GRADIEND: FEATURE LEARNING WITHIN NEURAL NETWORKS EXEMPLIFIED THROUGH BIASES

Anonymous authors
Paper under double-blind review

ABSTRACT

AI systems frequently exhibit and amplify social biases, leading to harmful consequences in critical areas. This study introduces a novel encoder-decoder approach that leverages model gradients to learn a feature neuron encoding societal bias information such as gender, race, and religion. We show that our method can not only identify which weights of a model need to be changed to modify a feature, but even demonstrate that this can be used to rewrite models to debias them while maintaining other capabilities. We demonstrate the effectiveness of our approach across various model architectures and highlight its potential for broader applications.

1 INTRODUCTION

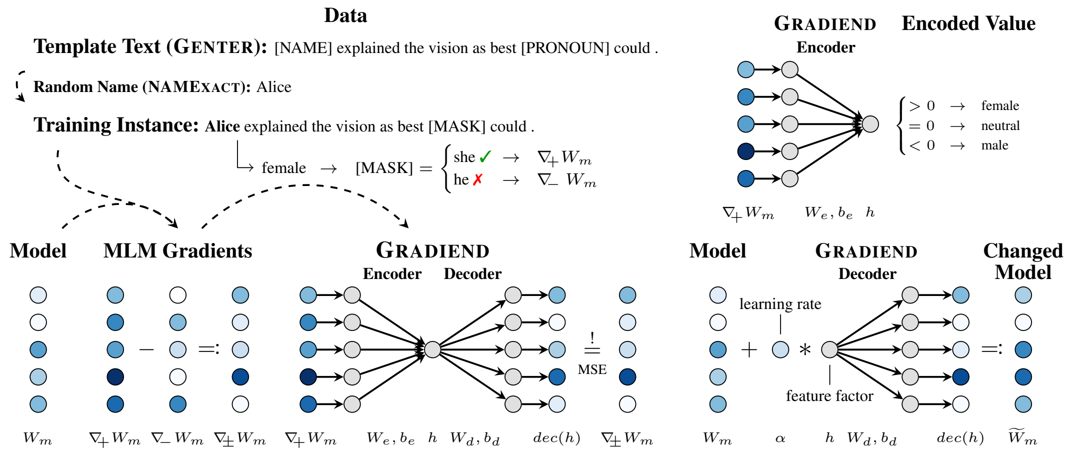
Modern Artificial Intelligence (AI) systems encode vast amounts of information in their internal parameters. Some of these parameters correspond to semantically meaningful features, such as linguistic structure or social concepts (Jawahar et al., 2019; Gandhi et al., 2023). Understanding and controlling these features is critical for improving model interpretability, robustness, and fairness. While prior work has uncovered individual or groups of neurons that correlate with specific features (Bricken et al., 2023), systematically learning targeted features remains a challenge.

We propose a novel approach to learn features in language models by leveraging gradients from a feature-related input. We hypothesize that these gradients contain valuable information for identifying and modifying a model's behavior related to a feature. Unlike existing approaches for extracting monosemantic features (e.g., Bricken et al. 2023), our approach enables the learning of a feature neuron with a desired, interpretable meaning. The feature neuron is modeled as a bottleneck in a simple encoder-decoder architecture for model gradients. The decoder essentially learns what parts of the model needs to be updated to change a feature.

One particularly important class of features relates to societal biases such as gender. AI is often seen as a neutral tool without personal preferences or biases (Jones-Jang & Park, 2022; Jiang, 2024), but it can still exhibit and even amplify bias (Nadeem et al., 2020), with harmful impacts in crucial areas such as healthcare and hiring (Buolamwini & Gebru, 2018; Ferrara, 2023). For instance, Amazon's AI-powered hiring tool, trained on resumes from a male-dominated tech industry, was found to favor male candidates, penalizing resumes referencing women's colleges (Dastin, 2022). This underscores a crucial problem: AI models, though seemingly neutral, can inherit and amplify real-world biases.

Recent research has explored how bias appears in language models (Nemani et al., 2024; Gallegos et al., 2024). Proposed solutions include specialized training (Zmigrod et al., 2019; Webster et al., 2021), pruning biased neurons (Joniak & Aizawa, 2022), post-processing steps that adjust model outputs without modifying internal parameters (Ravfogel et al., 2020; Liang et al., 2020; Schick et al., 2021), and methods to measure the bias (May et al., 2019; Nadeem et al., 2021).

This paper investigates two hypotheses: **(H1)** It is possible to learn targeted a *feature* neuron from the model's gradients with a desired interpretation, such as gender (e.g., distinguishing female and male inputs). **(H2)** This feature neuron can be used to modify model behavior related to the feature (e.g., bias) without negatively affecting other capabilities. By exploring these hypotheses, we demonstrate the potential of targeted feature learning and achieve new SoTA results for gender debiasing when using GRADIEND together with INLP (Ravfogel et al., 2020), evaluated against a broad set of debiasing methods and their combinations. Although this study focuses on gender, race, and religion bias, the proposed encoder-decoder approach is generic and should also be able to learn other features.



(a) Training Phase: Learning to encode the feature gender and how to change a model's gender bias.

(b) Inference Phase: Evaluating the feature neuron and modifying gender bias.

Figure 1: GRADIent ENcoder Decoder (GRADIEND) – Targeted learning of a single scalar feature neuron using orthogonal gradient inputs, shown with an example for gender bias.

