhibit high variance across identical runs (Aribandi et al., 2021). Goldfarb-Tarrant et al. (2021) show that intrinsic metrics do not reliably correlate with extrinsic metrics, meaning that a model could score well on SEAT, but still form unfair judgements in downstream conditions. This is especially concerning as many debiasing studies (Liang et al., 2020; Cheng et al., 2021) solely report on SEAT, which is shown to be unreliable and incoherent. For these reasons, we disregard SEAT as a main intrinsic metric in this work.[3]

Bias evaluation is critical as it is the first step towards detection and mitigation. Given that bias reflects across language in many ways, relying upon a single bias indicator is insufficient (Silva et al., 2021). Therefore, we benchmark not just MABEL, but also current task-agnostic methods against a diverse set of intrinsic and extrinsic indicators.

3 Method

MABEL attenuates gender bias in pre-trained language models by leveraging entailment pairs from natural language inference (NLI) data to produce general-purpose debiased representations. To the best of our knowledge, MABEL is the first method that exploits semantic signals from supervised sentence pairs for learning fairness.

3.1 Training Data

NLI data is shown to be especially effective in training discriminative and high-quality sentence representations (Conneau et al., 2017; Reimers and Gurevych, 2019; Gao et al., 2021). While previous works in fair representation learning use generic sentences from different domains (Liang et al., 2020; Cheng et al., 2021; Kaneko and Bollegala, 2021), we explore using sentence pairs with an *entailment* relationship: a hypothesis sentence that can be inferred to be true, based on a premise sentence. Since gender is our area of interest, we extract all entailment pairs that contain at least one gendered term in either the premise or the hypothesis from an NLI dataset. In our experiments, we explore using two well-known NLI datasets: the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018).

As a pre-processing step, we first conduct counterfactual data augmentation (Webster et al., 2020) on the entailment pairs. For any sensitive attribute term in a word sequence, we swap it for a word

[3] For comprehensiveness, we report MABEL's results on SEAT in Appendix G.

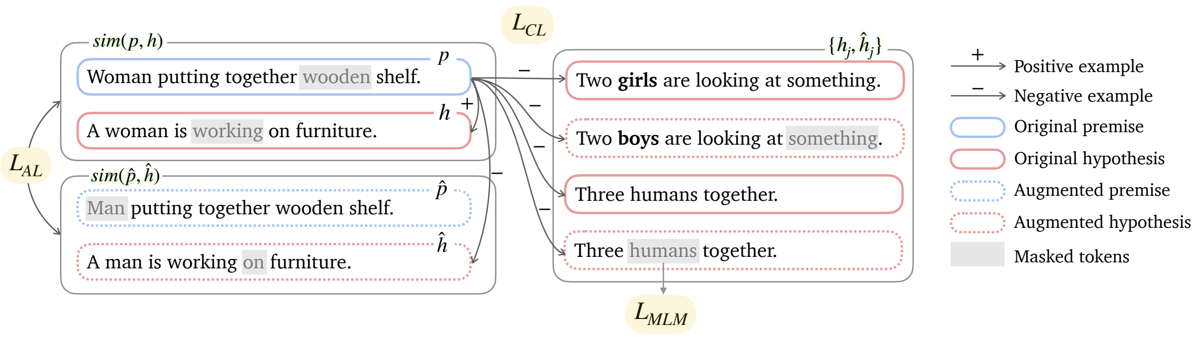

Figure 1: MABEL consists of three losses: 1) an entailment-based contrastive loss ($\mathcal{L}_{\text{CL}}$) that uses the premises's hypothesis as a positive sample and other in-batch hypotheses as negative samples; 2) an alignment loss ($\mathcal{L}_{\text{AL}}$) that minimizes the similarity difference between each original entailment pair and its gender-balanced counterpart; 3) a masked language modeling loss ($\mathcal{L}_{\text{MLM}}$) to recover $p = 15\%$ of the masked tokens.

along the opposite bias direction, i.e., *girl* to *boy*, and keep the non-attribute words unchanged.[4] This transformation is systematically applied to each sentence in every entailment pair. An example of this augmentation, with gender bias as the sensitive attribute, is shown in Figure 1.

3.2 Training Objective

Our training objective consists of three components: a contrastive loss based on entailment pairs and their augmentations, an alignment loss, and an optional masked language modeling loss.

Entailment-based contrastive loss. Training with a contrastive loss induces a more isotropic representation space, wherein the sentences' geometric positions can better align with their semantic meaning (Wang and Isola, 2020; Gao et al., 2021). We hypothesize that this contrastive loss would be conducive to bias mitigation, as concepts with similar meanings, but along opposite gender directions, move closer under this similarity measurement. Inspired by Gao et al. (2021), we use a contrastive loss that encourages the inter-association of entailment pairs, with the goal of the encoder also learning semantically richer associations.[5]

With p as the premise representation and h as the hypothesis representation, let $\{(p_i, h_i)\}_{i=1}^{n}$ be the sequence of representations for n original entailment pairs, and $\{(\hat{p}_i, \hat{h}_i)\}_{i=1}^{n}$ be n counterfactually-augmented entailment pairs. Each entailment pair (and its corresponding augmented pair) forms a positive pair, and the other in-batch sentences constitute negative samples. With m pairs and their augmentations in one training batch, the contrastive objective for an entailment pair i is defined as:

$$\mathcal{L}_{\text{CL}}^{(i)} = -\log \frac{e^{\text{sim}(p_i, h_i)/\tau}}{\sum_{j=1}^{m} e^{\text{sim}(p_i, h_j)/\tau} + e^{\text{sim}(p_i, \hat{h}_j)/\tau}} - \log \frac{e^{\text{sim}(\hat{p}_i, \hat{h}_i)/\tau}}{\sum_{j=1}^{m} e^{\text{sim}(\hat{p}_i, h_j)/\tau} + e^{\text{sim}(\hat{p}_i, \hat{h}_j)/\tau}},$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function, and τ is the temperature. $\mathcal{L}_{\text{CL}}$ is simply the average of all the losses in a training batch. Note that when $h_i = \hat{h}_i$ (i.e., when h_i does not contain any gender words and the augmentation is unchanged), we exclude $\hat{h}_i$ from the denominator to avoid h_i as a positive sample and $\hat{h}_i$ as a negative sample for p_i, and vice versa.

Alignment loss. We want a loss that encourages the intra-association between the original entailment pairs and their augmented counterparts. Intuitively, the features from an entailment pair and its gender-balanced opposite should be taken as positive samples and be spatially close. Our alignment loss minimizes the distance between the cosine similarities of the original sentence pairs (p_i, h_i) and the gender-opposite sentence pairs $(\hat{p}_i, \hat{h}_i)$:

$$\mathcal{L}_{\text{AL}} = \frac{1}{m} \sum_{i=1}^{m} \Big(\text{sim}(\hat{p}_i, \hat{h}_i) - \text{sim}(p_i, h_i) \Big)^2 .$$

We assume that a model is less biased if it assigns similar measurements to two gender-opposite pairs, meaning that it maps the same concepts along different gender directions to the same contexts.[6]

[4] We use the same list of attribute word pairs from Bolukbasi et al. (2016), Liang et al. (2020), and Cheng et al. (2021), which can be found in Appendix A.

[5] In this work, we only refer to the supervised SimCSE model, which leverages entailment pairs from NLI data.

[6] We also explore different loss functions for alignment and report them in Appendix J.

Masked language modeling loss. Optionally, we can append an auxiliary masked language modeling (MLM) loss to preserve the model's language modeling capability. Following Devlin et al. (2019), we randomly mask $p = 15\%$ of tokens in all sentences. By leveraging the surrounding context to predict the original terms, the encoder is incentivized to retain token-level knowledge.

