Measuring Social Biases in Masked Language Models by Proxy of Prediction Quality

Anonymous ACL submission

Abstract

Innovative transformer-based language models produce contextually-aware token embeddings and have achieved state-of-the-art performance for a variety of natural language tasks, but have been shown to encode unwanted biases for downstream applications. In this paper, we evaluate the social biases encoded by transformers trained with the masked language modeling objective using proposed proxy functions within an iterative masking experiment to measure the quality of transformer models' predictions, and assess the preference of MLMs towards disadvantaged and advantaged groups. We compare bias estimations with those produced by other evaluation methods using benchmark datasets, finding relatively high religious and disability biases across considered MLMs and low gender bias in one dataset relative to another. We find that our measures outperform others in their agreement with human annotators. We extend on previous work by evaluating social biases introduced after re-training an MLM under the masked language modeling objective (w.r.t. the model's pre-trained base), and find that proposed measures produce more accurate estimations of social biases introduced by re-training MLMs based on relative preference for biased sentences between transformers.

Warning: *This paper contains explicit statements of biased stereotypes and may be upsetting.*

1 Introduction

Word embeddings have proven useful in a variety of Natural Language Processing (NLP) tasks due to their ability to efficiently model complex semantic and syntactic word relationships. Token-level embeddings, such as those produced by Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) algorithms, learn a non-contextualized text representation and produce static word embeddings that can uncover linear semantic or syntactic relationships between tokens.

Masked language models (MLM(s); Brown et al., 2020; Devlin et al., 2019; Liu et al., 2019a; Radford et al., 2019) such as transformers BERT and RoBERTa incorporate bidirectionality and self-attention, producing contextually-aware embeddings. Unlike BERT_{unc}, RoBERTa (Robustly Optimized BERT) was pre-trained solely on the masked language modeling objective on a larger corpus of text. RoBERTa uses a dynamic masking strategy for a diverse set of representations during training and has achieved state-of-the-art results on GLUE, RACE and SQuAD (Liu et al., 2019b; Rajpurkar et al., 2016; Lai et al., 2017; Wang et al., 2018). Distilled model variants have been shown to train significantly faster for minor decreases in performance (Sanh et al., 2019). MLMs have produced state-of-the-art results for masked language modeling, named entity recognition and intent or topic classification, but also encode concerning social biases against disadvantaged groups that are undesirable in production settings. As MLMs become increasingly prevalent, researchers have been working on methods to measure biases embedded in these models (Nangia et al., 2020; Kaneko and Bollegala, 2022; Salutari et al., 2023).

In this work, we leverage pre-trained base transformer models available through the Transformers library (Wolf et al., 2020) to evaluate social biases of MLMs using proposed proxy measures for MLM prediction quality to assess the relative preferences of MLMs towards sentences biased against disadvantaged and advantaged groups. In particular, **we quantify mlm preference by proxy of relative masked token prediction quality given unmasked token context between encoded sentences within pairs**, where one sentence contains bias against disadvantaged groups (stereotypical) and the other contains bias against advantaged groups (anti-stereotypical). Proposed measures ΔP, CRR, CRRA and ΔPA (13, 14, 4 and 5) are computed under the iterative masking experiment

detailed in Section 3.2. Additionally, **we compute bias scores for benchmark datasets using functions 7 and 8 to measure transformer preference for biased sentences, along with relative preference for biased sentences between transformers**.

Our proposed methodology discussed in Section 4 could help others evaluate social biases encoded in an MLM after it is re-trained on the masked language modeling objective, such as for a downstream fill-mask task. We find that proposed measures outperform others in accuracy measured by ground truth human annotations, and for estimating biases introduced by re-training MLMs on sentence sets in a benchmark dataset.

2 Related Work

2.1 Biases in Non-contextual Word Embeddings

Non-contextual word (token) embeddings (Pennington et al., 2014; Mikolov et al., 2013), can be shifted in a direction to decompose bias embedded in learned text data representations. These could be analogies or biases along an axis, such as gender (e.g. doctor - man + woman = nurse) as used in Bolukbasi et al., 2016 or race as used in Manzini et al., 2019. The Word Embedding Association Test (WEAT; Caliskan et al., 2017) measures association between targets and attributes using cosine similarity between static word embeddings, but has been shown to overestimate biases by Ethayarajh et al., 2019 which proposed the more robust relational inner product association (RIPA) method, derived from the subspace projection method to debias vectors in Bolukbasi et al., 2016.

While WEAT has shown that token-level embeddings produced by GloVe and Word2Vec encode biases based on gender and race (Caliskan et al., 2017), the Sentence Encoder Association Test (SEAT; May et al., 2019) extends on WEAT to measure social biases in sentence-level encoders such as ELMo and BERT using template sentences with masked target tokens (e.g. "This is <mask>"), averaging token embeddings to form sentence-level embeddings on which cosine similarity is applied as a proxy for semantic association. As an alternative evaluation method under a different objective, Liang et al., 2020 assessed differences in log-likelihood between gender pronouns in a template sentence (e.g. "<mask> is a/an [occupation]") where occupations can uncover the directionality of the bias encoded by an MLM.

2.2 Evaluating Biases in MLMs

To measure the bias in masked language models, Salutari et al., 2023 tests MLMs in an iterative fill mask setting, where a model encodes or tokenizes input text to predict a masked token for combinations of token-masked text. In this setting, the model provides a set of tokens to fill the masked one, starting with the token of highest likelihood (and first rank) in the set of token predictions of the MLM, which is limited according to the MLM's embedding space. Salutari et al., 2023 proposes the Complementary Reciprocal Rank (CRR) for a masked token given its context, and uses the measure for a sentence (average of all single masked token CRRs with token ordering preserved) as a proxy for an MLM's prediction quality (preference).

Kaneko and Bollegala, 2022 reference the use of sentence-level embeddings produced by MLMs for downstream tasks such as sentiment classification to argue that biases associated with masked tokens should not influence the intrinsic bias evaluation of an MLM, as opposed to the evaluation of biases introduced after an MLM is fine-tuned. They propose evaluation metrics All Unmasked Likelihood (AUL) and AUL with Attention weights (AULA), where AUL and AULA are generated by requiring the MLM to predict all tokens (unmasked input) to eliminate biases associated with masked tokens under previously proposed pseudo-likelihood-based scoring methods (Nadeem et al., 2021, Nangia et al., 2020), which assumed that masked tokens are statistically independent. In contrast with Kaneko and Bollegala, 2022, we focus on an MLM's key pre-training objective, masked language modeling, to measure social biases of the MLM.[1] In addition, we measure relative changes in biases w.r.t the intrinsic biases of the same base MLM after re-training under the masked language modeling objective (as opposed to fine-tuning) in Section 4.2. Thus, we argue that biases associated with masked tokens are not undesirable in our case.

3 Experiments

3.1 Benchmark Datasets for Social Bias

The Crowdsourced Stereotype Pairs Benchmark (CPS; Nangia et al., 2020) contains biased sen-

[1] Masked language modeling was a pre-training objective for all transformers considered in this work. Next sentence prediction was not used for RoBERTa and similar variants due to relatively lower performance with its inclusion (Liu et al., 2019b).

tences towards historically advantaged and disadvantaged groups along nine forms of social biases. StereoSet (SS; Nadeem et al., 2021) contains intrasentence and intersentence (with context) pairs for four forms of social biases, using the likelihood of modified tokens given unmodified token contexts as proxy for an MLM's preference. Similarly, CPS contains characteristic words that distinguish sentences within pairs and define the nature of a particular bias as against advantaged or disadvantaged groups, but instead uses the relative likelihood of unmodified tokens being chosen by the MLM given a modified context (characteristic word) across a sentence pair.

The CPS and SS datasets contain biased sentences towards advantaged and disadvantaged groups, where CPS sentence pairs are categorized by bias types: race, age, socioeconomic, disability, religion, physical appearance, gender, sexual orientation and nationality, and SS sentence pairs are categorized by bias types: race, religion, gender and profession. To probe for biases of interest that are encoded in considered MLMs, the scope of our experiments include all bias categories and sentences pairs in CPS and intrasentence pairs in SS, since intersentence pairs are not masked for bias evaluation.[2] We attempt to estimate the preference of an MLM towards a stereotypical sentence over a less stereotypical one for each bias category in CPS and SS and report corresponding results.

3.1.1 Re-training Dataset

CPS provides a more diverse alternative to biases expressed by sentence pairs in SS. Biases widely acknowledged in the United States[3] are well represented in CPS compared to SS, and there is greater diversity of sentence structures in CPS (Nangia et al., 2020). CPS has been found to be a more reliable benchmark for pre-trained MLM bias measurement than SS, and the validation rate of CPS is 18% higher than SS. (Nangia et al., 2020). Given these findings, paired with the computational expense and time-consumption involved with re-training MLMs under the masked language modeling objective, and concerns regarding standard masked language modeling metric viability on SS, we proceed to use sentence sets in CPS to re-train MLMs and validate methods for estimating biases introduced by that re-training (Nangia et al., 2020). In particular, we re-train MLMs on $\forall s \in S^{\text{dis}}$ or $\forall s \in S^{\text{adv}}$, where s is a sentence biased towards either advantaged (S^{adv}) or disadvantaged groups (S^{dis}), and compare the accuracy of our proposed measures with others for measuring social biases introduced by re-training MLMs.

3.2 MLM Preference by Prediction Quality

When considering a sentence s containing tokens $\{t_1, t_2, ..., t_{l_s}\}$, where l_s is the number of tokens in s, (modified) token(s) of s can characterize its bias towards either disadvantaged or advantaged groups.[4] For a given sentence s with tokens $t \in s$ we denote all tokens besides t_x as $s_{\setminus t_x}$ (where $1 \leq x \leq l_s$). For a given sentence s, we denote modified tokens as M and unmodified tokens as U ($s = U \cup M$). For a given MLM with parameters θ, we denote a masked token as t_m and a predicted token as t_p.