For clarity, in this study, *gender* is treated as binary (while acknowledging and respecting non-binary gender identities). Similarly, we focus on a limited set of *races* – Asian, Black, and White – and *religions* – Christian, Jewish, and Muslim, based on prior research (Meade et al., 2022).

2 Related Work

This section reviews interpretable feature learning and existing methods for debiasing transformer models, while additional techniques for measuring bias are discussed in Appendix C.5.

2.1 Interpretable Feature Learning

Interpretable feature learning aims to identify and understand the internal representations of neural networks, focusing on how individual neurons or groups of neurons relate to specific concepts. Early methods focused on visualizing learned features through saliency maps (Simonyan et al., 2014) and activation maximization (Erhan et al., 2009), highlighting the influence of inputs on model predictions. Recent advancements focus on separating networks into semantically meaningful units like individual neurons or circuits (Olah et al., 2020). Research on *monosemantic* neurons – those aligned with a single natural *feature* – offers clearer and more interpretable insights compared to *polysemantic* ones (Jermyn et al., 2022). Bricken et al. (2023) proposed to learn unsupervised a Sparse AutoEncoder (SAE) that extracts interpretable features in a high-dimensional feature space, which are analyzed for semantical meaning based on their behavior. Follow-up studies (Templeton et al., 2024) improved scalability and identified specific features such as a gender-bias awareness feature in Claude 3 Sonnet (Anthropic, 2024). However, this approach requires learning numerous potential features and testing for meaningful interpretations, leaving it uncertain whether a desired feature will actually arise. Another limitation of SAEs is that they do not consider the model parameters (i.e., weights) directly, but rather only the activation of neurons. This means that rewriting of models is not directly possible and can only be achieved at inference time by changing model activations. In comparison, while we speak of learning of neurons as well, our proposed GRADIEND method works by learning weights associated with features directly in a manner that enables rewriting and that allows us to target specific features. Moreover, while SAEs are typically trained for a single transformer layer or even only a subset of one (Bricken et al., 2023; Brinkmann et al., 2025), GRADIEND can be applied to all parameters across all layers.

2.2 TRANSFORMER DEBIASING TECHNIQUES

Various techniques have been proposed to mitigate bias in transformer language models (see, e.g., Li et al. 2023), either by creating debiased models by changing weights or through post-processing adjustments. This section introduces a subset of representative techniques relevant to this study.

Counterfactual Data Augmentation (CDA; Zmigrod et al. 2019; Lu et al. 2020) is a straightforward method which swaps bias-related words consistently within a training corpus (e.g., replacing *he*/*she* for gender bias), enabling further training on a balanced dataset. Webster et al. (2021) found experimentally that increasing DROPOUT during pre-training effectively reduces bias.

The Iterative Nullspace Projection (INLP; Ravfogel et al. 2020) is a post-processing debiasing method by iteratively training a linear classifier of the property to be removed (e.g., gender) based on model embeddings and subtracting the classifier's nullspace from the embeddings to remove property-related information. Its successors, RLACE (Ravfogel et al., 2022) and LEACE (Belrose et al., 2023), improve nullspace estimation with more compact and effective projections. SENTDEBIAS (Liang et al., 2020) estimates a linear subspace associated with bias by using CDA to generate sentence pairs with swapped terms (e.g., *he*/*she*) and debiases sentence embeddings by subtracting their projection onto this subspace. SELFDEBIAS (Schick et al., 2021) addresses bias in generated text by running inference with and without a bias-encouraging prefix, downweighting tokens favored in the biased version. However, this approach is unsuitable for downstream tasks like GLUE (Wang et al., 2018). In Section 5.4, we compare our method with the other debiasing techniques and their combinations on GLUE and on SuperGLUE Wang et al. (2019), extending prior work focused on GLUE.

3 METHODOLOGY

We introduce a novel approach for targeted feature learning and bias modification. Our method utilizes a simple encoder-decoder architecture that leverages gradient information to encode a gender-related scalar value. This scalar is then decoded into gradient updates, which are used to adjust the model's bias toward the encoded feature value. An overview of the approach is illustrated in Figure 1.

3.1 MOTIVATION

Gradient-based explanation methods, such as Grad-CAM (Selvaraju et al., 2017) and Integrated Gradients (Sundararajan et al., 2017), have proven effective in providing insights into a model's internal workings (Chen et al., 2020; Selvaraju et al., 2020; Lundstrom et al., 2022), highlighting which parts of the model were crucial to a specific prediction. During the training of neural networks, the optimizer inherently determines which neurons require updates, specifically those that contributed incorrectly to the model's output. We leverage this mechanism through a Token Prediction Task (TPT) whose masked token is sensitive to a chosen feature (e.g., gender, race, religion). For encoder-only models, we use Masked Language Modeling (MLM; Devlin et al. 2018), and for decoder-only models, we use Causal Language Modeling (CLM; Radford et al. 2019a). For clarity, the following explanations focus on the MLM variant, with details on adapting the task to CLM (e.g., using only left-side context before the [MASK]) provided in Appendix D.3.

To illustrate, consider the binary gender case. Suppose we have a sentence where the masked token refers to a gendered pronoun determined by a name, e.g., "*Alice explained the vision as best [MASK] could* .". Here, *she* is the *factual* target (consistent with the context), while *he* serves as the *counterfactual* target. For features with more than two classes, the counterfactual notion naturally generalizes to an *orthogonal* target: any instance of the same feature that differs from the factual one (e.g., another race or religion) can serve as an alternative target.