In sum, our training objective is as follows:

$$\mathcal{L} = (1 - \alpha) \cdot \mathcal{L}_{\text{CL}} + \alpha \cdot \mathcal{L}_{\text{AL}} + \lambda \cdot \mathcal{L}_{\text{MLM}},$$

wherein the two contrastive losses are linearly interpolated by a tunable coefficient α, and the MLM loss is tempered by the hyper-parameter λ.

4 Evaluation Metrics

4.1 Intrinsic Metrics

StereoSet (Nadeem et al., 2021) queries the language model for stereotypical associations. Following Meade et al. (2022), we consider intra-sentence examples from the gender domain. This task can be formulated as a fill-in-the-blank style problem, wherein the model is presented with an incomplete context sentence, and must choose between a stereotypical word, an anti-stereotypical word, and an irrelevant word. The Language Modeling Score (LM) is the percentage of instances in which the model chooses a valid word (either the stereotype or the anti-stereotype) over the random word; the Stereotype Score (SS) is the percentage in which the model chooses the stereotype over the anti-stereotype. The Idealized Context Association Test (ICAT) score combines the LM and SS scores into a single metric.

CrowS-Pairs (Nangia et al., 2020) is an intra-sentence dataset of minimal pairs, where one sentence contains a disadvantaged social group that either fulfills or violates a stereotype, and the other sentence is minimally edited to contain a contrasting advantaged group. The language model compares the masked token probability of tokens unique to each sentence. Focusing only on gender examples, we report the stereotype score (SS), the percentage in which a model assigns a higher aggregated masked token probability to a stereotypical sentence over an anti-stereotypical one.

4.2 Extrinsic Metrics

As there has been some inconsistency in the evaluation settings in the literature, we mainly consider the fine-tuning setting for extrinsic metrics and leave the discussion of the linear probing setting to Appendix I.

Bias-in-Bios (De-Arteaga et al., 2019b) is a third-person biography dataset annotated by occupation and gender. We fine-tune the encoder, along with a linear classification layer, to predict an individual's profession given their biography. We report overall task accuracy and accuracy by gender, as well as two common fairness metrics (De-Arteaga et al., 2019b; Ravfogel et al., 2020): 1) GAP^{TPR}_{M}, the difference in true positive rate (TPR) between male- and female-labeled instances; 2) $GAP^{TPR}_{M,y}$, the root-mean square of the TPR gap of each occupation class.

Bias-NLI (Dev et al., 2020) is an NLI dataset consisting of neutral sentence pairs. It is systematically constructed by populating sentence templates with a gendered word and an occupation word with a strong gender connotation (e.g., The *woman* ate a bagel; The *nurse* ate a bagel). Bias can be interpreted as a deviation from neutrality and is determined by three metrics: Net Neutral (NN), Fraction Neutral (FN) and Threshold:τ (T:τ). A bias-free model should score a value of 1 across all 3 metrics. We fine-tune on SNLI and evaluate on Bias-NLI during inference.

WinoBias (Zhao et al., 2018) is an intra-sentence coreference resolution task that evaluates a system's ability to correctly link a gendered pronoun to an occupation across both pro-stereotypical and anti-stereotypical contexts. Coreference can be inferred based on syntactic cues in Type 1 sentences or on more challenging semantic cues in Type 2 sentences. We first fine-tune the model on the OntoNotes 5.0 dataset (Hovy et al., 2006) before evaluating on the WinoBias benchmark. We report the average F1-scores for pro-stereotypical and anti-stereotypical instances, and the true positive rate difference in average F1-scores, across Type 1 and Type 2 examples.

4.3 Language Understanding

To evaluate whether language models still preserve general linguistic understanding after bias attenuation, we fine-tune them on seven classification tasks and one regression task from the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019).[7]

[7] We also evaluate transfer performance on the SentEval tasks (Conneau et al., 2017) in Appendix E.

5 Experiments

5.1 Baselines & Implementation Details

We choose Sent-Debias[8] (Liang et al., 2020), Context-Debias[9] (Kaneko and Bollegala, 2021), and FairFil[10] (Cheng et al., 2021) as our primary baselines. By introducing a general-purpose method for producing debiased representations, these three approaches are most similar in spirit to MABEL. We consider FairFil to be especially relevant as it is also a task-agnostic, contrastive learning approach. Compared to FairFil, MABEL leverages NLI data, and also applies entailment-based and MLM losses to ensure that sentence- and token-level knowledge is preserved.

We evaluate the three aforementioned task-agnostic baselines across all bias benchmarks and also compare against other approaches in Table 1 by reporting the recorded numbers from their original work. Unless otherwise specified, all models, including MABEL, default to `bert-base-uncased` (Devlin et al., 2019) as the backbone encoder. In the standard setting, $\lambda = 0.1$ and $\alpha = 0.05$. Implementation details on MABEL and the task-agnostic baselines can be found in Appendix A and Appendix B, respectively. For our own implementations, we report the average across 3 runs.[11]

5.2 Results: Intrinsic Metrics

As Table 2 shows, MABEL strikes a good balance between language modeling and fairness with the highest ICAT score. Compared to BERT, MABEL retains and even exhibits an average modest improvement (from 84.17 to 84.80) in language modeling. MABEL also performs the best on CrowS-Pairs, with an average metric score of 50.76.

While MABEL does not have the best SS value for StereoSet, we must caution that this score should not be considered in isolation. For example, although FairFil shows a better stereotype score, its language modeling ability (as the LM score shows) is significantly deteriorated and lags behind other approaches. This is akin to an entirely random model that obtains a perfect SS of 50 as it does not contain bias, but would also have a low LM score as it lacks linguistic knowledge.

	StereoSet			CrowS-Pairs
Model	**LM ↑**	**SS ◇**	**ICAT ↑**	**SS ◇**
BERT	84.17	60.28	66.86	57.25 ↑7.25
BERT+Dropout*	83.04	60.66	65.34	55.34 ↑5.34
BERT+CDA*	83.08	59.61	67.11	56.11 ↑6.11
INLP*	80.63	57.25	68.94	51.15 ↑1.15
Sent-Debias*	84.20	59.37	68.42	52.29 ↑2.29
Context-Debias	**85.42**	59.35	69.45	58.01 ↑8.01
Auto-Debias‡	-	-	-	54.92 ↑4.92
FairFil	44.85	**50.93**	44.01	49.03 ↑0.97
MABEL (ours)	84.80	56.92	**73.07**	**50.76** ↑**0.76**

Table 2: Results on StereoSet and CrowS-Pairs (standard deviations are in Table 14). ⋆: the results are reported in Meade et al. (2022); ‡: the results are reported in Guo et al. (2022). ◇: the closer to 50, the better. LM: language modeling score, SS: Stereotype score, ICAT: combined score, defined as $\text{LM} \cdot (\min(\text{SS}, 100 - \text{SS}))/50$.

Model	**Acc. (All) ↑**	**Acc. (M) ↑**	**Acc. (F) ↑**	**TPR GAP ↓**	**TPR RMS ↓**
BERT	84.14	84.69	83.50	1.189	0.144
INLP$^{\flat\dagger}$	70.50	-	-	-	**0.067**
Con*†	81.69	-	-	-	0.168
DADV$^{\sharp\dagger}$	81.10	-	-	-	0.126
GATE$^{\sharp\dagger}$	80.50	-	-	-	0.111
R-LACE$^{\flat\dagger}$	**85.04**	-	-	-	0.115
Sent-Debias	83.56	84.10	82.92	1.180	0.144
Context-Debias	83.67	84.08	83.18	0.931	0.137
FairFil	83.18	83.52	82.78	0.746	0.142
MABEL (ours)	84.85	**84.92**	**84.34**	**0.599**	0.132

Table 3: Results on fine-tuning with the Bias in Bios dataset. ⋆: the results are reported in Shen et al. (2021); ♯: the results are reported in Han et al. (2021a); ♭: the results are reported in Ravfogel et al. (2022); †: the approaches depend on gender annotations.