Salazar et al., 2020 used pseudo log-likelihood scoring to approximate $P(U|M, \theta)$, or the probability of unmodified tokens conditioned on modified ones. Similarly, Nangia et al., 2020 reports CrowS-Pairs Scores (CSPS; Appendix B), a pseudo-log-likelihood score for an MLM selecting unmodified tokens given modified ones. Nadeem et al., 2021 reports a StereoSet Score (SSS; Appendix C), a pseudo-log-likelihood score for an MLM selecting modified tokens given unmodified ones. To measure the bias in MLMs, Salutari et al., 2023 tests them in an iterative fill mask setting where the model outputs a set of tokens (or the (log)softmax of model logits mapped to tokens) to fill the masked one, starting with the token of highest probability $P(t_p|c)$ and, as such, first rank $\rho(t_p|c) = 1$ in the set of token predictions provided by the model. $\Delta\text{P}(t|s_{\setminus t_m}; \theta)$ (Appendix D) is a metric for measuring the fairness (or bias) of an MLM and represents the difference between the probability of a predicted token t_p and a masked token t_m in a sentence s. It serves as a proxy of the MLM's prediction quality for a token given its context within an iterative masking experiment, or all tokens in s besides t_m (Salutari et al., 2023). $\text{CRR}(t|s_{\setminus t_m}; \theta)$ (Ap-

[2] Our experiments on the SS dataset include only intrasentence pairs as in experiment code used by Kaneko and Bollegala, 2022.

[3] CPS categories are a "narrowed" version of the US Equal Employment Opportunities Commission's list of protected categories (Nangia et al., 2020).

[4] For example, in the sentence pair $s_1 \in S^{dis}$ and $s_2 \in S^{adv}$ from the CPS dataset, where s_1 is "The poor are really ignorant about how to handle the money they do have" and s_2 is "The rich are really ignorant about how to handle the money they do have", the bias is characterized by modified tokens "poor" and "rich" respectively. All other tokens between s_1 and s_2 in this example are unmodified.

pendix E) is another metric for measuring bias of an MLM for a sentence s, where $\rho(t_p|s_{\backslash t_m})^{-1}$ is the reciprocal rank of the predicted token (and always equal to 1) and $\rho(t_m|s_{\backslash t_m})^{-1}$ is the reciprocal rank of the masked token. Thus, $\rho(t_m|s_{\backslash t_m})^{-1}$ provides a likelihood measure for t_m being chosen by the model as a candidate token to replace the ground truth (masked) one. Salutari et al., 2023 defines $\Delta\text{P}(s)$ (Appendix F) as the probability difference for a sentence (or utterance) s and $\text{CRR}(s)$ (Appendix G) as the complementary reciprocal rank for a sentence s.

Salutari et al., 2023 claims that metrics based on CRR for a sentence s are necessary to fully capture the biases embedded in MLMs, in contrast with the evaluation metrics AUL (Appendix H) and AULA (Appendix I) proposed by Kaneko and Bollegala, 2022. AUL and AULA are generated by predicting all of the tokens in a given unmasked input sentence s as referenced in Section 2. By requiring the MLM to simultaneously predict all of the unmasked tokens in a sentence, Kaneko and Bollegala, 2022 attempts to avoid any selectional biases from masking a subset of input tokens, such as high frequency words (which are masked more often during training). AUL and AULA were found to be sensitive to contextually meaningful inputs, and the CRR measure is also conditional to the unmasked token context by definition.[5] We argue measure sensitivity to unmasked token contexts is desirable when evaluating a given MLM's preference under a fill-mask task, or when estimating biases using contextualized token-level embeddings produced by an MLM. In contrast with previous methods such as SSS and CSPS, we refrain from using strictly modified or unmodified subsets of input tokens as context, and instead provide all tokens but the ground truth one as context for MLM prediction under each iteration of our masking experiment.

Existing benchmark datasets such as CPS are limited to one ground truth per masked token. Thus, an important consideration is an MLM's ability to predict multiple plausible tokens for a context that could qualify for concerning social bias but goes unrecognized during evaluation using previously proposed measures. However, the CRR measure could perform better than pseudo-(log)likelihood-based measures for this sensitivity that yields larger relative differences in $\Delta\text{P}(t|c)$ as opposed to $\text{CRR}(t|c)$,[6], and CRR is deemed critical for evaluation by Salutari et al., 2023 due to the possible uniformity of probabilities generated by a particular MLM with respect to others.

4 Methodology

We report and compare results from the following transformers available through the HuggingFace library (Wolf et al., 2020): BERT_{unc} (**bert-base-uncased**; Devlin et al., 2019), RoBERTa (**roberta-base**; Liu et al., 2019b), $\text{distilBERT}_{\text{unc}}$ (**distilbert-base-uncased**; Sanh et al., 2019) and distilRoBERTa (**distilroberta-base**; Liu et al., 2019a). For all results and transformers T, T_R denotes re-trained while T_P denotes pre-trained (out of box). The subscript $_{\text{unc}}$ denotes an uncased model.

We represent MLM bias based on the model's relative preference for ground truth tokens between two paired sentences along a social bias axis, as measured by the (attention-weighted) quality (or error) of its predictions under the masked language modeling objective. We propose and validate proxy functions that measure the likelihood an MLM will select a ground truth token to replace a masked one, such as attention-weighted functions 4 and 5. We propose function 7 to estimate the encoded social biases in pre-trained MLMs and 8 to estimate the relative change in biases after re-training an MLM.

4.1 Recovering Social Biases in Pre-trained MLMs

We probe for MLM preferences using an iterative masking training procedure, which masks one token at a time until all tokens have been masked once. Thus, we have n logits or predictions for a sentence with n tokens. Special start and end character tokens for MLMs are not included in the span of tokens considered in our experiments to eliminate noise.[7]

The Table in Appendix L shows an example of the iterative fill mask experiment for one model

[5] Kaneko and Bollegala, 2022 illustrates the sensitivity of AUL and AULA to contextually meaningful inputs by randomly shuffling tokens in input sentences and comparing accuracies with and without shuffle.

[6] The lowest rank of any possible ground truth token within a model's vocabulary is equal to the size of the vocabulary ($1 \leq \rho(t_m|c) \leq V$ and $0 \leq \text{CRR}(s) \leq 1 - (V)^{-1}$ for an MLM with embedding vocabulary of size V).

[7] Special start and end character tokens for MLMs are not considered by measures using attention weights and probabilities computed from the (log)softmax of model logits for the masked token index.

and text example, where the language model encodes (tokenizes) text t and predicts for a masked token w_i where i is the original token index. Tables reported in Appendix K show masked token predictions (those with first rank and highest probability) produced by MLMs for example input contexts. They contain the ground truth masked token rank, or $\rho(t_m|c)$, to illustrate the quality of the models' predictions relative to a biased sentence.

$$\begin{aligned}\Delta\text{P}(t|s_{\backslash t_m};\theta) &= \big(\log P(t_p|s_{\backslash t_m};\theta) \\ &\quad - \log P(t_m|s_{\backslash t_m};\theta)\big) \\ &= \Delta\text{P}(w;\theta)\end{aligned} \tag{1}$$

We redefine the Equation shown in Appendix D and propose a modified version $\Delta\text{P}(w)$ as shown in Equation 1.

$$\begin{aligned}\text{CRRA}(t|s_{\backslash t_m};\theta) &= a_m\big(1 - \log\rho(t_m|s_{\backslash t_m};\theta)^{-1}\big) \\ &= \text{CRRA}(w;\theta)\end{aligned} \tag{2}$$

$$\begin{aligned}\Delta\text{PA}(t|s_{\backslash t_m};\theta) &= a_m\big(\log P(t_p|s_{\backslash t_m};\theta) \\ &\quad - \log P(t_m|s_{\backslash t_m};\theta)\big) \\ &= \Delta\text{PA}(w;\theta)\end{aligned} \tag{3}$$

For a given MLM with parameters θ, we propose attention-weighted measures $\text{CRRA}(w)$ and $\Delta\text{PA}(w)$ defined in Equations 2 and 3 respectively, where a_w is the average of all multi-head attentions associated with the ground truth token w, and $P(t_m|s_{\backslash t_m})$ and $\rho(t_m|s_{\backslash t_m})$ are the probability score and rank of the masked token respectively. We extend these definitions for a sentence s as shown in Equations 4 and 5.

$$\text{CRRA}(s) := \frac{1}{l_s}\sum_{w\in s}\text{CRRA}(w;\theta) \tag{4}$$

$$\Delta\text{PA}(s) := \frac{1}{l_s}\sum_{w\in s}\Delta\text{PA}(w;\theta) \tag{5}$$

We compute measure $f, \forall f \in \{\text{CRR}_T(s), \Delta\text{P}_T(s), \text{CRRA}_T(s), \Delta\text{PA}_T(s)\}$, where T is an MLM transformer and s is a sentence, $\forall s \in S^{\text{dis}}$ and $\forall s \in S^{\text{adv}}$.[8] The Table in Appendix Q shows these measures (likelihood scores) for an example sentence s in the SS and CPS datasets.[9]

We define sets of measures M_1 and M_2 ($M_1 \cap M_2 = \emptyset$), where $M_1 = \{\Delta\text{P}, \text{CRR}, \text{CRRA}, \Delta\text{PA}\}$ and $M_2 = \{\text{AUL}, \text{AULA}, \text{CSPS}, \text{SSS}\}$.

$$\Delta f_T(i) = \begin{cases} f_T(S_i^{\text{adv}}) - f_T(S_i^{\text{dis}}), & \text{if } f \in M_1 \\ f_T(S_i^{\text{dis}}) - f_T(S_i^{\text{adv}}), & \text{if } f \in M_2 \end{cases} \tag{6}$$

We apply Equation 6 to estimate the preference of a transformer T for $s \in S^{\text{dis}}$ relative to $s \in S^{\text{adv}}$ for paired sentence s and measure $f, \forall f \in \{\text{CRR}, \text{CRRA}, \Delta\text{P}, \Delta\text{PA}, \text{CSPS}, \text{SSS}, \text{AUL}, \text{AULA}\}$. We define a **bias score for a pre-trained MLM as BSPT** in Equation 7.

$$\text{BSPT}(f) := \frac{100}{N}\sum_{i=1}^{N}\mathbb{1}\big(\Delta f_T(i) > 0\big) \tag{7}$$

$\mathbb{1}$ is a per-model indicator function which returns 1 if transformer T has a larger preference for S^{dis} relative to S^{adv} and 0 otherwise, as estimated for a measure f by Equation 6. BSPT represents the proportion of sentences with higher relative bias against disadvantaged groups for a given measure, where values greater than 50 indicate greater relative bias against disadvantaged groups for an MLM. We apply BSPT to compare preferences for a given MLM, and report results for all considered measures and MLMs in the Tables in Appendix N, where we also provide a detailed analysis of results for the CPS and SS Datasets and race bias in particular. We report overall bias scores in Table 1.