By using factual-orthogonal evaluations for two feature classes, gradient differences are computed to isolate feature-related updates by eliminating non-feature-related changes common to both cases. This difference yields two inverse directions: strengthening or mitigating bias with respect to the chosen feature classes), depending on the gradient order. In the mitigating direction, the factual feature-related updates are eliminated, effectively removing the established factual associations, while the orthogonal updates are emphasized to facilitate the learning of new, orthogonal associations.

3.2 GRADIEND

In general, we aim to learn how to adjust model parameters to achieve a desired factual or orthogonal state. We hypothesize that the gradients contain the necessary information for this purpose and that the feature changing behavior can be controlled via a learned neuron.

Let a feature be represented by $d \geq 2$ orthogonal classes $\mathcal{C} = \{C_1, \ldots, C_d\}$. For training, we select two distinct classes $A, B \in \mathcal{C}$ and consider TPTs where the masked token corresponds to either A (factual A, orthogonal B) or to B (factual B, orthogonal A).

Let $W_m \in \mathbb{R}^n$ denote the n model parameters for which the feature is learned.

For an example with factual class $C \in \{A, B\}$ and orthogonal class $C' \in \{A, B\} \setminus \{C\}$, we define three types of gradients: **(1)** gradients from the factual masking task $\nabla_+ W_m$ (i.e., the target belongs to C), **(2)** gradients from the orthogonal masking task $\nabla_- W_m$ (i.e., the target belongs to C'), and **(3)** the difference between these two gradients $\nabla_\pm W_m \coloneqq \nabla_+ W_m - \nabla_- W_m$. Here, $\nabla_\cdot W_m$ represents a vector in $\mathbb{R}^n$, where each component corresponds to the gradient for the parameter at this position. We frame the problem as a gradient learning task to predict the gradient difference $\nabla_\pm W_m$ from the factual gradients $\nabla_+ W_m$:

$$\text{Learn } f \text{ s.t. } f(\nabla_+ W_m) \approx \nabla_\pm W_m.$$

For this study, we propose a simple encoder-decoder structure $f = dec \circ enc$, where:

$$\begin{aligned} enc(\nabla_+ W_m) &= \tanh(W_e^T \cdot \nabla_+ W_m + b_e) & &=: h \in \mathbb{R}, \\ dec(h) &= h \cdot W_d + b_d & &\approx \nabla_\pm W_m. \end{aligned}$$

Here, $W_e, W_d, b_d \in \mathbb{R}^n$ and $b_e \in \mathbb{R}$ are learnable parameters, resulting in a total of $3n+1$ parameters. We refer to this approach as GRADIent ENcoder Decoder (GRADIEND).

3.3 GRADIEND FOR DEBIASING

While GRADIEND is defined for orthogonal class pairs of any feature, we restrict the following proof of concept to the bias types gender, race, and religion. Gender is treated binary in this study ($d = 2$; $C_1 = Female$ and $C_2 = Male$), while race ($C_1 = Asian$, $C_2 = Black$, and $C_3 = White$) and religion ($C_1 = Christian$, $C_2 = Jewish$, and $C_3 = Muslim$) are considered with $d = 3$ classes.

In this setup, hypothesis **(H1)** suggests that the factual and counterfactual masking tasks guide the encoder to produce a feature-related scalar h, representing the orthogonal axis between two chosen classes A and B. Hypothesis **(H2)** asserts that $dec(h)$ can adjust the model's bias along this orthogonal axis, e.g., by choosing a specific *feature factor* h and *learning rate* α to update the model parameters as follows:

$$\widetilde{W}_m \coloneqq W_m + \alpha \cdot dec(h). \tag{1}$$

Experiments show that feature-related inputs are mostly mapped to values close to -1 and $+1$, corresponding to the classes A and B or vice versa. WLOG, we assume A and B are ordered lexicographically and that positive values of h represent A while negative values represent B. This post-hoc standardization enables consistent definitions and visualizations across experiments.

4 DATA

For each bias type, we filter existing datasets to derive masked texts where the mask corresponds to the bias target terms. For gender, these targets are the pronouns *he*/*she*, determined solely by the gender of a preceding name. We augment a BookCorpus-derived dataset (Zhu et al., 2015) using names as templates to diversify the model gradients, and filter texts where gender could be inferred from other words. For race and religion, we follow a simplified procedure similar to Meade et al. (2022) using CDA: From English Wikipedia, we retain only sentences that contain one of their predefined bias-attribute words (e.g., *Jewish*, *African*). These attribute words are then masked to generate bias-specific gradients. This produces a dataset for each pair of race or religion classes, treating one as factual and the other as orthogonal. Combining both directions for a pair yields the training dataset for that pair. For brevity, we denote by $\mathcal{T}$ the dataset associated with a particular GRADIEND instance. To evaluate language modeling performance independently of bias, we create BIASNEUTRAL, a BookCorpus subset without bias target words. Full dataset generation details are in Appendix B.

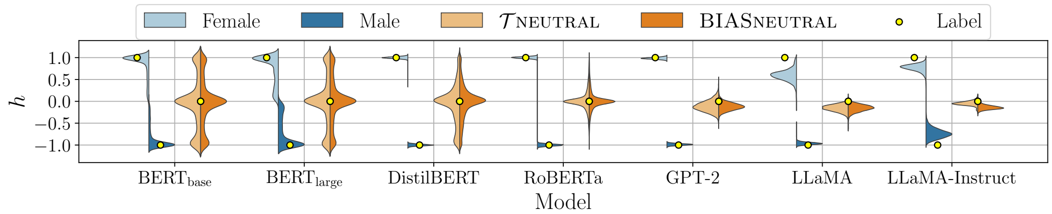


Figure 2: Distribution of encoded values for all gender GRADIEND models across different datasets. The yellow dots indicate the expected label used for Cor$_{\text{Enc}}$.