5.3 Results: Extrinsic Metrics

Bias-in-Bios. As Table 3 indicates, MABEL exhibits the highest overall and individual accuracies, as well as the smallest TPR-GAP when compared against the task-agnostic baselines and BERT.

Still, MABEL and the other task-agnostic models are close in performance to BERT on Bias-in-Bios, which suggests that the fine-tuning process can significantly change a pre-trained model's representational structure to suit a specific downstream task. Furthermore, Kaneko et al. (2022) finds that debiased language models can still re-learn social biases after standard fine-tuning on downstream tasks, which may explain why the task-agnostic methods, which operate upstream, fare worse on

[8] `https://github.com/pliang279/sent_debias`

[9] `https://github.com/kanekomasahiro/context-debias`

[10] As there is no code released, we use our own implementation without an auxiliary regularization term.

[11] Standard deviations can be found in Appendix D.

Model	**TN** ↑	**FN** ↑	**T:0.5** ↑	**T:0.7** ↑
BERT	0.799	0.879	0.874	0.798
ADELE⋆‡	0.557	0.504	-	-
Sent-Debias	0.793	0.911	0.897	0.788
Context-Debias	0.858	0.906	0.902	0.857
Context-Debias⋆‡	0.878	0.968	-	0.893
FairFil	0.829	0.883	0.846	0.845
MABEL (ours)	**0.900**	**0.977**	**0.974**	**0.935**

Table 4: Results on Bias-NLI. We fine-tune the models on SNLI and then evaluate on Bias-NLI. ⋆: results are reported from original papers; ‡: the models are fine-tuned on MNLI.

this particular manifestation of gender bias than on others. Our results also show that task-specific interventions fare better fairness-wise on this task. Methods such as INLP (Ravfogel et al., 2020), GATE (Han et al., 2021a), and R-LACE (Ravfogel et al., 2022) exhibit better TPR RMS scores, although sometimes at the expense of task accuracy. As these methods operate directly on the downstream task, they may have a stronger influence on the final prediction (Jin et al., 2021).

Bias-NLI. We next move to Bias-NLI, where Table 4 indicates that MABEL outperforms BERT and other baselines across all metrics. Unlike in Bias-in-Bios, the results have a greater spread, and MABEL's comparative advantage becomes clear. The FN score denotes that, on average, MABEL correctly predicts neutral 97.7% of the time. MABEL is also more confident in predicting the correct answer, surpassing the 0.7 threshold 93.5% of the time. Other approaches, such as Sent-Debias and FairFil, do not show as clear-cut of an improvement over BERT, despite scoring well on other bmetrics such as SEAT.

As natural language inference requires robust semantic reasoning capabilities to deduce the correct answer, it is a more challenging problem than classification. Therefore, for this task, the models' initialization weights—which store the linguistic knowledge acquired in pre-training—may play a larger impact on the final task accuracy than in Bias-in-Bios.

WinoBias. On this token-level extrinsic task (Table 5), MABEL, and the other bias mitigation baselines, achieve very similar average F1-scores on OntoNotes. However, performance on WinoBias becomes variegated. MABEL shows the best task improvement on anti-stereotypical tasks, with an average 7.25% and 10.58% increase compared to BERT on Type 1 and Type 2 sentences, respectively. The strong performance on anti-stereotypical examples implies that MABEL can effectively weaken the stereotypical token-level associations between occupation and gender. Though MABEL exhibits a marginally lower F1-score on Type 1 pro-stereotypical examples (a 1.64% average decrease compared to the best-performing model, BERT), it has the highest F1-scores across all other categories. Furthermore, it has the best reduction in fairness, with the smallest average TPR-1 and TPR-2 by a clear margin (respectively, 23.73 and 3.41, compared to the next-best average TPR scores at 26.14 and 9.57).

5.4 Results: Language Understanding

As the GLUE benchmark results indicate (Table 6), MABEL preserves semantic knowledge across downstream tasks. On average, MABEL performs marginally better than BERT (82.0% vs. 81.8%), but not as well as BERT fine-tuned beforehand on the NLI task with MNLI and SNLI data (BERT-NLI), at 82.0% vs. 82.1%. Other bias mitigation baselines lag behind BERT, but the overall semantic deterioration remains minimal.

6 Analysis

6.1 Qualitative Comparison

We perform a small qualitative study by visualizing the t-SNE (van der Maaten and Hinton, 2008) plots of sentence representations from BERT, supervised SimCSE (Gao et al., 2021), and MABEL. Following Liang et al. (2020); Cheng et al. (2021), we plot averaged sentence representations of a gendered or neutral concept across different contexts (sentence templates). We re-use the list of gender words, and neutral words with strong gender connotations, from Caliskan et al. (2017).

From Figure 2, in BERT, certain concepts from technical fields such as 'technology' or 'science' are spatially closer to 'man,' whereas concepts from the humanities such as 'art' or 'literature' are closer to 'woman.' After debiasing with MABEL, we observe that the gendered tokens (e.g., 'man' and 'woman,' or 'girl' and 'boy') have shifted closer in the embedding space, and away from the neutral words. While SimCSE shows a similar trend in pulling gendered words away from the neutral words, it also separates the masculine and feminine terms into two distinct clusters. This is

Model	OntoNotes ↑	1A ↑	1P ↑	2A ↑	2P ↑	TPR-1 ↓	TPR-2 ↓
BERT	**73.53**	53.96	**86.57**	82.20	94.67	32.79	12.48
SENT-DEBIAS	72.36	54.11	85.09	83.29	94.73	30.98	11.44
CONTEXT-DEBIAS	73.16	59.40	85.54	83.63	93.20	26.14	9.57
FAIRFIL	71.79	53.24	85.77	77.37	91.40	32.43	14.03
MABEL (ours)	73.48	**61.21**	84.93	**92.78**	**96.20**	**23.73**	**3.41**

Table 5: Average F1-scores OntoNotes and WinoBias, and TPR scores across Winobias categories. 1 = Type 1; 2 = Type 2. A=anti-stereotypical; P=pro-stereotypical.

	CoLA ↑	SST-2 ↑	MRPC ↑	QQP ↑	MNLI ↑	QNLI ↑	RTE ↑	STS-B ↑	Avg. ↑
Model	(mcc.)	(acc.)	(f1/acc.)	(acc./f1)	(acc.)	(acc.)	(acc.)	(pears./spear.)	
BERT	56.5	92.3	**89.5/85.3**	90.7/87.5	84.3	**92.2**	65.0	88.4/88.2	81.8
BERT-NLI	**58.6**	**93.6**	89.4/85.1	90.4/86.8	83.3	89.0	**69.0**	88.3/87.9	**82.1**
SENT-DEBIAS	50.5	89.1	87.5/81.6	87.5/90.7	83.9	91.4	63.2	88.1/87.9	79.4
CONTEXT-DEBIAS	55.2	92.0	85.1/77.5	90.7/87.4	84.6	89.9	57.0	88.4/88.1	79.4
FAIRFIL	55.5	92.4	87.5/80.6	**91.2/88.1**	**84.8**	91.3	63.2	88.4/88.1	80.9
MABEL (ours)	57.8	92.2	**89.5**/85.0	**91.2/88.1**	84.5	91.6	64.3	**89.6/89.2**	82.0

Table 6: Fine-tuning results on the GLUE benchmark. BERT-NLI denotes that we fine-tune pre-trained BERT on NLI data first before fine-tuning on a GLUE task. For the average, we report the Matthew's correlation coefficient for CoLA, the Spearman's rank correlation coefficient for STS-B, and the accuracy for all other tasks.