CPS Dataset							
Model	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
RoBERTa$_P$	59.35	58.75	58.09	58.89	60.68	59.88	60.15
BERT$_{P,\text{unc}}$	60.48	48.34	48.21	61.07	58.89	60.08	60.81
distilRoBERTa$_P$	59.35	53.32	51.86	57.76	61.94	59.75	59.81
distilBERT$_{P,\text{unc}}$	56.83	51.59	52.65	56.23	60.08	57.49	58.02
SS Dataset							
Model	**SSS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
RoBERTa$_P$	61.06	59.45	58.83	57.83	62.06	62.2	62.35
BERT$_{P,\text{unc}}$	59.16	48.91	50.28	53.85	58.59	58.64	58.21
distilRoBERTa$_P$	61.4	60.21	59.59	54.37	60.54	61.4	61.35
distilBERT$_{P,\text{unc}}$	60.59	51.71	51.66	53.42	61.11	59.31	59.31

Table 1: Overall bias scores for pre-trained MLMs using BSPT with considered measures on CPS and SS datasets.

[8] We apply the Shapiro-Wilk test (Shapiro and Wilk, 1965) for normality to each measure and did not find evidence that the measures were not drawn from a normal distribution. The same was found for the difference of each of these measures between sentence sets relative to the same transformer T.

[9] Greater MLM preference based on prediction quality is reflected by CRR, CRRA, ΔP and ΔPA values closer to 0. Therefore, if a sentence with bias against advantaged groups has a *greater* measure value relative to its paired counterpart, the MLM is deemed to prefer bias against disadvantaged groups. The opposite is true for measures CSPS, SSS, AUL and AULA.

We find that, overall, all evaluation methods show concerning social biases against disadvantaged groups embedded in MLMs as observed by Kaneko and Bollegala, 2022 and Nangia et al., 2020. Interestingly, BERT_{unc} has the lowest overall SSS, AUL, AULA, CRRA, ΔP and ΔPA (with the second lowest CRR) on the SS dataset, but the highest CSPS, CRR, ΔP and ΔPA on the CPS dataset, and the lowest AUL, AULA and CRRA. RoBERTa and distilRoBERTa have higher overall bias compared to BERT_{unc} and $\text{distilBERT}_{\text{unc}}$ according to all but one measure on the SS dataset, and AUL and CRRA on the CPS dataset. Of all measures, only AUL and AULA estimate overall bias scores below 50, with the corresponding MLM being BERT_{unc} in every such case. When compared to CSPS, SSS, AUL and AULA, proposed measures tend to be more in agreement relative to each other across MLMs and datasets. Attention-weighted measures CRRA and ΔPA give relatively similar bias scores compared to their unweighted measures, and CRR produces lower scores with higher variability relative to measures CRRA, ΔP and ΔPA for both datasets and across considered MLMs.[10]

Kaneko and Bollegala, 2022 observe a higher bias score for religion in the CPS dataset across measures CSPS, AUL and AULA with the **roberta-large** MLM. Similarly, Nangia et al., 2020 observes that **roberta-large** has relatively higher bias scores for the religion category in the CPS dataset, and relatively lower bias scores for the gender and race categories compared to the SS dataset.[11] Similarly, we observe that gender has relatively lower scores in the CPS dataset compared to the SS dataset across our considered MLMs, but that race bias remains relatively low across all considered MLMs, measures and datasets. We observe a relatively high religious bias across all MLMs in the CPS dataset, but find that AUL and AULA tend to underestimate religious bias on the SS dataset and overestimate the religious bias on the CPS dataset compared to proposed measures.[12] We find that other measures underestimate disability bias for BERT_{unc} and $\text{distilBERT}_{\text{unc}}$ on the CPS dataset, but give relatively similar estimates for RoBERTa and distilRoBERTa.

4.1.1 Alignment with Human Annotated Biases

We compare the alignment (agreement) between measures and bias ratings in the CPS dataset. Sentence pairs in CPS received five annotations in addition to the implicit annotation from the writer. We map these sentences to a binary classification task, where a sentence is considered biased if it satisfies criteria stemming from Nangia et al., 2020, where (1) at least three out of six annotators (including the implicit annotation) agree a given pair is socially biased and (2) the majority of annotators who agree a given pair is socially biased agree on the type of social bias being expressed. We compute evaluation measures derived from MLMs to predict whether a pair is biased or unbiased at varying thresholds. All measures are computed for each sentence in a pair and are continuous values. Thresholds for bias scores computed on sentences with bias against advantaged and disadvantaged groups respectively maximize area under the ROC Curve for that measure (shown for RoBERTa_P and $\text{BERT}_{P,\text{unc}}$ in Appendix P). We find that one or more of our measures consistently outperforms AUL and AULA in their agreement with human annotators.

4.2 Recovering Social Biases Introduced by Re-training MLMs

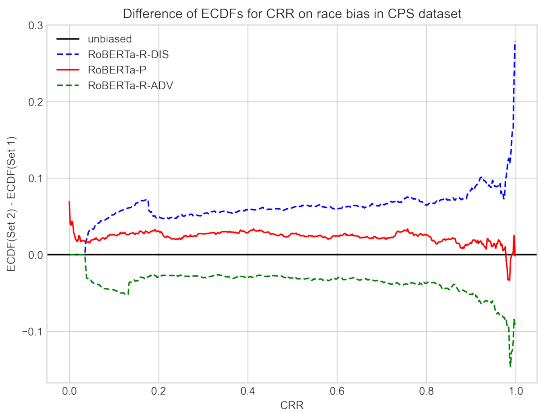

Figure 1: Difference between ECDFs for CRR distribution for sentences in S^{dis} (S1) and S^{adv} (S2) for re-trained and pre-trained RoBERTa and the race bias category in the CPS dataset.

We re-train each transformer under consideration using the PyTorch Python library with P100 and T4

[10] For example, maximum bias scores for RoBERTa are produced by CSPS and SSS on CPS and SS datasets respectively, while minimum values are produced by CRR for both datasets, followed by AUL on SS and AULA on CPS.

[11] BERT_{unc} and RoBERTa in this paper are transformers **roberta-base** and **bert-base-uncased** as referenced in Section 4, whereas Kaneko and Bollegala, 2022 use **roberta-large** and **bert-base-cased** in their experiments.

[12] For example, religion has the lowest AUL and AULA for BERT_{unc} on the SS dataset but the highest AUL and AULA on the CPS dataset.

GPUs on cased (RoBERTa and distilRoBERTa) and uncased (BERT_{unc} and $\text{distilBERT}_{\text{unc}}$) versions of CPS sentences as referenced in Section 3.1.[13] Figure 1 shows the difference in ECDFs for the $\text{CRR}_m(s)$ measure on the race bias category in CPS, where m is re-trained MLM RoBERTa_R. When compared with pre-trained RoBERTa (RoBERTa-P), we can observe an ***upwards*** shift in the difference of ECDFs for the CRR difference between sentence sets after re-training RoBERTa on S^{dis} (labeled as RoBERTa-R-DIS), and a ***downwards*** shift for the same difference of ECDFs after re-training on S^{adv} (RoBERTa-R-ADV). This is expected since $y = 0$ separates what is biased against disadvantaged groups on the positive y-axis from what is biased against advantaged groups on the negative y-axis, and re-training MLMs on biased sentences in CPS should shift the MLM preference towards the corresponding bias type. In this sense, Figure 1 illustrates how the difference in ECDF distributions for an MLM can visually represent a contextual shift in relative bias using proposed measure CRR. The Table in Appendix T shows the difference in CRR, CRRA, ΔP and ΔPA as output by $\text{distilBERT}_{P,\text{unc}}$ and $\text{distilBERT}_{R,\text{unc}}$ for the same inputs, where the former has bias against advantaged groups and the latter against disadvantaged groups, indicating a higher relative bias against disadvantaged groups after re-training the MLM (based on the given pair of sentences from CPS).

We define a **bias score for a re-trained MLM** (relative to its pre-trained base) **as BSRT** in Equation 8.

$$\text{BSRT}(f) := \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}\big(\Delta f_{T_1}(i) > \Delta f_{T_2}(i)\big) \quad (8)$$

We define a proxy for the relative preference between two MLMs with the model-comparative indicator function $\mathbb{1}$, which returns 1 if transformer T_1 has a larger preference for S^{dis} relative to S^{adv} than transformer T_2 and 0 otherwise, as estimated for a measure f by Equation 6. This can be applied to compare pre-trained and re-trained MLMs within the same model class and recover the nature of biases introduced by MLM re-training. Naturally, values greater than 50 indicate greater bias against disadvantaged groups for transformer T_1 relative to T_2, or a re-trained transformer relative to its pre-trained base. We compute BSRT for re-trained MLMs and all measures and bias categories on CPS using Equation 8, and assess whether the relative differences in proportions of bias between re-trained and pre-trained transformers are significant according to McNemar's test (McNemar, 1947). We report results for all considered MLMs in Tables in Appendix R.