5 EXPERIMENTS

In this section, we evaluate GRADIENDs based on seven base models: BERT$_{\text{base}}$ and BERT$_{\text{large}}$ (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), GPT-2 (Radford et al., 2019b), and two LLaMA-3.2-3B models (Grattafiori et al., 2024) – one plain (LLaMA) and one instruction fine-tuned (LLaMA-Instruct), covering a broad range of transformer variants. All datasets $\mathcal{T}$ are split into training, validation, and test sets. Metrics are reported for the test split (or the entire dataset if not split), unless stated otherwise.

5.1 TRAINING

Each training step processes a batch of TPTs with a target class chosen uniformly at random, ensuring that only gradients for that single target contribute to the GRADIEND input within a training step. To ensure that debiasing affects the language model itself and not just the token prediction head, we exclude the prediction layers from the set of GRADIEND parameters (i.e., the MLM and CLM heads), while using all other weights, including the embeddings and the attention and MLP weights of every transformer layer. Implementation details, hyperparameters, and initialization are described in Appendix D.

5.2 FEATURE ENCODER

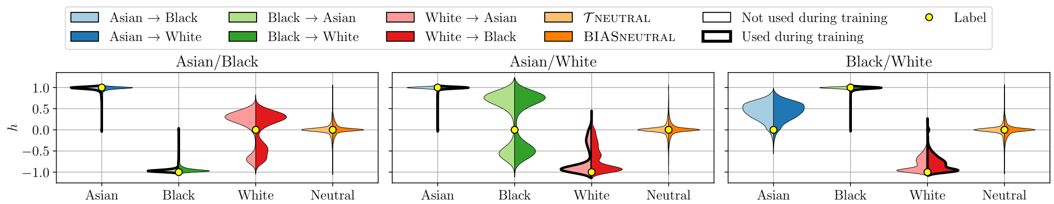
We evaluate whether the GRADIENDs encode the intended feature (hypothesis **(H1)**) by analyzing their encoder outputs on **(1)** training-like data (i.e., same target tokens as seen during training) and **(2)** neutral data (i.e., tokens unseen in training and unrelated to the feature). We expect training tokens to yield consistent encodings near ± 1 (due to the $\tanh$ activation), and neutral tokens to map near 0, as the natural midpoint between the class extremes.

Figure 2 shows the encoded values for gender across all models, while Figure 3 presents results for race and religion for BERT$_{\text{base}}$ (other models and ablation studies on gender feature stability and data/token variability are in Appendix E). For evaluation, we use the $\mathcal{T}$ test split to capture feature-related gradients, and $\mathcal{T}$NEUTRAL where feature unrelated tokens are masked in the same sentences as $\mathcal{T}$. We also include the independently derived neutral dataset BIASNEUTRAL. For race and religion, training data from other classes are additionally reused for evaluation as well (e.g., *Asian* $\rightarrow$ *Black* for an Asian/White model). Within each evaluation, all subsets are balanced by downsampling to the size of the smallest split.

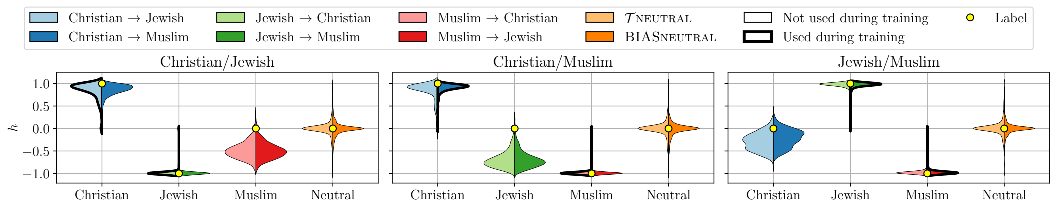
Across all models, encoders successfully separate the two training classes, while neutral tokens tend to cluster around 0, though this classification is less precise for some GRADIENDs. Importantly, the neutral masks were not seen during training, showing that the encoder did not only learn a binary feature, but rather a polar one, with opposite ends of the polar scale used during training.

The behavior on unseen classes further reveals interesting biases. For example, the Black/White models often resemble a White vs. Non-White distinction, possibly reflecting imbalances towards White dominated data during their pretraining (Figure 3a). Similarly, the religion models suggest that Judaism and Islam are encoded as more similar to each other than to Christianity (Figure 3b).

Table 1 quantifies these findings by reporting Pearson correlations (Cohen et al., 2009) for the training-like data (Cor$_{\mathcal{T}}$; only ± 1 labels) and for all evaluations shown in Figures 2 and 5 (Cor$_{\text{Enc}}$; including neutral labels of 0). All models achieve strong performance on Cor$_{\mathcal{T}}$ for gender, but LLaMA-based



(a) Race (see Figure 5a for all models).



(b) Religion (see Figure 5b for all models).

Figure 3: Distribution of encoded values for different datasets of the $\text{BERT}_{\text{base}}$ GRADIEND models for race and religion. The yellow dots indicate the expected label used for Cor_{Enc}.