	StereoSet			CSP	Bias-NLI		
	LM↑	SS◇	ICAT↑	SS◇	NN↑	FN↑	TN:0.5↑
DEFAULT	84.5	56.2	74.0	**50.8**	0.917	**0.983**	**0.983**
SNLI ENT.	84.1	58.9	69.1	51.5	0.885	0.973	0.972
MNLI ENT.	**85.8**	**55.7**	**76.1**	53.8	0.915	0.927	0.971
SNLI NEU.	82.8	58.9	68.3	55.0	**0.935**	0.945	0.945
SNLI CON.	76.9	58.0	64.6	56.5	0.710	0.723	0.722
QQP	76.9	57.9	64.6	53.1	0.917	0.938	0.938
PARA-NMT	79.3	57.8	67.0	53.4	0.756	0.783	0.782
DROPOUT	78.4	57.3	67.0	52.7	0.780	0.809	0.807

Table 7: Data ablation results for MABEL. Positive pair constructions include entailment (Ent.), neutral (Neu.) and contradictory (Con.) pairs, paraphrastic examples from QQP and Para-NMT, and general sentences from the corpora used by Liang et al. (2020). Default: SNLI+MNLI entailment data. Dropout: the same sentence is passed through the encoder twice with standard dropout. CSP: CrowS-Pairs. ◇: the closer to 50, the better.

undesirable behavior, as it suggests that identical concepts along opposite gender directions are now further apart in latent space, and are have become more differentiated in the same contexts.

6.2 Ablations

We perform extensive ablations to show that every component of MABEL benefits the overall system. We use StereoSet, CrowS-Pairs, and Bias-NLI as representative tasks.

Comparing other supervised pairs. Since leveraging entailment examples as positive pairs is conducive to high-quality representation learning (Gao et al., 2021), we believe that this construction is particularly suitable for semantic retention. To justify our choice, we further train on neutral pairs and contradiction pairs from the SNLI dataset. We also consider paraphrase pairs from the Quora Question Pairs (QQP) dataset (Wang et al., 2017) and the Para-NMT dataset (Wieting and Gimpel, 2018). Finally, we try individual unlabeled sentences from the same multi-domain corpora used by Liang et al. (2020) and Cheng et al. (2021). In this setting, standard dropout is applied: positive pairs are constructed by encoding the same sentence twice with different masks, resulting in two minimally different embeddings (Gao et al., 2021).

From Table 7, entailment pairs are a critical data choice for preserving language modeling ability, and result in the highest ICAT scores. Interestingly, the SS is consistent across the board. The MNLI dataset produces the best LM and SS scores, likely as it is a semantically richer dataset with sentences harvested across diverse genres. In contrast, SNLI consists of short, artificial sentences harvested from image captions. The exposure to MNLI's diverse vocabulary set may have helped MABEL learn bet-

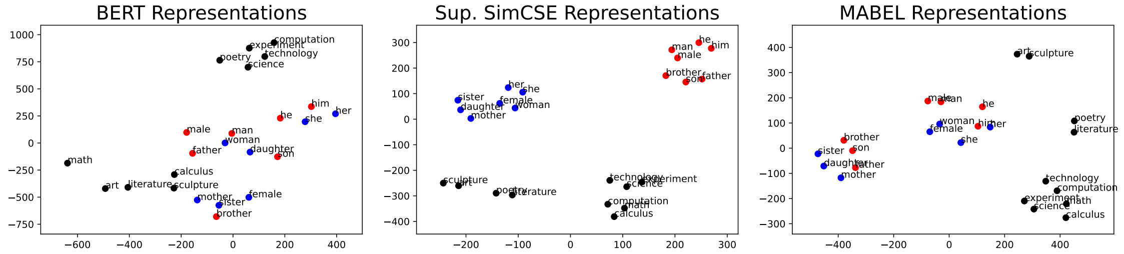

Figure 2: t-SNE plots of sentence representations encoded with BERT, sup. SimCSE, and MABEL. Male-aligned terms (man, male, he, brother, son, father) are in red, female-aligned terms (woman, female, she, her, sister, daughter, mother) are in blue. Neutral terms (e.g., math, art, calculus, poetry, science) are in black.

ter language modeling and greater fairness across a broad range of semantic contexts. The stereotype scores from StereoSet and CrowS-Pairs do not correlate well; for instance, MABEL trained on SNLI entailment pairs shows the worst stereotype score on StereoSet, but among the best on CrowS-Pairs. With only 266 examples, CrowS-Pairs is significantly smaller than StereoSet (which has 2313 examples), and tends to be a more equivocal metric. Entailment pairs, and neutral pairs to a lesser extent, demonstrate the best language retention on the Bias-NLI metric, although the QQP dataset also performs well. One possible explanation is that the QQP paraphrases hold a similar, albeit weaker, directional relationship to NLI pairs.

Disentangling objectives. Results of MABEL trained with ablated losses are in Table 8. Without the MLM objective, the LM score collapses along with the ICAT score. Though the SS score becomes very close to 50, it seems to be more indicative of randomness than model fairness; the off-the-shelf LM head is no longer compatible with the trained encoder. When the contrastive loss is omitted, MABEL's performance on Bias-NLI drops from the 0.9 range to the 0.8 range, showing that it is key to preserving sentence-level knowledge. Removing the alignment loss also leads to a similar decrease in performance on Bias-NLI. As this particular objective does not directly optimize semantic understanding, we attribute this drop to a reduction in fairness knowledge.

Impact of batch size. Table 9 shows the effect of batch size on StereoSet performance. Encouragingly, although contrastive representation learning typically benefits from large batch sizes (Chen et al., 2020), an aggregated batch size of 128 already works well. MABEL is very lightweight and trains in less than 8 hours on a single GPU.

	StereoSet			**CSP**	**Bias-NLI**		
	LM↑	SS◇	ICAT↑	SS◇	NN↑	FN↑	TN:0.5↑
MABEL	84.6	56.2	**74.0**	**50.8**	0.917	**0.983**	**0.982**
$-\mathcal{L}_{\text{MLM}}$	55.8	**51.1**	54.6	44.3	**0.970**	0.976	0.976
$-\mathcal{L}_{\text{CL}}$	84.9	57.2	72.6	54.6	0.858	0.884	0.883
$-\mathcal{L}_{\text{AL}}$	**85.0**	57.3	72.6	54.2	0.878	0.890	0.889

Table 8: Objective ablation results for MABEL. CSP: CrowS-Pairs. ◇: the closer to 50, the better.

Batch Size	**LM ↑**	**SS ◇**	**ICAT ↑**
64	82.43	56.42	71.85
128	84.55	**56.25**	**73.98**
256	**84.62**	57.46	72.00

Table 9: StereoSet results on different cumulative batch sizes. ◇: the closer to 50, the better.

7 Conclusion

In this work, we propose MABEL, a simple bias mitigation technique that harnesses supervised signals from entailment pairs in NLI data to create informative and fair contextualized representations. We systematically compare MABEL and other recent task-agnostic debiasing baselines across a range of intrinsic and extrinsic bias metrics, wherein MABEL demonstrates a markedly better performance-fairness tradeoff. Its capacity for language understanding is also minimally impacted, rendering it suitable for general-purpose use. Additionally, systematic ablations show that both the choice of data and individual objectives are integral to MABEL's good performance. Our contribution and findings are complementary to the bias transfer hypothesis (Jin et al., 2021), which suggests that upstream bias mitigation effects may be transferable to downstream settings. We hope that MABEL can add a new perspective toward creating fairer language models.

Acknowledgements

We thank Tianyu Gao, Sadhika Malladi, Howard Yen, and the other members of the Princeton NLP group for their helpful discussions and support. We are also grateful to the anonymous reviewers for their valuable feedback. This research is partially supported by the Peter and Rosalind Friedland Endowed Senior Thesis Fund from the Princeton School of Engineering & Applied Sciences.