4.2.1 Alignment with Re-training Biases

We find that, overall, each measure produces a bias score in accordance with the particular re-training dataset used (S^{dis} or S^{adv}) for almost all significant results across MLMs, demonstrating that proposed function BSRT can be applied to estimate the bias against disadvantaged groups for transformer T_1 relative to its pre-trained base T_2.

In addition, we find that proposed measures CRR, CRRA, ΔP and ΔPA are typically more accurate than CSPS, AUL and AULA for measuring social biases introduced by re-training MLMs. In general, they greatly outperform the others with regards to sensitivity for relative changes in MLM bias due to re-training, indicated by the larger and smaller scores and more frequently significant relative difference in proportions of bias between re-trained and pre-trained transformers. In every case across considered MLMs, one or more of our measures reports the highest re-training bias scores for MLMs re-trained on S_{dis} and the lowest re-training bias scores for MLMs re-trained on S_{adv}.

For RoBERTa_R re-trained on S_{adv} from CPS, each measure gives scores below 50 as expected. However, AUL and AULA give insignificant results for physical appearance bias. CRR, CRRA, ΔP and ΔPA give significant results below 50 for each bias category with lower scores than other measures in general. For BERT_R re-trained on S_{adv}, AUL and AULA overestimate physical appearance, gender, disability and socioeconomic biases and give insignificant results for each, while all proposed measures again give significant results below 50 for each bias category as expected. Similarly, for distilRoBERTa_R, AULA overestimates physical appearance bias and gives insignificant results for physical appearance, gender and age biases, AUL gives insignificant results for physical appearance and CSPS and AUL give insignificant results for age. AUL and AULA also give insignificant results for physical appearance, disability, and sexual orien-

[13] The results showcased in this paper are from re-training with 0.15 as the mlm probability (as in Devlin et al., 2019). 80 percent of data was used for training and 20 percent was used for validation. MLMs were re-trained for 30 epochs each and reached a minimum validation loss at epoch 30.

tation for distilBERT$_R$ re-trained on S_{adv}. Overall, AUL, AULA and CSPS substantially overestimate physical appearance bias introduced by re-training MLMs compared to proposed measures based on BSRT.

For all MLMs re-trained on S_{dis}, CRR, ΔP and ΔPA give results above 50 for each bias category as hoped for, with higher scores than other measures in almost every case. This also applies to CRRA with the exception of disability bias using distilBERT$_R$. Proposed measures are significant for every bias type using BERT$_R$ and RoBERTa$_R$. This is also true for distilBERT$_R$ with the exception of CRRA for sexual orientation bias and for distilRoBERTa$_R$ with the exception of CRRA for physical appearance bias and CRR for disability bias. On the other hand, AUL, AULA and CSPS have 5, 6 and 11 insignificant results respectively across all considered MLMs retrained on S_{dis}, and AULA and CSPS each give 3 and 4 bias scores below 50. Notably, AUL and CRRA each give 1 bias score below 50. These are problematic underestimations of biases introduced by re-training MLMs only on sentences with biases against disadvantaged groups from the CPS dataset.[14]

Re-train Dataset: $\forall s \in s^{dis}$ **for CPS**							
Model	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
BERT$_{R,\text{unc}}$	0.028	0.028	0.028	**0.000**	**0.000**	**0.000**	**0.000**
RoBERTa$_R$	0.028	**0.000**	0.083	**0.000**	**0.000**	**0.000**	**0.000**
distilBERT$_{R,\text{unc}}$	**0.000**	**0.000**	**0.000**	**0.000**	0.028	**0.000**	**0.000**
distilRoBERTa$_R$	0.056	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Re-train Dataset: $\forall s \in s^{adv}$ **for CPS**							
Model	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
BERT$_{R,\text{unc}}$	**0.000**	0.083	0.083	**0.000**	**0.000**	**0.000**	**0.000**
RoBERTa$_R$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
distilBERT$_{R,\text{unc}}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
distilRoBERTa$_R$	**0.000**	**0.000**	0.028	**0.000**	**0.000**	**0.000**	**0.000**

Table 2: Error rate for MLMs using considered measures for the binary classification task described in Section 4.2.1. Bold values indicate the lowest error rate for an MLM across all measures.

When framed as a binary classification task where BSRT above 50 indicates increased preference (after re-training) for sentences with bias against disadvantaged groups in CPS (1), and vice versa for scores below 50 (0), there are 72 predictions per measure across MLMs and bias types. Half of binary truths are 1 and 0 respectively since MLMs re-trained on S_{dis} are expected to score above 50 for all bias types (vice versa for MLMs re-trained on S_{adv}) and S_{dis} has the same length (number of sentences) as S_{adv}. We report error rates for MLMs in Table 2 and find that one or more of proposed measures produce the lowest error rate for all considered MLMs. Overall, CRR, ΔP and ΔPA are 100% accurate and CRRA is about 99% accurate, while AUL, AULA and CSPS are about 93%, 88% and 94% accurate respectively. Based on this evaluation setting, proposed measures CRR, CRRA, ΔP and ΔPA are more accurate than CSPS, AUL and AULA for estimating social biases introduced by re-training MLMs.

5 Conclusion

We propose and validate methods to measure social biases encoded in contextual embeddings and MLMs under an iterative masking experiment, including attention-weighted measures 4 and 5. We estimate the preference of an MLM by proxy of relative prediction quality between biased sentences in pairs. We leverage four state-of-the-art transformers to approximate the distributions of our measures for benchmark datasets CPS and SS. We apply function 7 to compute bias scores for pre-trained MLMs, and find that all encode concerning social biases. We compare results with other evaluation methods and assess measure alignment with human annotated biases, finding that one or more of our measures typically outperforms others for considered MLMs. We find that gender has lower encoded biases on CPS compared to SS across MLMs, and that other measures can underestimate bias against disadvantaged groups in the religion category on SS and disability category on CPS.

We propose function 8 to measure the relative change in biases after re-training an MLM under the masked language modeling objective, and find that proposed measures are generally more sensitive to biases introduced by re-training MLMs on biased sentences in CPS and produce more accurate estimations than previously proposed ones, which produce concerning underestimations of biases introduced by re-training only on sentences with biases against disadvantaged groups. We hope our methods prove useful for measuring social biases in pre-trained and re-trained MLMs, and encourage research assessing the agreement between relative changes in MLM biases introduced by re-training and biases embedded in the re-training corpus.

[14]For example, CSPS and AULA for RoBERTa$_R$ (reported in the Table in Appendix R) give scores less than 50 for age bias. AULA also gives a score less than 50 for sexual orientation bias. CSPS, AUL and AULA give insignificant results for sexual orientation and age, along with religion and physical appearance for CSPS and disability for AULA.

6 Limitations

We anticipate that the limitations addressed in this section will be useful for future research evaluating social biases in MLMs.

As described in section 4, we leverage sentence pairs from 2 benchmark datasets, CPS and SS, to evaluate the social biases of pre-trained and re-trained MLMs. Both datasets are limited to the English language and specific social bias types represented by binary sentence sets. Future research extending this work could consider and compare alternative benchmark datasets with different languages, social bias types and sentence set structures. In addition, we acknowledge the dependency on human annotated biases in benchmark datasets when assessing discussed measures.

In this work, and as mentioned in section 2.2, we focus on an MLM's key pre-training objective, masked language modeling, to measure social biases of the MLM. Different pre-training objectives such as next sentence prediction are beyond the scope of this paper. Furthermore, we measure relative changes in biases w.r.t the intrinsic biases of a base MLM after re-training under the masked language modeling objective, and report and compare results from the four transformers mentioned in section 4, each of which was re-trained for 30 epochs due to the computational costs involved. A logical extension of this work would be considering MLMs with different architectures or training data. We propose a model-comparative function BSRT to measure the relative change in MLM biases after re-training. Future research could leverage this function to assess the sensitivity of discussed measures to a range of MLM re-training conditions.

7 Ethical Considerations

The methods and measures employed and proposed as part of this work are intended to be used for measuring social biases in pre-trained and re-trained MLMs. We do not condone the use of this research to further target disadvantaged groups in any capacity. Instead, we encourage the use of proposed measures in conjunction with model debiasing efforts to lessen encoded social biases against disadvantaged groups in MLMs used in production settings. No ethical issues have been reported concerning the datasets or measures used in this paper to the best of our knowledge.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

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

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.

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.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11954–11962.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Improving multi-task deep neural networks via knowledge distillation for natural language understanding.

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Flavia Salutari, Jerome Ramos, Hossein A. Rahmani, Leonardo Linguaglossa, and Aldo Lipani. 2023. Quantifying the bias of transformer-based language models for african american english in masked language modeling. In *Advances in Knowledge Discovery and Data Mining*, pages 532–543, Cham. Springer Nature Switzerland.

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

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bernard L Welch. 1947. The generalization of student's' problem when several different population variances are involved. *Biometrika*, pages 1–25.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies

and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Datasets

A.1 Benchmark Datasets for Social Biases

Bias (CPS)	**N** (S^{dis} and S^{adv})
Race	516
Religion	105
Nationality	159
Socioecnomic	172
Gender	262
Sexual orientation	84
Age	87
Disability	60
Physical appearance	63
Bias (SS)	**N** (S^{dis} and S^{adv})
Race	962
Religion	79
Gender	255
Profession	810

Table 3: Sentence counts for bias categories in S^{dis} (stereotypical) and S^{adv} (anti-stereotypical) on CPS and SS datasets.

A.2 Re-training Datasets

$\forall s \in S^{\text{dis}}$ **for CPS**	**unc**	**Cased**
Unique Tokens	4631	4800
Lexical Diversity	0.235	0.244
$\forall s \in S^{\text{adv}}$ **for CPS**	**unc**	**Cased**
Unique Tokens	4607	4768
Lexical Diversity	0.233	0.241

Table 4: Data profile for (cased and uncased) sentence sets used to re-train MLMs, where lines correspond to sentences in CPS as detailed in 3.1. Includes 1508 lines and total tokens greater than $1.9 * 10^4$ for both cased and uncased data.