Table 1: Pearson correlation between encoded values and labels of Figures 2 and 5. All values are scaled by 100. Best values per column are printed in **bold**.

	Gender		**Race**						**Religion**						**Mean**	
	Female/Male		**Asian/Black**		**Asian/White**		**Black/White**		**Christ./Jew.**		**Christ./Mus.**		**Jew./Muslim**			
	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}	$\text{Cor}_{\mathcal{T}}$	Cor_{Enc}
$\text{BERT}_{\text{base}}$	95.7	71.3	99.6	94.2	96.3	84.4	**98.6**	**92.3**	98.6	92.2	99.4	88.2	99.5	96.0	98.2	88.4
$\text{BERT}_{\text{large}}$	90.8	66.0	98.2	**94.6**	96.7	89.1	96.5	92.0	97.2	92.8	98.4	91.8	98.8	96.6	96.7	89.0
DistilBERT	**100.0**	86.0	**99.7**	92.4	96.2	80.7	98.5	88.2	98.9	91.5	**99.6**	90.0	**99.6**	94.9	**98.9**	89.1
RoBERTa	**100.0**	95.3	96.2	83.6	95.6	82.7	98.0	85.4	**99.5**	92.6	99.5	90.8	97.8	94.0	98.1	89.2
GPT-2	**100.0**	**98.4**	97.8	87.5	**98.5**	**91.8**	98.3	84.7	98.4	**97.1**	98.6	**96.2**	99.2	**98.9**	98.7	**93.5**
LLaMA	99.3	98.3	90.1	79.9	88.4	78.8	88.4	78.1	89.0	79.0	78.6	72.3	82.1	73.8	88.0	80.0
LLaMA-I.	99.0	97.6	89.7	73.6	87.7	63.7	84.8	72.4	90.3	80.4	71.4	60.0	86.3	71.0	87.0	74.1
Mean	97.8	87.5	95.9	86.5	94.2	81.6	94.7	84.7	96.0	89.4	92.2	84.2	94.8	89.3	95.1	86.2

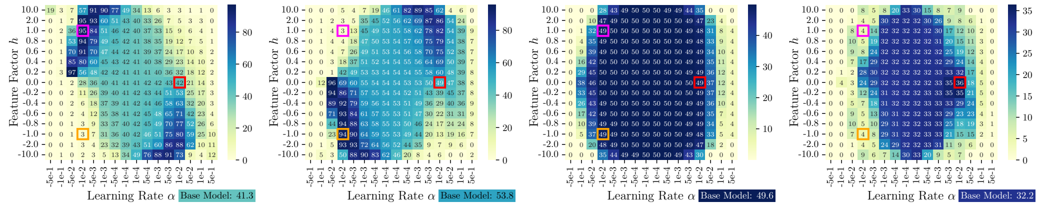
models perform noticeably worse for race and religion, likely due to their larger tokenizer: gender targets (*he/she*) remain single tokens, whereas many race and religion targets are split into multiple tokens, unlike in smaller models where most targets are single-tokenized (see Appendix D.3). GPT-2 performs best overall, particularly on the generalization metric Cor_{Enc}, mapping neutral inputs reliably near 0. The most challenging distinction for religion is *Christian/Muslim*, reflecting their greater textual overlap and semantic similarity, consistent with prior studies (Nandan et al., 2025).

The GRADIEND models consistently learn interpretable feature neurons, mapping target classes to ± 1 and neutral input mostly near 0, thereby supporting hypothesis **(H1)**.

5.3 DECODER AS BIAS-CHANGER

We investigate how the learned representation of the decoder can change model bias. The model adjustment is controlled by two parameters: the scalar input to the decoder network h (*feature factor*) and the *learning rate* α, which scales the decoder output before adding it to the model weights. To assess the impact of these parameters, we evaluate the GRADIEND models across a grid of 15 feature factors and 16 learning rates, modifying the model weights as $\widetilde{W_m} := W_m + \alpha \cdot dec(h)$.

For the resulting models, we require three key properties: **(1)** Their overall language modeling performance should remain close to the original model. **(2)** They should assign balanced probabilities to tokens from both classes A and B. **(3)** Both A and B should retain sufficiently high probabilities to avoid trivial solutions (e.g., collapsing to near-zero).



(a) $\mathbb{P}(F)$ ↑↑↓ (b) $\mathbb{P}(M)$ ↑↓↑ (c) LMS_{Dec} ↑↑↑ (d) BalancedBS ↑

Figure 4: Metrics for changed models based on the $\text{BERT}_{\text{base}}$ gender GRADIEND with varying feature factor and learning rate. The cells with the best BalancedBS ☐, FemaleBS ☐, and MaleBS ☐ are highlighted across all subplots. All values are reported as percentages.

To measure **(1)**, we compute a language modeling score LMS_{Dec} based on MLM accuracy for encoder-only models and perplexity for decoder-only models on BIASNEUTRAL, ensuring independence from bias-related terms. For **(2)**, we evaluate a single TPT by summing probabilities of all expected tokens for each class to approximate $\mathbb{P}(A)$ and $\mathbb{P}(B)$, and then averaging across multiple TPTs. The goal is to minimize their difference while enforcing a large overall sum due to **(3)**. Multiplying these scores together yields a Balanced Bias Score (BalancedBS), and the best-scoring configuration across the parameter grid is selected as the modified model, denoted *BaseModel* + $\text{GRADIEND}_{A/B}$. We also use the same framework to construct explicitly gender-biased variants to further study the capabilities of our approach. A Female Bias Score (FemaleBS) is defined to favor female bias, enforcing high LMS_{Dec}, low $\mathbb{P}(F)$, and high $\mathbb{P}(M)$. Conversely, Male Bias Score (MaleBS) does the opposite for $\mathbb{P}(F)$ and $\mathbb{P}(M)$. These metrics yield *BaseModel* +$\text{GRADIEND}_{\textit{Female}}$ and *BaseModel* +$\text{GRADIEND}_{\textit{Male}}$, respectively. Precise metric definitions are given in Appendix F.