Limitations

Following prior bias mitigation work (Cheng et al., 2021; Liang et al., 2020), our framework relies on a curated list of gender word pairs for counterfactual data augmentation. While we believe that our general list is broad enough to cover the majority of gendered terms in a dataset, this lexicon is nevertheless non-exhaustive and cannot completely remove all bias directions (Ethayarajh et al., 2019). One possible improvement would be to use automatic perturbation augmentation on the entailment pairs (Qian et al., 2022) (concurrent work), a more expansive technique that counterfactually augments data along multiple demographic axes.

Another consideration is that we primarily juxtapose against task-agnostic approaches in our work, even though some task-specific procedures, specifically R-LACE (Ravfogel et al., 2022) and INLP (Ravfogel et al., 2020), show excellent gains in occupation classification, an extrinsic task. Recently, Meade et al. (2022) has successfully adapted INLP to a task-agnostic setting by mining on an unlabeled corpus. We believe that other task-specific methods can be similarly adapted to train task-agnostic encoders, though we leave this comparison to future work. In light of recent findings that bias can re-enter the model during any stage of the training pipeline (Jin et al., 2021; Kaneko et al., 2022), one interesting direction would be to pair MABEL, which is task-agnostic, with task-specific procedures. Essentially, by debiasing at both ends—first upstream in the encoder, then downstream in the classifier—MABEL could potentially achieve a greater reduction in bias across some of the extrinsic benchmarks.

Although MABEL shows exciting performance across an extensive range of evaluation settings, these results should not be construed as a complete erasure of bias. For one, our two main intrinsic metrics, StereoSet and CrowS-Pairs, are skewed towards North American social biases and only reflect positive predictive power. They can detect the presence, not the *absence* of bias (Meade et al., 2022). Aribandi et al. (2021) furthers that these likelihood-based diagnostics can vary wildly across identical model checkpoints trained on different random seeds. Blodgett et al. (2021) points to the unreliability of several benchmarks we use, including StereoSet, CrowS-Pairs, and WinoBias, which inadequately articulate their assumptions of stereotypical behaviors. Additionally, MABEL's gains in fairness are not universally strong—it handles some operationalizations of gender bias more effectively than others. One reason for this inconsistency is that bias metrics have been found to correlate poorly; desirable performance on one bias indicator does not necessarily translate to equivalently significant gains on other evaluation tasks (Goldfarb-Tarrant et al., 2021; Orgad et al., 2022). The lack of clarity and agreement in existing evaluation frameworks is a fundamental challenge in this field.

Ethics Statement

There are several ethical points of consideration to this work. As our contribution is entirely methodological, we rely upon an existing range of well-known datasets and evaluation tasks that assume a binary conceptualization of gender. In particular, the over-simplification of gender identity as a dichotomy, not as a spectrum, means that MABEL does not adequately address the full range of stereotypical biases expressed in real life. We fully acknowledge and support the development of more inclusive methodological tools, datasets, and evaluation mechanisms.

Furthermore, we restrict the definitonal scope of bias in this work to allocational and representational bias (Barocas et al., 2017). *Allocational bias* is the phenomenon in which models perform systematically better for some social groups over others, e.g., a coreference resolution system that successfully identifies male coreferents at a higher rate over female ones. *Representational* bias denotes the spurious associations between social groups and certain words or concepts. An example would be the unintentional linkage of genders with particular occupations, as captured by contextualized word representations.

We neglect other critical types of biases under this framework, in particular intersectional biases. As per Subramanian et al. (2021), most existing debiasing techniques only consider sensitive attributes, e.g., race or gender, in isolation. However, a truly fair model does not and cannot operate in a vacuum, and should be able to handle a complex combination of various biases at once.

Another consideration is that MABEL is entirely English-centric. This assumption is symptomatic of a larger problem, as most gender bias studies are situated in high-resource languages. Given that conceptualizations of gender and language are a function of societal and cultural norms, it is imperative that the tools we create can generalize beyond an English context. For instance, some languages such as Spanish or German contain grammatical gender, meaning that nouns or adjectives can have masculine or feminine forms. The need to account for both linguistic gender and social gender significantly complicates the matter of bias detection and elimination.

For these reasons, practitioners should exercise great caution when applying MABEL to real-world use cases. At its present state, MABEL should not be viewed as a one-size-fits-all solution to gender bias in NLP, but moreso as a preliminary effort to illuminate and attenuate aspects of a crucial, elusive, and multi-faceted problem.

References

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. In *SIGCIS Conference*, Online.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. 2022. Conditional supervised contrastive learning for fair text classification.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of* *the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Ido Dagan and Oren Glickman. 2005. The pascal recognising textual entailment challenge. In *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

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

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

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *TAC*. Citeseer.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021a. Balancing out bias: Achieving fairness through training reweighting. *ArXiv preprint*, abs/2109.08253.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021b. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150. ijcai.org.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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

A Implementation Details of MABEL

We use an aggregation of entailment pairs from the SNLI and MNLI datasets, and augment pairs with opposite gender directions, drawing from the same list of attribute word pairs used by Bolukbasi et al. (2016), Liang et al. (2020), and Cheng et al. (2021): `(man, woman)`, `(boy, girl)`, `(he, she)`, `(father, mother)`, `(son, daughter)`, `(guy, gal)`, `(male, female)`, `(his, her)`, `(himself, herself)`, `(John, Mary)`, alongside plural forms.

We implement MABEL using the HuggingFace Trainer in PyTorch (Paszke et al., 2019) and train for 2 epochs. We take the last-saved checkpoint. Training MABEL takes less than 2 hours across 4 NVIDIA GeForce RTX 3090 GPUs.

Ablation details. Dataset sizes from our ablation study are in Table 10.

Dataset	**Type**	**Original #**	**Final #**
MNLI	Entailment	130.9K	21.5K
SNLI	Entailment	190.1K	112.7K
SNLI	Neutral	189.2K	126.6K
SNLI	Contradiction	189.7K	127.2K
QQP	Paraphrase	149.2K	23.9K
Para-NMT	Paraphrase	5M	1.3M

Table 10: Information about dataset sizes.

Besides batch size, we also tune for learning rate $\in \{1e^{-5}, 3e^{-5}, 5e^{-5}\}$ and $\alpha \in \{0.01, 0.05, 0.1\}$.

As Table 11 indicates, increasing the learning rate improves fairness as the stereotype score approaches 50, but also seems to slightly diminish the model's language modeling ability. Furthermore, a larger α results in a fairer stereotype score, which corroborates our intuition as this parameter adjusts the influence of our alignment loss. Unfortunately, increasing α also monotonically decreases the language modeling score.

Therefore, we take a learning rate of $5e^{-5}$, a batch size of 32, and an $\alpha = 0.05$ as our default hyper-parameters; these result in the best trade-off between fairness and language modeling ability. We use $\lambda = 0.1$ in all the experiments.

B Baseline Implementation

Context-Debias. We use the model checkpoint provided by Kaneko and Bollegala (2021), and treat it as a regular encoder for downstream evaluation.

	LM ↑	SS ◇	ICAT ↑
LR $= 1e^{-5}$	85.13	59.71	68.60
LR $= 3e^{-5}$	85.03	58.29	70.92
LR $= 5e^{-5}$	84.54	**56.73**	**73.98**
$\alpha = 0.01$	**85.29**	59.67	68.80
$\alpha = 0.05$	84.54	**56.73**	**73.98**
$\alpha = 0.1$	83.34	56.94	72.52

Table 11: StereoSet results on different hyper-parameter settings. Unless otherwise stated, the default configuration is a learning rate (LR) of $5e^{-5}$, a batch size of 32, and an α of 0.05. ◇: the closer to 50, the better.

Sent-Debias. We use the code and data provided by Liang et al. (2020) to compute the gender bias subspace. For downstream evaluation, the debiasing step (subtracting the subspace from the representations) is applied directly after encoding.