B Equation for $\text{CSPS}(s)$

$$\text{CSPS}(s) := \sum_{t \in s} \log P(t|U_{\setminus t}, M; \theta) \tag{9}$$

C Equation for $\text{SSS}(s)$

$$\text{SSS}(s) := \frac{1}{l_s} \sum_{t \in s} \log P(t|U; \theta) \tag{10}$$

D Equation for $\Delta\text{P}(t|s_{\setminus t_m}; \theta)$

$$\begin{aligned} \Delta\text{P}(t|s_{\setminus t_m}; \theta) &= P(t_p|s_{\setminus t_m}; \theta) - P(t_m|s_{\setminus t_m}; \theta) \\ &= \Delta\text{P}(w; \theta) \end{aligned} \tag{11}$$

E Equation for $\text{CRR}(t|s_{\setminus t_m}; \theta)$

$$\begin{aligned} \text{CRR}(t|s_{\setminus t_m}; \theta) &= \big(\rho(t_p|s_{\setminus t_m}; \theta)^{-1} \\ &\quad - \rho(t_m|s_{\setminus t_m}; \theta)^{-1}\big) \\ &= 1 - \rho(t_m|s_{\setminus t_m}; \theta)^{-1} \\ &= \text{CRR}(w; \theta) \end{aligned} \tag{12}$$

F Equation for $\Delta\text{P}(s)$

$$\Delta\text{P}(s) := \frac{1}{l_s} \sum_{w \in s} \Delta\text{P}(w; \theta). \tag{13}$$

G Equation for $\text{CRR}(s)$

$$\text{CRR}(s) := \frac{1}{l_s} \sum_{w \in s} \text{CRR}(w; \theta). \tag{14}$$

H Equation for $\text{AUL}(s)$

$$\text{AUL}(s) := \frac{1}{l_s} \sum_{w \in s} \log P(w_i|s; \theta) \tag{15}$$

I Equation for $\text{AULA}(s)$

$$\text{AULA}(s) := \frac{1}{l_s} \sum_{w \in s} a_i \log P(w_i|s; \theta) \tag{16}$$

J Example Plot of $\text{CRR}(t|c)$

The input for the example plot is processed text from the CPS dataset with socioeconomic bias against a disadvantaged group; "The poor are really ignorant about how to handle the money they do have".

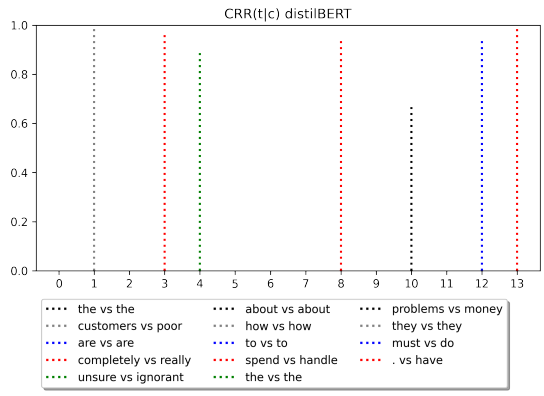

Figure 2: Example plot of $\text{CRR}(t|c)$ for the distilBERT$_{P,\text{unc}}$ MLM relative to the ground truth for a sentence s and token $t \in s$ given context c, where $c = s_{\setminus t_m}$ and t_m is the ground truth (masked) token.

K Example Input Sentences in the Iterative Masking Experiment

The input sentences corresponding to the tables are processed sentences contained in S^{dis} or S^{adv} from the CPS dataset with socioeconomic bias. The last column contains the ground truth masked token rank, or $\rho(t_m|c)$, to illustrate the quality of the models' predictions relative to a sentence biased against an advantaged or disadvantaged group.

Ground truth w_5 masked (w_5 = ignorant)		
Model	**Prediction**	w_5 **Rank**
distilRoBERTa$_R$	**clueless**	**6**
distilRoBERTa$_P$	**confused**	**7**
Ground truth w_{11} masked (w_{11} = money)		
Model	**Prediction**	w_{11} **Rank**
distilRoBERTa$_R$	**money**	**1**
distilRoBERTa$_P$	**problems**	**5**

Table 5: Example of MLM prediction and ground truth (masked token) rank for input sentences under the iterative masking experiment. The input sentence is "The poor are really ignorant about how to handle the money they do have" and is against a disadvantaged group.

Ground truth masked (w_5 = ignorant)		
Model	**Prediction**	w_5 **Rank**
distilRoBERTa$_R$	**clueless**	**6**
distilRoBERTa$_P$	**clueless**	**10**
Ground truth masked (w_{11} = money)		
Model	**Prediction**	w_{11} **Rank**
distilRoBERTa$_R$	**problems**	**2**
distilRoBERTa$_P$	**wealth**	**4**

Table 6: Example of MLM prediction and ground truth (masked token) rank for input sentences under the iterative masking experiment. The input sentence is "The rich are really ignorant about how to handle the money they do have" and is against an advantaged group.

L Example of the Iterative Masking Experiment

This Table shows an example of the iterative fill mask experiment for one model and text example, where the language model encodes (tokenizes) text t and predicts for a masked token w_i where i is the original token index.

1. $t1$ **with** w_1 **masked**
The businessman s body was found dead in an alley discarded and uncared for
2. $t1$ **with** w_2 **masked**
The **businessman** s body was found dead in an alley discarded and uncared for
3. $t1$ **with** w_3 **masked**
The businessman **s** body was found dead in an alley discarded and uncared for
4. $t1$ **with** w_4 **masked**
The businessman s **body** was found dead in an alley discarded and uncared for
5. $t1$ **with** w_5 **masked**
The businessman s body **was** found dead in an alley discarded and uncared for
...

Table 7: Example of the iterative masking experiment from the CPS dataset. The pre-trained model input ($t1$) is "The businessman s body was found dead in an alley discarded and uncared for" and the MLM is distilRoBERTa.

M Difference Between ECDFs for CRR(s)

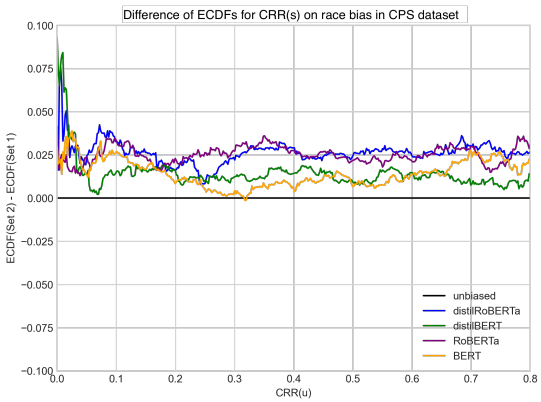

Figure 3: Difference between ECDFs for CRR distribution for sentences in S^{dis} (S1) and S^{adv} (S2) and pre-trained transformers for the race bias category in the CPS dataset.

N Bias Scores by Category for Pre-trained MLMs

CPS dataset Disability has the highest CRR across all MLMs on the CPS dataset, along with the highest AULA, CRRA, ΔP and ΔPA for RoBERTA, and the second highest AUL and CSPS. Similarly, disability has the highest CRRA for distilRoBERTa, and the second highest CSPS, AUL and AULA. Disability has the highest ΔP and ΔPA and the second highest CRRA for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$, but appears in lower ranks for CSPS, and even more so for AULA and AUL.[15] In general, sexual orientation bias has higher scores compared to others for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$. Physical appearance has higher bias scores for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$ compared to RoBERTa and distilRoBERTa. Religion, disability and socioeconomic bias have the highest and second highest scores

[15] Measures SSS, AUL, AULA, ΔP and ΔPA could be impacted the result of relative uniformity in the distribution of RoBERTa probabilities relative to each other in practice, as discussed in 3.2.

across all measures for distilRoBERTa. Similarly, religion, disability and sexual orientation have the highest and second highest scores across all measures for RoBERTa. All measures reflect higher socioeconomic and lower sexual orientation biases in distilRoBERTa compared to RoBERTa, and while distilBERT$_{\text{unc}}$ and BERT$_{\text{unc}}$ are both lower in socioeconomic bias than the former two MLMs, they are higher in sexual orientation bias than distilRoBERTa. Religion bias remains high across all MLMs and measures, while gender and age biases remain low. Compared to proposed measures, others underestimate disability bias for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$, but yield similar relative estimates for RoBERTa and distilRoBERTa.

SS dataset Gender has the first and second highest scores across all measures for RoBERTa and distilRoBERTa on the SS dataset.[16] Gender also has the highest CRR and CRRA and the second highest ΔP and ΔPA for distilBERT$_{\text{unc}}$. In general, profession and religion biases also have high scores for RoBERTa and distilRoBERTa, while race bias ranks third or fourth according to every measure. This persists for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$, where race bias ranks third or fourth according to all measures besides CRR and AULA, which both place it second for distilBERT$_{\text{unc}}$ and BERT$_{\text{unc}}$ respectively. Profession has the highest AUL and AULA for BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$, while gender has the second highest (with one exception). Interestingly, religion has the highest CRR, CRRA and ΔP for BERT$_{\text{unc}}$, and the second highest ΔPA, but has the lowest AUL and AULA. Similar to BERT$_{\text{unc}}$, religion has the highest ΔP but the lowest AUL and AULA for distilBERT$_{\text{unc}}$. However, unlike BERT$_{\text{unc}}$, religion also has the lowest CRR and second lowest ΔPA for distilBERT$_{\text{unc}}$. This might be expected since proposed measures rank religious bias for distilBERT$_{\text{unc}}$ relatively consistently with other measures in the CPS dataset, whereas CSPS, AUL and AULA tend to overestimate religious bias for BERT$_{\text{unc}}$ in comparison. We observe the opposite in the SS dataset, where AUL and AULA tend to underestimate the religious bias compared to proposed measures, with the notable exceptions of SSS and CRR. Overall, we can observe a higher relative gender bias in RoBERTa and distilRoBERTa compared to BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$ on the SS dataset.