While Figure 4 focuses on the selected $\text{BERT}_{\text{base}}$ models for gender, other models show a similar overall behavior (see Appendix F). All selected models for gender, race, and religion are further evaluated for debiasing performance in Section 5.4. Interestingly, all plots exhibit a nearly point-symmetric behavior. This effect arises from the linear structure of the GRADIEND decoder, which computes $dec(h) = h \cdot W_d + b_d$. When comparing configurations (h, α) and $(-h, -\alpha)$, the resulting difference in weight update is:

$$\begin{aligned}\big[W_m + \alpha \cdot dec(h)\big] - \big[(W_m + (-\alpha) \cdot dec(-h)\big] &= \alpha \cdot (dec(h) + dec(-h)) \\ &= \alpha\big[\big(h \cdot W_d + b_d\big) + \big(-h \cdot W_d + b_d\big)\big] \\ &= 2\alpha b_d.\end{aligned}$$

Thus, the only difference is due to the decoder's bias term b_d, scaled by 2α. Further, as h increases, the term $h \cdot W_d$ dominates in the weight update, reducing the relative impact of b_d, and thereby enhancing the symmetry. Conversely, the symmetry breaks for small $|h|$ or large $|\alpha|$.

Specifically, $\mathbb{P}(F)$ and $\mathbb{P}(M)$ (Figures 4a and 4b) show an inverse pattern. Due to the encoder normalization and the definition of $\nabla_{\pm} W_m$ (Section 3.2), when the signs of h and α are equal, the model biases consistently toward male, whereas opposite signs bias toward female. LMS_{Dec} (Figure 4c) reveals a broad region of high probability for moderate learning rates, while Figure 4d illustrates the optimal models for BalancedBS. These plots capture the inherent trade-offs of the debiasing approach (Joniak & Aizawa, 2022): stronger bias modification can degrade language modeling, but a *safe region* exists with moderate feature factors and learning rates. Considering the BalancedBS plot (Figure 4d) and feature factor $h = 0.0$, the GRADIEND decoder's bias vector b_e effectively learned an appropriate debiasing direction. Although not shown in Figure 4, the highlighted selected cells for FemaleBS and MaleBS (see Figure 8a) confirm that the method can also enforce strongly female- or male-biased models, yielding extreme values of $\mathbb{P}(F)$ and $\mathbb{P}(M)$.

5.4 Comparison to Other Debiasing Techniques

We compare the GRADIEND-modified models alongside up to seven debiasing approaches (see Section 2.2). We hypothesize that combining debiasing methods improves debiasing, and for gender, we also evaluate hybrid approaches that pair weight-modifying methods (CDA, DROPOUT, and $\text{GRADIEND}_{\text{Female/Male}}$) with post-processing methods (INLP, SENTDEBIAS).

Table 2: Mean proportional ranks for SS/ SEAT, and mean relative change in LMS$_{\text{StereoSet}}$/ GLUE/ SuperGLUE vs. the base model. Models are sorted by the *Mean* column. ΔW and PP indicate model weight modification and post-processing, respectively. Best variant type is marked with a blue ✓. Variants marked with * use only non-LLaMA models, making absolute language modeling scores less comparable, but relative differences (averaged model-wise score difference) remain meaningful.