FairFil. As code for this work is not available, we re-implement FairFil, the main contrastive approach, without the additional information-theoretic regularizer. Note that the reported performance difference from including the regularizer or not (0.150 vs. 0.179 on SEAT) is marginal. We checked all the implementation details carefully and report our reproduced and original SEAT effect size results in Table 12.

SEAT	**FF (O)**	**FF (R)**
Category	**ES ‡**	**ES ‡**
Names, Career/Family 6	0.218	0.279±0.147
Terms, Career/Family 6b	0.086	0.155±0.139
Terms, Math/Arts 7	0.133	0.046±0.008
Names, Math/Arts 7b	0.101	0.061±0.046
Terms, Science/Arts 8	0.218	0.055±0.050
Names, Science/Arts 8b	0.320	0.530±0.092
Avg. Abs. Effect Size	0.179	0.188

Table 12: Absolute average effect sizes (ES) on the 6 gender-associated SEAT categories, for original (O) and reproduced (R) results on FairFil (FF). We report the average and standard deviation for our reproduction. ‡: the closer to 0, the better.

During evaluation, we fix the FairFil layer upon initialization so that its parameters no longer update. We debias by feeding encoded representations through the layer.

C Evaluation Details

C.1 Intrinsic Metrics

StereoSet. StereoSet unifies the language modeling (LM) score and the stereotype score (SS) into a single metric, the Idealized Context Association Test (ICAT) score, which is as follows:

$$\mathrm{ICAT} = \mathrm{LM} \cdot \frac{\min(\mathrm{SS}, 100 - \mathrm{SS})}{50}.$$

In the ideal scenario, a perfectly fair and highly performative language model would have an LM score of 100, an SS score of 50, and thus an ICAT score of 100. Therefore, the higher the ICAT score, the better.

CrowS-Pairs. While CrowS-Pairs originally used pseudo log-likelihood MLM scoring, this form of measurement is found to be error-prone (Meade et al., 2022). Therefore, we follow Meade et al. (2022)'s evaluation approach, and compare the masked token probability of tokens unique to each sentence. The stereotype score (SS) for this task is the percentage of instances for which a language model computes a greater masked token probability to a stereotypical sentence over an anti-stereotypical sentence. An impartial language model without stereotypical biases should score an SS of 50.

C.2 Extrinsic Metrics

Bias-in-Bios. GAP_M^{TPR} is denoted as (observe that the closer the value is to 0, the better)

$$GAP_M^{TPR} = |TPR_M - TPR_F|.$$

Merely taking the difference in overall accuracies does not account for the highly imbalanced nature of the Bias-in-Bios dataset. In line with Ravfogel et al. (2020), we also calculate the root-mean square of $GAP_{M,y}^{TPR}$ to obtain a more robust metric. Taking y as a profession in C, the set of all 28 professions, we can compute

$$GAP_M^{TPR,RMS} = \sqrt{\frac{1}{|C|}\sum_{y\in C}(GAP_{M,y}^{TPR})^2}.$$

Following the suggestion of De-Arteaga et al. (2019a), our train-val-test split of the Bias-in-Bios dataset is 65/25/10. We were able to scrape 206,511 biographies.[12]

In the fine-tuning setting, we train for 5 epochs and evaluate every 1000 steps on the validation set. The model checkpoint is saved if the validation accuracy has improved. We use the `AutoModelForSequenceClassification` class from the `transformers` package (Wolf et al., 2020), which extracts sentence representations by taking the last-layer hidden state of the `[CLS]` token and feeding it through a linear layer with *tanh* activation. We use a batch size of 128, a learning rate of $\lambda = 1e^{-5}$, and a maximum sequence length of 128.

Bias-NLI. The three evaluation metrics used in Bias-NLI task are calculated as follows:

1. **Net Neutral (NN)**: The average probability of the neutral label across all instances.

2. **Fraction Neutral (FN)**: The fraction of sentence pairs accurately labeled as neutral.

3. **Threshold:**τ **(T:**τ**)**: The fraction of instances with the probability of neutral above τ.

In the linear probing setting, we construct an updating linear layer on top of the frozen encoder. Following Dev et al. (2020), sentence representations are extracted from the `[CLS]` token of the last hidden state. We use a batch size of 64, a learning rate of $\lambda = 5e^{-5}$, and a maximum sequence length of 128. We fine-tune for 3 epochs and evaluate every 500 steps, saving the checkpoint if the validation accuracy improves. We randomly sub-sample 10,000 elements from Dev et al. (2020)'s evaluation dataset during inference.

WinoBias. Each WinoBias example contains exactly two mentions of professions and one pronoun, which co-refers correctly to one of the profession (Table 13). Type 1 sentences are syntactically ambiguous and require world knowledge to be correctly resolved, while Type 2 sentences are easier and can be inferred through only syntactic cues. Examples are presented in in Table 13.

Following previous gender bias analyses (Orgad et al., 2022), we borrow the PyTorch re-implementation of the end-to-end c2f-coref model from Xu and Choi (2020). We use the *cased* version of encoders, a significant performance difference exists between cased and uncased variants. Models

[12] `https://github.com/microsoft/biosbias`

	Type 1 Sentence	Type 2 Sentence
Pro-stereotypical	The developer argued with the designer because he did not like the design.	The guard admired the secretary and wanted her job.
Anti-stereotypical	The developer argued with the designer because his design cannot be implemented.	The secretary called the mover and asked her to come.

Table 13: Example of Type I and Type II sentences from the WinoBias dataset (Zhao et al., 2018). The colored text indicates the pronoun and the correct coreferent.

are trained for 24 epochs with a dropout rate of 0.3 and a maximum sequence length of 384. Encoder parameters and task parameters have separate learning rates (1×10^{-5} and 3×10^{-4}), separate linear decay schedules, and separate weight decay rates (1×10^{-2} and 0).

We report the averaged F1-score of three coreference evaluation metrics: MUC, B^3, and CEAF, following Xu and Choi (2020).

C.3 Language Understanding

GLUE. CoLA (Warstadt et al., 2019) and SST-2 (Socher et al., 2013) are single-sentence tasks; MRPC (Dolan and Brockett, 2005) and QQP are paraphrase detection tasks; MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), and RTE (Dagan and Glickman, 2005; Haim et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009) are inference tasks; STS-B (Cer et al., 2017) is a sentence similarity task. We report the accuracy for SST-2, MNLI, QNLI, RTE, and STS-B, and the Matthews correlation coefficient for CoLA. Both the accuracy and the F-1 score are included for MRPC and QQP.

We use the `run_glue.py` script provided by HuggingFace (Wolf et al., 2020), and follow their exact hyper-parameters. For all tasks, we use a batch size of 32, a maximum sequence length of 128, and a learning rate of 2×10^{-5}. We train for 3 epochs for all tasks except for MRPC, which is trained for 5.

D Standard Deviation

As many fairness benchmarks tend to exhibit high variance, we report the standard deviation across 3 runs for each of our implementations from the main results. Standard deviations for intrinsic tasks can be found in Table 14, and for extrinsic tasks in Table 15.

	StereoSet		CrowS-Pairs
Model	LM	SS	SS
CONTEXT-DEBIAS	0.07	0.24	0.77
FAIRFIL	1.47	0.71	3.43
MABEL (ours)	0.39	0.69	0.77

Table 14: Standard deviation on intrinsic tasks from StereoSet and CrowS-Pairs (Table 2).

E SentEval Transfer Evaluation

To more thoroughly evaluate our models' capacity for NLU retention, we additionally test on 9 transfer tasks provided by the SentEval toolkit (Conneau and Kiela, 2018).