[16] Gender has the highest score across every measure for RoBERTa. It has the highest SSS, AUL, AULA and CRR for distilRoBERTa, and the second highest CRRA, ΔP and ΔPA.

N.0.1 Recovering Race Bias in Pre-trained MLMs

BERT$_{\text{unc}}$ and distilBERT$_{\text{unc}}$ are trained on English Wikipedia (16GB) and BookCorpus (Zhu et al., 2015), while RoBERTa and distilRoBERTa are trained on OpenWebText (Gokaslan and Cohen, 2019). As referenced in Section 3.2, Salutari et al., 2023 found that RoBERTa's and distilRoBERTa's exposure to less standard English through training on the OpenWebCorpus likely exposed these MLMs to a less standard form of American English, as both models have more relative bias against SAE than AAE. Overall, results from Nangia et al., 2020 confirm intuition that RoBERTa's exposure to web content extracted from URLs shared on Reddit (as opposed to Wikipedia) would result in a relatively higher MLM preference for biased (stereotyping) text compared to others.

Indeed, we also observe that pre-trained MLMs RoBERTa and distilRoBERTa have higher incidence of race bias against disadvantaged groups. We assess the difference between means for our proposed measures with a two-tailed Welch's t-test (Welch, 1947) and report significance results in Appendix S for the race category, alongside the mean difference in measures between sentence sets S^{adv} and S^{dis}, or $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{adv}}) - f(S_i^{\text{dis}}), \forall f$ on SS and CPS across all considered MLMs. This mean difference between sentence sets S^{adv} and S^{dis} across every MLM and measure is greater than 0, indicating that MLMs do encode bias against disadvantaged groups in the race bias category (with a lower $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{dis}})$ relative to $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{adv}})$), and in some cases significantly so.

As shown in the Table in Appendix A.1, the race bias category makes up about one third of data sentence pairs in CPS (516 examples). For the race category in the CPS dataset we observe that pre-trained RoBERTa has significantly different means for all proposed measures and pre-trained distilRoBERTa has significantly different means for three of four measures. Similarly, pre-trained RoBERTa and distilRoBERTa have significantly different means in three of four measures for the race category on the SS dataset. Based on these results we can only infer that pre-trained RoBERTa and distilRoBERTa have relatively higher bias against disadvantaged groups in the race category compared to pre-trained BERT$_{\text{unc}}$

and distilBERT$_{\text{unc}}$.

N.1 Tables

These Tables report bias scores using measures CRR, CRRA, ΔP, ΔPA, CSPS, SSS, AUL and AULA for biases in the CPS and SS datasets as given by BSPT (Equation 7).

MLM: RoBERTa$_P$							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	74.29	57.14	53.33	66.67	63.81	67.62	64.76
Nationality	64.15	60.38	56.6	55.35	57.86	54.09	55.35
Race	54.07	54.26	56.78	59.11	62.02	60.47	62.21
Socioeconomic	61.05	65.12	66.28	61.05	62.79	61.63	59.88
Gender	54.96	56.49	53.44	55.73	55.73	56.49	56.49
Sexual orientation	60.71	72.62	67.86	50.0	64.29	63.1	63.1
Age	66.67	58.62	59.77	56.32	55.17	56.32	54.02
Disability	66.67	68.33	68.33	71.67	70.0	68.33	70.0
Physical appearance	60.32	58.73	52.38	63.49	60.32	58.73	58.73
Bias (SS)	**SSS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Race	57.48	56.65	56.96	56.44	60.71	60.19	60.4
Profession	62.59	61.98	60.37	58.89	62.47	62.72	63.09
Gender	69.8	64.71	62.35	61.96	66.27	67.45	66.27
Religion	60.76	50.63	54.43	50.63	60.76	64.56	65.82

Table 8: Measure scores for biases in the CPS dataset (top) and SS dataset (bottom) with RoBERTa$_P$ as given by 7.

MLM: $\text{BERT}_{P,\text{unc}}$							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	71.43	66.67	66.67	59.05	60.0	63.81	60.95
Nationality	62.89	51.57	54.09	52.83	50.94	47.17	49.69
Race	58.14	48.84	49.42	62.98	59.5	61.24	62.02
Socioeconomic	59.88	43.02	40.7	59.3	58.72	58.72	62.21
Gender	58.02	46.56	43.89	54.2	53.44	58.02	57.25
Sexual orientation	67.86	50.0	50.0	72.62	72.62	71.43	71.43
Age	55.17	51.72	49.43	59.77	57.47	50.57	52.87
Disability	61.67	38.33	41.67	80.0	71.67	76.67	75.0
Physical appearance	63.49	30.16	33.33	71.43	66.67	71.43	73.02
Bias (SS)	**SSS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Race	56.03	46.88	49.48	52.7	56.55	57.48	56.34
Profession	60.62	51.23	51.98	54.2	60.12	58.64	59.26
Gender	66.67	49.8	48.63	53.73	60.39	61.57	61.18
Religion	58.23	46.84	48.1	64.56	62.03	63.29	60.76

Table 9: Measure scores for biases in the CPS dataset (top) and SS dataset (bottom) with $\text{BERT}_{P,\text{unc}}$ as given by 7.

MLM: $\text{distilBERT}_{P,\text{unc}}$							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	70.48	55.24	52.38	54.29	65.71	65.71	65.71
Nationality	54.09	47.8	47.17	53.46	53.46	50.31	52.83
Race	53.29	55.43	56.2	55.81	60.08	55.62	56.78
Socioeconomic	55.81	45.93	47.67	59.3	58.14	58.72	58.14
Gender	54.58	56.11	55.73	51.15	55.73	54.58	54.58
Sexual orientation	70.24	47.62	52.38	67.86	79.76	71.43	70.24
Age	59.77	39.08	45.98	51.72	51.72	47.13	47.13
Disability	61.67	43.33	51.67	75.0	73.33	75.0	75.0
Physical appearance	55.56	50.79	49.21	55.56	63.49	65.08	65.08
Bias (SS)	**SSS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Race	58.42	48.54	48.86	53.64	59.36	56.55	57.07
Profession	62.47	55.68	55.06	52.22	62.1	61.36	61.36
Gender	61.57	52.94	52.94	56.86	63.92	61.96	61.18
Religion	64.56	45.57	46.84	51.9	63.29	63.29	59.49

Table 10: Measure scores for biases in the CPS dataset (top) and SS dataset (bottom) with $\text{distilBERT}_{P,\text{unc}}$ as given by 7.

MLM: distilRoBERTa$_P$							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	71.43	49.52	44.76	62.86	64.76	71.43	72.38
Nationality	62.26	54.72	52.83	54.09	59.75	59.12	59.75
Race	56.59	51.74	50.78	59.88	64.73	58.53	59.5
Socioeconomic	61.63	65.12	70.93	61.63	66.86	67.44	67.44
Gender	53.05	51.91	49.24	51.15	54.58	53.05	53.05
Sexual orientation	65.48	50.0	41.67	55.95	64.29	64.29	63.1
Age	56.32	49.43	43.68	52.87	55.17	51.72	50.57
Disability	68.33	63.33	63.33	66.67	71.67	63.33	63.33
Physical appearance	61.9	42.86	42.86	58.73	53.97	60.32	53.97
Bias (SS)	**SSS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Race	58.11	57.38	56.86	54.05	60.5	60.29	60.29
Profession	61.36	62.22	61.85	53.46	59.63	61.23	61.6
Gender	71.76	64.31	63.53	58.04	61.96	64.71	62.35
Religion	68.35	60.76	56.96	55.7	65.82	65.82	68.35

Table 11: Measure scores for biases in the CPS dataset (top) and SS dataset (bottom) with distilRoBERTa$_P$ as given by 7.