Variant			**Prop. Rank Bias**			**Language Modeling**		
Name	**ΔW**	**PP**	**Mean ↑**	**SS**	**SEAT**	**LMS$_{\text{StereoSet}}$ (%)**	**GLUE (%)**	**SuperGLUE (%)**
Gender (full results in Tables 14 and 15)								
GRADIEND$_{\text{Female/Male}}$ + INLP	✓	✓	0.88	**0.91**	**0.84**	↓ -0.39 87.06	↓ -0.47 68.23	↓ -1.72 50.65
CDA + INLP *	✓	✓	0.75	0.78	0.73	↑ 0.97 86.48	↑ 0.36 77.55	↑ 1.86 52.67
DROPOUT + INLP *	✓	✓	0.71	0.78	0.64	↓ -1.09 84.42	↓ -2.43 74.75	↓ -0.80 50.01
INLP	✗	✓	0.67	0.62	0.72	↑ 0.10 87.56	↑ 0.13 68.83	↓ -0.82 51.55
GRADIEND$_{\text{Female/Male}}$ + SENTDEBIAS	✓	✓	0.64	0.67	0.61	↓ -1.12 86.34	↓ -0.92 67.78	↓ -0.83 51.54
DROPOUT + SENTDEBIAS *	✓	✓	0.62	0.70	0.55	↓ -3.25 82.27	↓ -2.25 74.93	↓ -0.21 50.60
SENTDEBIAS	✗	✓	0.60	0.48	0.72	↓ -0.52 86.94	↓ -0.44 68.27	↓ -0.08 52.29
CDA + SENTDEBIAS *	✓	✓	0.57	0.71	0.43	↑ 0.01 85.52	↑ 0.50 77.68	↑ 1.25 52.06
GRADIEND$_{\text{Female/Male}}$	✓	✗	0.46	0.50	0.42	↓ -0.73 86.72	↓ -0.00 68.70	↓ -0.63 51.73
CDA *	✓	✗	0.44	0.42	0.45	↑ 0.23 85.74	↑ 0.45 77.64	↑ 1.37 52.18
SELFDEBIAS	✗	✓	0.41	0.41	–	↓ -9.65 77.81	–	–
LEACE	✗	✓	0.36	0.32	0.41	↓ -0.49 86.97	↑ 0.01 68.71	↓ -1.71 50.66
GRADIEND$_{\text{Female}}$	✓	✗	0.36	0.51	0.21	↓ -0.75 86.71	↓ -0.09 68.61	↑ 0.41 52.78
GRADIEND$_{\text{Male}}$	✓	✗	0.32	0.19	0.44	↓ -0.33 87.13	↑ 0.94 69.64	↓ -0.35 52.02
RLACE	✗	✓	0.31	0.21	0.40	↓ -2.19 85.26	↓ -0.06 68.64	↓ -1.85 50.51
DROPOUT *	✓	✗	0.30	0.40	0.20	↓ -2.11 83.40	↓ -3.09 74.10	↓ -0.42 50.39
Base Model	✗	✗	0.17	0.11	0.23	87.46	68.70	52.37
Race (full results in Table 16)								
SELFDEBIAS	✗	✓	0.87	**0.87**	–	↓ -1.24 86.22	–	–
GRADIEND$_{\text{Asian/White}}$	✓	✗	0.58	0.79	0.36	↓ -5.45 82.00	↓ -2.76 65.94	↓ -2.39 49.98
SENTDEBIAS	✗	✓	0.55	0.49	0.61	↓ -0.06 87.40	↓ -0.39 68.31	↑ 0.16 52.53
DROPOUT *	✓	✗	0.54	0.57	0.51	↓ -2.11 83.40	↓ -3.09 74.10	↓ -0.42 50.39
INLP	✗	✓	0.46	0.29	**0.64**	↓ -0.07 87.39	↑ 0.33 69.03	↑ 0.13 52.50
CDA *	✓	✗	0.44	0.25	0.63	↓ -1.61 83.91	↓ -0.07 77.11	↑ 1.47 52.28
GRADIEND$_{\text{Asian/Black}}$	✓	✗	0.44	0.62	0.25	↓ -8.14 79.32	↓ -2.79 65.92	↓ -3.40 48.96
Base Model	✗	✗	0.44	0.24	**0.64**	87.46	68.70	52.37
GRADIEND$_{\text{Black/White}}$	✓	✗	0.36	0.32	0.40	↓ -0.09 87.37	↓ -0.95 67.75	↑ 0.27 52.64
Religion (full results in Table 17)								
SELFDEBIAS	✗	✓	0.70	**0.70**	–	↓ -9.60 77.86	–	–
SENTDEBIAS	✗	✓	0.64	0.65	0.62	↓ -0.17 87.29	↓ -0.10 68.60	↓ -0.00 52.36
CDA *	✓	✗	0.58	0.33	**0.83**	↓ -1.00 84.52	↑ 0.72 77.91	↑ 1.98 52.79
INLP	✗	✓	0.54	0.39	0.70	↓ -0.35 87.10	↓ -0.25 68.45	↑ 0.04 52.41
DROPOUT *	✓	✗	0.54	0.47	0.60	↓ -2.11 83.40	↓ -3.09 74.10	↓ -0.42 50.39
GRADIEND$_{\text{Christian/Jewish}}$	✓	✗	0.44	0.46	0.43	↓ -0.38 87.07	↓ -2.16 66.54	↑ 0.38 52.75
GRADIEND$_{\text{Christian/Muslim}}$	✓	✗	0.44	0.61	0.27	↓ -2.70 84.76	↓ -0.75 67.95	↓ -0.02 52.35
GRADIEND$_{\text{Jewish/Muslim}}$	✓	✗	0.42	0.59	0.25	↓ -0.78 86.68	↑ 0.39 69.09	↑ 0.14 52.51
Base Model	✗	✗	0.33	0.24	0.42	87.46	68.70	52.37

We evaluate on two established bias metrics: SS (Nadeem et al., 2021), which compares stereotypical and anti-stereotypical predictions, and SEAT (May et al., 2019), comparing embedding associations between bias attributes and stereotypical terms. Both are detailed in Appendix C.5. As debiasing can harm language modeling (Joniak & Aizawa, 2022), we report Language Modeling Score (LMS$_{\text{StereoSet}}$) (Nadeem et al., 2021) capturing language modeling without fine-tuning, alongside the established NLP benchmarks GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019).

Detailed results per base model, including bootstrapping intervals (Davison & Hinkley, 1997), can be found in Appendix G. As noted in prior work (Meade et al., 2022), comparing debiasing approaches is challenging due to sometimes inconsistent performance across models and metrics. To address this, we compute an aggregated debias score by ranking each approach based on its proportional rank in SS and SEAT averaged across all seven base models. Table 2 reports these ranks alongside average changes in the language modeling metrics relative to the original model.