While some task overlap exists between SentEval and GLUE, the evaluation setup is different. By freezing the encoder and training a logistic regression classifier, the default SentEval implementation focuses on evaluating the knowledge stored in a fixed-size frozen sentence embedding. Whereas when testing on GLUE, the parameters in the entire encoder are allowed to freely update. GLUE also emphasizes high-resource downstream tasks, which use training data with hundreds of thousands of samples. Comparatively, SentEval mostly focuses on low-resource transfer, with smaller downstream classification tasks such as Movie Review (MR) or Product Review (CR) (Conneau and Kiela, 2018).

We use the evaluation toolkit provided by Conneau and Kiela (2018), following the standard settings, and form sentence representations by extracting the `[CLS]` token of the last hidden state. 10-fold cross-validation was used for MR, CR, SUBJ, MPQA, and SST-2, and cross-validation for TREC. For MRPC, a 2-class classifier learns to predict the probability distribution of relatedness scores; the Pearson correlation coefficient is reported.

Table 16 shows the performance across the downstream SentEval transfer tasks. As this evaluation

	Bias-in-Bios					Bias-NLI				Coreference Resolution						
Model	Acc. (All)	Acc. (M)	Acc. (F)	TPR GAP	TPR RMS	TN	FN	T:0.5	T:0.7	ON	1A	1P	2A	2P	TPR-1	TPR-2
BERT	0.56	0.56	0.58	0.14	0.01	0.03	0.05	0.05	0.05	0.05	1.12	0.57	1.02	1.15	1.17	1.21
SENT-DEBIAS	0.09	0.10	0.20	0.25	0.00	0.05	0.04	0.05	0.09	0.15	2.88	0.27	1.47	0.29	3.02	1.53
CONTEXT-DEBIAS	0.36	0.32	0.42	0.17	0.01	0.03	0.05	0.05	0.04	0.19	1.00	1.25	1.13	1.33	0.75	0.74
FAIRFIL	0.07	0.25	0.16	0.41	0.00	0.05	0.07	0.01	0.05	0.64	0.41	0.47	2.31	1.49	0.16	0.84
MABEL (ours)	0.40	0.51	0.47	0.04	0.00	0.02	0.01	0.01	0.03	0.37	1.12	1.11	0.30	0.41	1.88	0.44

Table 15: Standard deviation on Bias-in-Bios (Table 3), Bias-NLI (Table 4), and coreference resolution (OntoNotes and WinoBias) (Table 5).

regime does not involve fine-tuning, we notice less uniformity across the baselines' results. While MABEL and MABEL w/o MLM have a higher average performance than BERT, the only task with an obvious gain is in MRPC (68.87% and 71.48% vs. 65.57%). This makes sense as MRPC is a semantic similarity task, which leveraging supervised signals from NLI entailment pairs happens to benefit (Gao et al., 2021).

F Bias Benchmarks in Other Works

Table 17 shows a comprehensive compilation of gender bias metrics used by other bias mitigation methods in recent literature.

G SEAT Evaluation

The Sentence Encoder Association Test (SEAT) (May et al., 2019) is the sentence-level extension of the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), a hypothesis-driven diagnostic that checks whether two sets of target words (for instance, `[artist, musician,...]` and `[scientist, engineer...]`) are equally similar to two sets of attribute words (for instance, `[man, father,...]` and `[woman, mother,...]`). In SEAT, the concept words from WEAT are inserted into semantically bleached sentence templates such as "This is a[n] <word>," effectively allowing for the comparison of *sentence* representations.

For both WEAT and SEAT, the null hypothesis postulates that no difference exists in the relative similarity between the sets of target words X, Y and the sets of attribute words A, B. The effect size, $s(X, Y, A, B)$, quantifies the difference in mean cosine similarity between representations of the target concept pair (X, Y), and representations of the attribute concept pair (A, B), through the following equations:

$$s(w, A, B) = \underset{a \in A}{\text{mean}} \cos(\overrightarrow{w}, \overrightarrow{a}) - \underset{b \in B}{\text{mean}} \cos(\overrightarrow{w}, \overrightarrow{b})$$

$$s(X, Y, A, B) = \textstyle\sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

We report our results (alongside pre-trained BERT's) in Table 18. Following Liang et al. (2020), we extract the sentence representation as the `[CLS]` token fed through a linear layer and `tanh` activation (e.g., the pooled output).

H A Comparison to SimCSE

A non-debiasing analogue to MABEL is supervised SimCSE (Gao et al., 2021), a state-of-the-art representation learning approach that generates sentence representations with good semantic textual similarity (STS) performance. Like MABEL, supervised SimCSE also trains on entailment pairs from NLI data using an contrastive learning objective.

One potential concern is that MABEL performs well on tasks such as Bias-NLI not due to greater fairness, but because it has already been trained on NLI data. To test this assumption, we repeat the Bias-NLI task on SimCSE, which has been trained on un-augmented entailment pairs from the SNLI dataset. In Table 19, the tangible increase across all metrics from SimCSE to MABEL indicates that MABEL's good performance cannot be solely attributed to NLI knowledge retention.

I Linear Probing Experiments

To better illustrate the effects of MABEL, we conduct a suite of probing experiments, in which the entire *encoder* parameters are frozen during training. We summarize task details and results below.

I.1 Bias-in-Bios

We follow the same procedure as in the fine-tuning setting, except freeze the encoder and only update the linear classification layer. From Table 20, both

Model	MR ↑	CR ↑	SUBJ ↑	MPQA ↑	SST-2 ↑	TREC ↑	MRPC ↑	Avg. ↑
BERT	**80.99**	85.67	**95.31**	87.40	**86.99**	84.20	65.57	83.73
Context-Debias	78.37	85.22	94.11	85.99	84.73	**85.60**	66.55	82.94
Sent-Debias	69.85	68.96	89.12	80.78	81.44	60.00	70.14	74.33
FairFil	76.94	80.34	92.82	83.54	81.88	79.20	69.28	80.57
MABEL	78.33	85.83	93.78	**89.13**	85.50	85.20	68.87	83.81
MABEL w/o MLM	80.01	**86.41**	94.50	89.29	85.45	84.80	**71.48**	**84.56**

Table 16: Downstream transfer task results for BERT, MABEL models, and bias baselines from the SentEval benchmark (Conneau et al., 2017).

Method	WEAT/ SEAT	CrowS- Pairs	Stereo- Set	Bias- in-Bios	Bias- NLI	Wino- Bias	Other int.	Other ext.
Task-specific approaches								
INLP (Ravfogel et al., 2020)				✓*				
R-LACE (Ravfogel et al., 2022)				✓*				
Con (Shen et al., 2021)				✓*				
DADV (Han et al., 2021b)				✓*				
GATE (Han et al., 2021a)				✓*				
Task-agnostic approaches								
CDA (Webster et al., 2020)				✓			DisCo	STS-B, WinoGender
Dropout (Webster et al., 2020)				✓			DisCo	STS-B, WinoGender
ADELE (Lauscher et al., 2021)	✓				✓		BEC-Pro, DisCo	STS-B
Bias Projection (Dev et al., 2020)					✓			
OSCaR (Dev et al., 2021)	✓				✓		ECT	SIRT
Sent-Debias (Liang et al., 2020)	✓							
Context-Debias (Kaneko and Bollegala, 2021)	✓				✓			
Auto-Debias (Guo et al., 2022)	✓	✓*						
FairFil (Cheng et al., 2021)	✓							
MABEL (ours)		✓*	✓*	✓*	✓*	✓*		

Table 17: Gender bias metrics used for each baseline, as reported from the original work. A * means that the metrics are directly comparable to those in our main results. DisCo (Webster et al., 2020) is a template-based likelihood metric; STS-B is a semantic similarity task adapted by Webster et al. (2020) for measuring gender bias; WinoGender (Rudinger et al., 2018) is a small-scale coreference resolution dataset that links pronouns and occupations; BEC-Pro (Bartl et al., 2020) is a template-based metric that measures the influence of an occupation word on a gender word; ECT (Dev and Phillips, 2019) applies the Spearman's correlation coefficient to calculate the association between gender words and attribute-neutral words, SIRT (Dev et al., 2021) uses NLI data to evaluate for gendered information retention.