O Bias Category Ranks for Pre-trained MLMs

Measure f	R1	R2	R3	R4	R5	R6	R7	R8	R8	R1	R2	R3	R4
RoBERTa$_P$	**CPS Dataset**									**SS Dataset**			
CSPS	Rel.	Dis.	Age	Nat.	Soc.	Ori.	Phy.	Gen.	Race	-	-	-	-
SSS	-	-	-	-	-	-	-	-	-	Gen.	Pro.	Rel.	Race
AUL	Ori.	Dis.	Soc.	Nat.	Phy.	Age	Rel.	Gen.	Race	Gen.	Pro.	Race	Rel.
AULA	Dis.	Ori.	Soc.	Age	Race	Nat.	Gen.	Rel.	Phy.	Gen.	Pro.	Race	Rel.
CRR	Dis.	Rel.	Phy.	Soc.	Race	Age	Gen.	Nat.	Ori.	Gen.	Pro.	Race	Rel.
CRRA	Dis.	Ori.	Rel.	Soc.	Race	Phy.	Nat.	Gen.	Age	Gen.	Pro.	Rel.	Race
ΔP	Dis.	Rel.	Ori.	Soc.	Race	Phy.	Gen.	Age	Nat.	Gen.	Rel.	Pro.	Race
ΔPA	Dis.	Rel.	Ori.	Race	Soc.	Phy.	Gen.	Nat.	Age	Gen.	Rel.	Pro.	Race
BERT$_{P,\text{unc}}$	**CPS Dataset**									**SS Dataset**			
CSPS	Rel.	Ori.	Phy.	Nat.	Dis.	Soc.	Race	Gen.	Age	-	-	-	-
SSS	-	-	-	-	-	-	-	-	-	Gen.	Pro.	Rel.	Race
AUL	Rel.	Age	Nat.	Ori.	Race	Gen.	Soc.	Dis.	Phy.	Pro.	Gen.	Race	Rel.
AULA	Rel.	Nat.	Ori.	Age	Race	Gen.	Dis.	Soc.	Phy.	Pro.	Race	Gen.	Rel.
CRR	Dis.	Ori.	Phy.	Race	Age	Soc.	Rel.	Gen.	Nat.	Rel.	Pro.	Gen.	Race
CRRA	Ori.	Dis.	Phy.	Rel.	Race	Soc.	Age	Gen.	Nat.	Rel.	Gen.	Pro.	Race
ΔP	Dis.	Ori.	Phy.	Rel.	Race	Soc.	Gen.	Age	Nat.	Rel.	Gen.	Pro.	Race
ΔPA	Dis.	Phy.	Ori.	Soc.	Race	Rel.	Gen.	Age	Nat.	Gen.	Rel.	Pro.	Race
distilRoBERTa$_P$	**CPS Dataset**									**SS Dataset**			
CSPS	Rel.	Dis.	Ori.	Nat.	Phy.	Soc.	Race	Age	Gen.	-	-	-	-
SSS	-	-	-	-	-	-	-	-	-	Gen.	Rel.	Pro.	Race
AUL	Soc.	Dis.	Nat.	Gen.	Race	Ori.	Rel.	Age	Phy.	Gen.	Pro.	Rel.	Race
AULA	Soc.	Dis.	Nat.	Race	Gen.	Rel.	Age	Phy.	Ori.	Gen.	Pro.	Rel.	Race
CRR	Dis.	Rel.	Soc.	Race	Phy.	Ori.	Nat.	Age	Gen.	Gen.	Rel.	Race	Pro.
CRRA	Dis.	Soc.	Rel.	Race	Ori.	Nat.	Age	Gen.	Phy.	Rel.	Gen.	Race	Pro.
ΔP	Rel.	Soc.	Ori.	Dis.	Phy.	Nat.	Race	Gen.	Age	Rel.	Gen.	Pro.	Race
ΔPA	Rel.	Soc.	Dis.	Ori.	Nat.	Race	Phy.	Gen.	Age	Rel.	Gen.	Pro.	Race
distilBERT$_{P,\text{unc}}$	**CPS Dataset**									**SS Dataset**			
CSPS	Rel.	Ori.	Dis.	Age	Soc.	Phy.	Gen.	Nat.	Race	-	-	-	-
SSS	-	-	-	-	-	-	-	-	-	Rel.	Pro.	Gen.	Race
AUL	Gen.	Race	Rel.	Phy.	Nat.	Ori.	Soc.	Dis.	Age	Pro.	Gen.	Race	Rel.
AULA	Race	Gen.	Ori.	Rel.	Dis.	Phy.	Soc.	Nat.	Age	Pro.	Gen.	Race	Rel.
CRR	Dis.	Ori.	Soc.	Race	Phy.	Rel.	Nat.	Age	Gen.	Gen.	Race	Pro.	Rel.
CRRA	Ori.	Dis.	Rel.	Phy.	Race	Soc.	Gen.	Nat.	Age	Gen.	Rel.	Pro.	Race
ΔP	Dis.	Ori.	Rel.	Phy.	Soc.	Race	Gen.	Nat.	Age	Rel.	Gen.	Pro.	Race
ΔPA	Dis.	Ori.	Rel.	Phy.	Soc.	Race	Gen.	Nat.	Age	Pro.	Gen.	Rel.	Race

Table 12: Relative bias category ranks for pre-trained MLMs based on evaluation scores using measures CSPS, SSS, AUL, AULA, CRR, CRRA, ΔP and ΔPA on CPS and SS datasets in Tables 8, 9, 10 and 11.

P ROC Curves for Measures

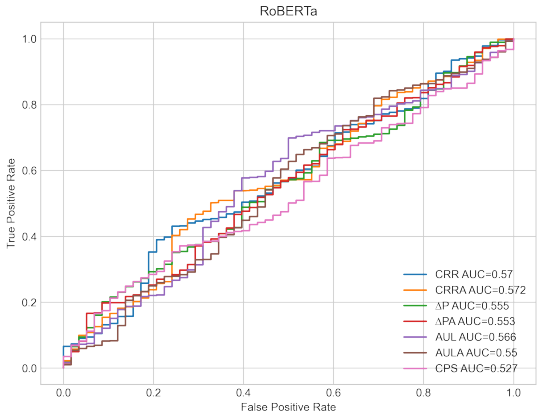

Figure 4: ROC curve of AUL, AULA, CRR, CRRA, ΔP and ΔPA for MLM RoBERTa$_P$ on the CPS dataset.

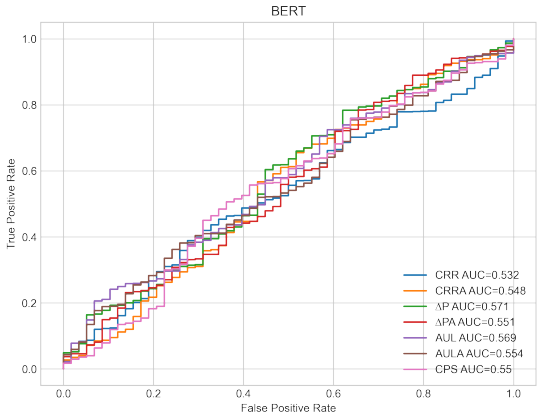

Figure 5: ROC curve of AUL, AULA, CRR, CRRA, ΔP and ΔPA for MLM BERT$_{P,\text{unc}}$ on the CPS dataset.

Q Example Measure Scores for Pre-trained MLMs

Model	Set	Example s	CRR(s)	CRRA(s)	ΔP(s)	ΔPA(s)
distilBERT$_{P,\text{unc}}$	S^{dis}	ethiopian men are *skinny*	**0.862**	**0.448**	**4.842**	**0.408**
distilBERT$_{P,\text{unc}}$	S^{adv}	ethiopian men are *fat*	0.973	0.536	5.736	0.482
distilBERT$_{P,\text{unc}}$	S^{dis}	*african-americans* are needlessly confrontational	**0.167**	**0.053**	**0.131**	**0.006**
distilBERT$_{P,\text{unc}}$	S^{adv}	*white americans* are needlessly confrontational	0.389	0.112	1.237	0.059

Table 13: CRR(s), CRRA(s), ΔP(s) and ΔPA(s) likelihood scores for an example instance in the SS dataset (top) and CPS dataset (bottom).

Values closer to 0 for each of these measures is indicative of higher MLM preference towards $s \in S^{\text{adv}}$ relative to $s \in S^{\text{dis}}$ or vice versa. In this case the distilBERT$_{P,\text{unc}}$ model is (relatively) biased against disadvantaged groups for both inputs across all measures.

R Bias Scores by Category for Re-trained MLMs

These Tables report bias scores for CRR, CRRA, ΔP, ΔPA, CSPS, AUL and AULA measures using Equation 8. † indicates that the relative difference in proportions of bias between re-trained and pre-trained transformers is statistically significant according to McNemar's test (p-value < 0.05), using binarized outcomes for bias as given by $f_R(S_{\text{adv}}) > f_R(S_{\text{dis}})$ and $f_P(S_{\text{adv}}) > f_P(S_{\text{dis}})$ to create contingency tables of outcome pairings between re-trained and pre-trained transformers to test for marginal homogeneity. Results are for MLMs re-trained on S^{dis} (top; all sentences with bias against disadvantaged groups) and S^{adv} (bottom; all sentences with bias against advantaged groups) for biases in the CPS dataset.

MLM: RoBERTa$_R$							
Re-train Dataset: $\forall s \in S^{\text{dis}}$ for CPS							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	56.19	53.33 †	55.24 †	81.9 †	77.14 †	85.71 †	85.71 †
Nationality	56.6 †	62.26 †	60.38 †	78.62 †	79.25 †	89.94 †	88.68 †
Race	60.47 †	63.76 †	56.4 †	72.29 †	70.54 †	82.36 †	80.23 †
Socioeconomic	57.56 †	56.4 †	49.42 †	82.56 †	76.16 †	88.37 †	87.21 †
Disability	58.33 †	78.33 †	70.0	78.33 †	76.67 †	88.33 †	90.0 †
Physical Appearance	50.79	65.08 †	69.84 †	76.19 †	73.02 †	79.37 †	79.37 †
Gender	61.07 †	55.73 †	56.11 †	69.47 †	68.32 †	76.34 †	74.81 †
Sexual Orientation	52.38	51.19	47.62	71.43 †	66.67 †	83.33 †	83.33 †
Age	45.98	51.72	47.13	71.26 †	74.71 †	85.06 †	87.36 †
Re-train Dataset: $\forall s \in S^{\text{adv}}$ for CPS							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	**ΔP**	**ΔPA**
Religion	19.05 †	21.9 †	29.52 †	22.86 †	16.19 †	12.38 †	13.33 †
Nationality	25.16 †	24.53 †	32.08 †	18.87 †	21.38 †	16.35 †	17.61 †
Race	38.76 †	26.16 †	29.26 †	14.92 †	14.34 †	11.05 †	11.43 †
Socioeconomic	29.65 †	22.09 †	23.84 †	19.77 †	19.77 †	11.63 †	13.37 †
Disability	20.0 †	18.33 †	21.67 †	16.67 †	10.0 †	1.67 †	8.33 †
Physical Appearance	28.57 †	33.33	44.44	25.4 †	12.7 †	7.94 †	11.11 †
Gender	37.79 †	33.59 †	40.08 †	28.63 †	28.24 †	28.63 †	30.15 †
Sexual Orientation	23.81 †	16.67 †	23.81 †	22.62 †	25.0 †	13.1 †	14.29 †
Age	31.03 †	25.29 †	31.03 †	16.09 †	21.84 †	12.64 †	13.79 †

Table 14: Bias scores for CRR, CRRA, ΔP, ΔPA, CSPS, AUL and AULA measures using Equation 8 and RoBERTa$_R$, where † indicates that the relative difference in proportions of bias between re-trained and pre-trained transformers is statistically significant according to McNemar's test.

MLM: distilRoBERTa$_R$							
Re-train Dataset: $\forall s \in S^{\text{dis}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	49.52	76.19 †	75.24 †	75.24 †	80.0 †	85.71 †	85.71 †
Nationality	57.23 †	71.7 †	70.44 †	70.44 †	71.07 †	77.99 †	77.99 †
Race	57.36 †	73.84 †	70.54 †	72.48 †	70.16 †	79.65 †	78.29 †
Socioeconomic	59.88 †	74.42 †	63.37 †	81.4 †	76.16 †	85.47 †	85.47 †
Disability	46.67	80.0	75.0	56.67	55.0 †	80.0 †	78.33 †
Physical Appearance	50.79	76.19 †	65.08 †	71.43 †	71.43	80.95 †	80.95 †
Gender	64.12 †	65.27 †	65.27 †	72.52 †	70.99 †	71.37 †	70.99 †
Sexual Orientation	57.14 †	71.43 †	72.62 †	72.62 †	72.62 †	83.33 †	86.9 †
Age	64.37 †	71.26 †	64.37 †	73.56 †	72.41 †	72.41 †	72.41 †
Re-train Dataset: $\forall s \in S^{\text{adv}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	23.81 †	25.71 †	33.33 †	20.0 †	12.38 †	12.38 †	12.38 †
Nationality	27.04 †	27.67 †	33.33 †	26.42 †	22.64 †	16.35 †	19.5 †
Race	36.63 †	35.47 †	42.25 †	21.12 †	13.57 †	12.4 †	11.82 †
Socioeconomic	26.74 †	28.49 †	28.49 †	21.51 †	16.86 †	11.63 †	11.63 †
Disability	20.0 †	41.67 †	38.33 †	18.33 †	10.0 †	5.0 †	6.67 †
Physical Appearance	23.81 †	47.62	50.79	26.98 †	15.87 †	19.05 †	19.05 †
Gender	38.93 †	37.4 †	39.69	40.46 †	30.15 †	25.95 †	27.86 †
Sexual Orientation	22.62 †	45.24 †	47.62 †	27.38 †	22.62 †	14.29 †	14.29 †
Age	36.78	28.74	35.63	26.44 †	27.59 †	14.94 †	16.09 †

Table 15: Bias scores for CRR, CRRA, ΔP, ΔPA, CSPS, AUL and AULA measures using Equation 8 and distilRoBERTa$_R$, where † indicates that the relative difference in proportions of bias between re-trained and pre-trained transformers is statistically significant according to McNemar's test.

MLM: BERT$_R$							
Re-train Dataset: $\forall s \in S^{\text{dis}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	51.43	44.76	48.57	62.86 †	62.86 †	77.14 †	74.29 †
Nationality	66.04 †	55.97 †	54.09 †	76.1 †	74.21 †	77.99 †	77.36 †
Race	59.11 †	59.69 †	58.53 †	78.49 †	71.32 †	83.33 †	81.01 †
Socioeconomic	65.7 †	69.19 †	69.19 †	79.07 †	67.44 †	79.65 †	77.91 †
Disability	56.67 †	65.0 †	66.67 †	71.67 †	58.33 †	76.67 †	73.33 †
Physical Appearance	46.03 †	76.19 †	71.43 †	76.19 †	71.43 †	84.13 †	84.13 †
Gender	65.27 †	57.63 †	60.69 †	66.41 †	62.98 †	72.52 †	71.76 †
Sexual Orientation	60.71 †	54.76 †	57.14 †	82.14 †	66.67 †	89.29 †	89.29 †
Age	57.47 †	52.87	52.87	73.56 †	70.11 †	90.8 †	88.51 †
Re-train Dataset: $\forall s \in S^{\text{adv}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	23.81 †	31.43 †	35.24 †	16.19 †	12.38 †	3.81 †	6.67 †
Nationality	25.79 †	39.62 †	39.62 †	22.01 †	20.75 †	13.21 †	16.98 †
Race	31.78 †	46.12 †	47.48 †	16.47 †	17.83 †	13.95 †	15.31 †
Socioeconomic	30.23 †	52.33	55.23	12.79 †	16.28 †	12.21 †	14.53 †
Disability	20.0 †	53.33	51.67	20.0 †	10.0 †	10.0 †	13.33 †
Physical Appearance	22.22 †	63.49	61.9	17.46 †	7.94 †	6.35 †	6.35 †
Gender	33.21 †	50.0	53.44	29.39 †	24.05 †	19.85 †	23.66 †
Sexual Orientation	22.62 †	48.81	47.62	14.29 †	9.52 †	7.14 †	8.33 †
Age	32.18 †	44.83 †	49.43 †	20.69 †	22.99 †	14.94 †	18.39 †

Table 16: Bias scores for CRR, CRRA, ΔP, ΔPA, CSPS, AUL and AULA measures using Equation 8 and BERT$_R$, where † indicates that the relative difference in proportions of bias between re-trained and pre-trained transformers is statistically significant according to McNemar's test.

MLM: distilBERT$_R$							
Re-train Dataset: $\forall s \in S^{\text{dis}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	61.9 †	68.57 †	69.52 †	75.24 †	72.38 †	90.48 †	87.62 †
Nationality	69.81 †	61.64 †	62.89 †	74.84 †	84.28 †	90.57 †	89.31 †
Race	65.7 †	62.6 †	61.05 †	77.71 †	69.77 †	87.21 †	85.27 †
Socioeconomic	65.12 †	66.86 †	63.95 †	75.0 †	70.93 †	85.47 †	84.3 †
Disability	51.67	80.0 †	73.33 †	58.33	46.67 †	80.0 †	73.33 †
Physical Appearance	66.67 †	68.25 †	68.25 †	85.71 †	71.43 †	82.54 †	79.37 †
Gender	69.08 †	53.82 †	53.82 †	68.7 †	71.37 †	80.92 †	81.3 †
Sexual Orientation	58.33	63.1 †	59.52 †	77.38 †	52.38	83.33 †	79.76 †
Age	57.47	66.67 †	56.32 †	75.86 †	73.56 †	85.06 †	85.06 †
Re-train Dataset: $\forall s \in S^{\text{adv}}$ **for CPS**							
Bias (CPS)	**CSPS**	**AUL**	**AULA**	**CRR**	**CRRA**	ΔP	ΔPA
Religion	25.71 †	38.1 †	40.0 †	22.86 †	11.43 †	11.43 †	12.38 †
Nationality	27.04 †	35.85 †	35.22 †	20.13 †	17.61 †	14.47 †	15.09 †
Race	34.11 †	35.66 †	35.47 †	17.44 †	16.09 †	13.76 †	12.98 †
Socioeconomic	26.16 †	49.42	44.77	19.19 †	18.6 †	13.95 †	13.37 †
Disability	23.33 †	46.67	40.0	11.67 †	6.67 †	8.33 †	8.33 †
Physical Appearance	25.4 †	47.62	47.62	30.16	20.63 †	12.7 †	14.29 †
Gender	33.59 †	41.22 †	39.69 †	32.82 †	26.34 †	19.47 †	18.7 †
Sexual Orientation	17.86 †	47.62	42.86	14.29 †	13.1 †	7.14 †	7.14 †
Age	24.14 †	49.43 †	48.28 †	24.14 †	18.39 †	17.24 †	19.54 †
AA	34.46 †	35.81 †	37.16 †	14.53 †	14.19 †	9.12 †	8.78 †

Table 17: Bias scores for CRR, CRRA, ΔP, ΔPA, CSPS, AUL and AULA measures using Equation 8 and distilBERT$_R$, where † indicates that the relative difference in proportions of bias between re-trained and pre-trained transformers is statistically significant according to McNemar's test.

S Significance Results for the Difference Between Measure Means with a Two-Tailed Welch's t-test

Race Bias	$f = \textbf{CRR}(s)$		$f = \textbf{CRRA}(s)$		$f = \Delta\textbf{P}(s)$		$f = \Delta\textbf{PA}(s)$	
Model	**CPS**	**SS**	**CPS**	**SS**	**CPS**	**SS**	**CPS**	**SS**
BERT$_{P,\text{unc}}$	0.016	0.011	0.004	0.009	0.111	0.168 †	0.006	0.011
RoBERTa$_P$	0.023 †	0.017	0.004 †	0.007 †	0.14 †	0.217 †	0.006 †	0.01 †
distilBERT$_{P,\text{unc}}$	0.013	0.008	0.004	0.01 †	0.053	0.166 †	0.003	0.011 †
distilRoBERTa$_P$	0.024 †	0.014	0.005	0.007 †	0.113 †	0.166 †	0.005 †	0.008 †

Table 18: $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{adv}}) - f(S_i^{\text{dis}})$; The mean difference in measure f between sentence sets S^{adv} and S^{dis} for pre-trained transformers. † indicates that the difference between the means $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{adv}})$ and $\frac{1}{N}\sum_{i=1}^{N} f(S_i^{\text{dis}})$ for a transformer is statistically significant according to the two-tailed Welch's t-test (p-value < 0.05), where N is the total number of sentences and equal across sentence sets S^{adv} and S^{dis} within bias categories on CPS and SS datasets.

T Example Measure Scores for Re-trained MLMs

Model	**Set**	**Example** s	**CRR**(s)	**CRRA**(s)	Δ**P**(s)	Δ**PA**(s)
distilBERT$_{P,\text{unc}}$	S^{dis}	he was a *black* man with a bad temper	0.388	0.096	0.915	0.044
distilBERT$_{P,\text{unc}}$	S^{adv}	he was a *white* man with a bad temper	**0.314**	**0.092**	**0.881**	**0.042**
distilBERT$_{R,\text{unc}}$	S^{dis}	he was a *black* man with a bad temper	**<0.001**	**0.044**	**<0.001**	**<0.001**
distilBERT$_{R,\text{unc}}$	S^{adv}	he was a *white* man with a bad temper	0.093	0.053	0.295	0.013

Table 19: CRR(s), CRRA(s), ΔP(s) and ΔPA(s) likelihood scores for an example instance in the CPS dataset showing a shift in MLM preference (measure values closer to 0) indicating greater relative bias against disadvantaged groups after re-training.