MABEL and MABEL without the MLM loss show better overall and gender-specific task accuracy on the probe, in comparison to BERT and other bias mitigation baselines. MABEL has the lowest TPR-GAP and the second lowest TPR-RMS. While INLP has a very low TPR-RMS of 0.069, it also has significantly worse accuracy, whereas MABEL achieves a better fairness-accuracy balance.

I.2 Bias-NLI

Dev et al. (2020) originally formulated this task as a probing experiment. Accordingly, we only update a linear layer on top of a frozen encoder when training on the SNLI dataset. We freeze the entire model when evaluating on the test dataset. Our results are shown in Table 21.

Figure 3 shows the validation accuracy of BERT,

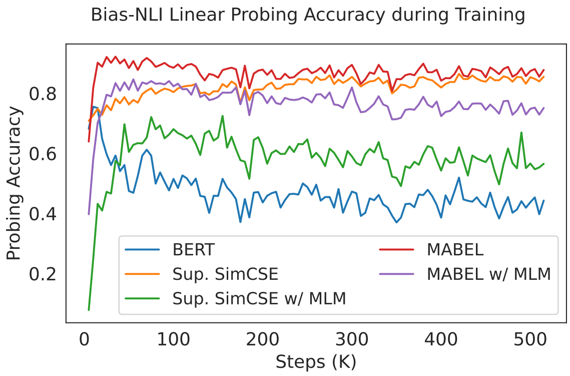

Figure 3: Linear probing accuracy on Bias-NLI of model checkpoints at various timesteps when training on SNLI.

SEAT Category	BERT ES ‡	MABEL ES ‡
Names, Career/Family 6	0.477	0.664±0.313
Terms, Career/Family 6b	0.108	0.167±0.196
Terms, Math/Arts 7	0.253	0.479±0.488
Names, Math/Arts 7b	0.254	0.647±0.254
Terms, Science/Arts 8	0.399	0.465±0.288
Names, Science/Arts 8b	0.636	0.570±0.296
Avg. Abs. Effect Size	0.354	0.499±0.090

Table 18: BERT's and MABEL effect sizes (ES) on the gender-associated SEAT categories. For MABEL, we report the absolute average and standard deviation across 3 runs. ‡: the closer to 0, the better.

Model	TN↑	FN↑	T:0.5 ↑	T:0.7 ↑
SUP. SIMCSE	0.830	0.951	0.945	0.839
MABEL (ours)	**0.917**	**0.983**	**0.983**	**0.968**

Table 19: Results on Bias-NLI for supervised SimCSE (Gao et al., 2021) and MABEL.

Model	Acc. (All) ↑	Acc. (M) ↑	Acc. (F) ↑	TPR GAP ↓	TPR RMS ↓
		Bias-in-Bios - Linear Probe			
BERT	79.63	80.27	78.84	1.436	0.200
CONTEXT-DEBIAS	78.27	78.98	77.39	1.595	0.214
SENT-DEBIAS	75.55	76.19	74.74	1.452	0.195
FAIRFIL	74.69	75.30	73.94	1.357	0.225
INLP	72.36	73.36	71.10	2.261	**0.069**
MABEL W/O MLM	80.68	81.22	80.00	1.220	0.172
MABEL	**80.98**	**81.43**	**80.41**	**1.012**	0.159

Table 20: Linear probing results on Bias-in-Bios across the MABEL models and different baselines.

Model	TN ↑	FN ↑	T:0.5 ↑
Bias-NLI - Linear Probe			
BERT⋆	0.409	0.512	0.239
SUP. SIMCSE	0.502	0.792	0.516
SUP. SIMCSE + MLM	0.412	0.551	0.264
BIAS PROJECTION⋆ - Test	0.396	0.371	0.341
BIAS PROJECTION⋆- Train + Test	0.516	0.526	0.501
OSCAR⋆†	0.566	0.588	-
SENT-DEBIAS	0.351	0.319	0.020
CONTEXT-DEBIAS	0.240	0.078	0.023
FAIRFIL	0.348	0.318	0.055
MABEL W/O MLM	**0.571**	**0.853**	**0.710**
MABEL	0.538	0.837	0.653

Table 21: Natural language inference results for pre-trained BERT, the baselines, and MABEL. Best numbers in **bold**. W/O = without; ⋆: results are from original papers; †: the encoder model is RoBERTa$_{\text{base}}$; ‡: the models are fine-tuned on MNLI. BERT and BIAS PROJECTION results are from Dev et al. (2020); OSCAR is from Dev et al. (2021).

MABEL models, and supervised SimCSE models on the NLI-Bias evaluation set at various timesteps when training on the SNLI dataset. BERT consistently struggles with the task—its accuracy starts off high before degrading noticeably. MABEL outperforms SimCSE both with and without the MLM loss, which shows that its performance on Bias-NLI is not entirely due to enhanced semantic understanding, but greater fairness as well. The MLM objective steadily drops the performance of both MABEL and SimCSE, which shows that it harms sentence-level knowledge retention. Interestingly, the opposite trend holds true in the fine-tuning setting—including the MLM loss leads to *better* NLI performance across all three metrics.

J Alignment Objectives

Beside our default alignment loss (**Alignment Loss 1**), we experiment with other losses that maximize the similarity between original and gender-augmented representations. We describe them and report their results across Bias-in-Bios and Bias-NLI tasks.

Alignment Loss 2 is a contrastive objective that is similar to FairFil, but takes cosine similarity as a scoring function instead of a two-layer fully-connected neural network. Augmented sentence pairs, either (p, p') (or (h, h')), form positive pairs, and other sentences form in-batch negatives. Let x_i be any original premise or hypothesis representation, and x' be the augmented counterpart of x:

$$\mathcal{L}_{AL2} = -\log \frac{e^{\text{sim}(x_i, x_i')/\tau}}{\sum_{j=1}^{2n} e^{\text{sim}(x_i, x_j)/\tau}}.$$

Alignment Loss 3 tries to maximize the cosine similarities between original and augmented sentences. Given the pairs (p_i, p_i') and (h_i, h_i'):

$$\mathcal{L}_{AL3} = \frac{1}{n} \sum_{i=1}^{n} -(\text{sim}(p_i, p_i') - \text{sim}(h_i, h_i')).$$

The results for MABEL trained with each alignment loss are in Table 22. Our default alignment loss (AL1) returns consistently better results.

Metric	**BERT**	**AL1**	**AL2**	**AL3**
		Bias-in-Bios - Linear Probe		
Overall Acc. ↑	79.63	**80.68**	79.94	75.30
Acc. (M) ↑	80.27	**81.22**	80.63	77.32
Acc. (F) ↑	78.83	**80.00**	79.08	72.77
TPR GAP ↓	1.436	**1.220**	1.550	4.543
TPR RMS ↓	0.200	**0.172**	0.180	0.234
		Bias-NLI - Linear Probe		
NN ↑	0.409	**0.571**	0.509	0.354
FN ↑	0.512	**0.853**	0.737	0.318
T:0.5 ↑	0.239	**0.710**	0.538	0.069
		Bias-NLI - Fine-tuning		
NN ↑	0.762	**0.917**	0.897	0.844
FN ↑	0.900	**0.983**	0.948	0.927
T:0.5 ↑	0.718	**0.983**	0.949	0.920

Table 22: Results on Bias-in-Bios (linear probe setting) and Bias-NLI (linear probe setting and fine-tuning setting), for MABEL trained with different alignment losses—AL1 (default), AL2, and AL3. All models are trained only on SNLI data in this ablation study.