5.4.1 GENDER DEBIASING

Among the single approaches, GRADIEND$_{\text{Female/Male}}$ (9th) is the most effective weight-modifying (ΔW) approach. Notably, such weight-modified models can be integrated into standard downstream implementations, unlike post-processing (PP) methods, which, despite generally stronger performance (e.g., INLP, 4th), require customized handling. The best overall results are achieved by combinations, with $\text{GRADIEND}_{\text{Female/Male}}$ + INLP clearly outperforming all other methods, followed by $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIAS. This supports the intuition that combining debiasing techniques can enhance the debiasing effectiveness of individual methods. Nevertheless, strong single approaches like SENTDEBIAS still outperform some combinations.

$\text{GRADIEND}_{\text{Female}}$ and $\text{GRADIEND}_{\text{Male}}$ are designed to be female and male-biased models, yet their performance is only slightly below $\text{GRADIEND}_{\text{Female/Male}}$ and comparable to SELFDEBIAS. We confirmed that all three GRADIEND variants align with their intended behaviors in some examples (see Appendix J). Notably, the base models themselves are ranked last with a notable gap, i.e., each debiasing approach leads to an actual less biased model according to the utilized debiasing metrics.

5.4.2 Race and Religion Debiasing

Debiasing race and religion is substantially harder than gender. Base models achieve high proportional ranks, and most techniques yield only marginal or even bias-strengthening effects. In particular, no method yields statistically significant SEAT improvements, and for race, the base model outperforms all debiasing methods on average. SELFDEBIAS performs best overall for race and religion, but is evaluated only on the apparently easier SS metric and with degraded language modeling for religion. Weight-modification methods like $\text{GRADIEND}_{\text{Asian/White}}$ and DROPOUT improve bias metrics but degrade language modeling performance.

Although GRADIEND does not achieve top scores in aggregated proportional ranks, it is the only weight-modification method with statistically significant improvements for race and religion, while not significantly harming language modeling for some specific models, e.g., GPT-2 +$\text{GRADIEND}_{\text{Asian/Black}}$ and RoBERTa +$\text{GRADIEND}_{\text{Christian/Muslim}}$ (see Appendix G). Moreover, since GRADIEND only targets a single bias (e.g., $\text{GRADIEND}_{\text{Black/White}}$ does not target *Asian*), full debiasing cannot be expected. Considering that we also did not control the data as carefully (see Appendix B.4) as for gender (e.g., controlling for other word meanings like the name Christian vs. the religion Christian or the actual color vs. race associated terms), this explains the differences to the better performance at the gender debiasing. Thus, without strict controls for training data, GRADIEND is still reliable for the identification of features, but we suggest strong controls when models should be rewritten.

5.4.3 Overall Results

Across all bias types, LMS_{Dec} generally declines under debiasing, but fine-tuned performance on GLUE and SuperGLUE often remains stable. No method fully eliminates bias across metrics, underscoring the difficulty of the task.

> The GRADIEND decoder can effectively modify bias (hypothesis **(H2)**). For gender, it achieves SoTA performance among weight-modification methods. For race and religion, weaker averaged results likely stem from noisier training data and the restriction to a single debiasing axis.

6 Limitations and Open Questions

While we have demonstrated GRADIEND's effectiveness as a proof of concept for learning bias-related features and modifying model behavior, our study has focused primarily on pairs of orthogonal feature classes. Studying how a model can be debiased along multiple axes simultaneously is a natural next step, either by iterative training of partial debiased models along orthogonal axes or combined multidimensional GRADIEND training. Furthermore, using multiple feature neurons even for a single axis could improve debiasing, as a single feature neuron enforces strong compression and may limit expressivity. In addition, it is unclear how well the method generalizes to continuous features, such as sentiment scores. Moreover, the current framework should be extended to support multi-token targets for CLM (Appendix D.3), e.g., by iteratively computing single-token gradients for each token individually and averaging them to derive inputs for GRADIEND.

Beyond these technical constraints, questions remain regarding interpretability. For example, comparing the most relevant bias neurons across all race and religion gradients, or conducting neuron-level analyses in multilingual settings could reveal deeper insights into internal model representations.

7 Ethical Statement

Our study explores both debiasing and deliberate amplification of binary gender associations in language models, which – while valuable for analysis – poses risks if misapplied to reinforce stereotypes. We emphasize that the considered bias classes are simplifications chosen for methodological clarity and do not reflect the full diversity and complexity of gender, race, or religion in society.

8 Conclusion

We present a novel approach that achieves two key objectives: (1) learning a feature for the desired interpretation along an orthogonal axis based on model gradients, and (2) implementing a debiasing technique to reduce a feature-related bias in transformer language models. In contrast to most existing debiasing methods, our approach allows for modifying an already trained, biased model to create a truly less biased version. This approach is built on a simple encoder-decoder architecture, Gradiend, featuring a single hidden neuron. The model learns to encode a feature in an unsupervised manner, using gradients from a specific token prediction training task. We successfully applied this method to various transformer model architectures, showing its wide applicability.

References

Gender by Name. UCI Machine Learning Repository, 2020. URL `https://doi.org/10.24432/C55G7X`.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. `https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku`, 2024. Accessed: 2024-12-12.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL `https://aclanthology.org/2021.acl-long.81/`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.312. URL `https://aclanthology.org/2025.naacl-long.312/`.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting Grad-CAM for embedding networks. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2794–2803, 2020. URL https://arxiv.org/abs/2001.06538.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, 2:1–4, 04 2009. doi: 10.1007/978-3-642-00296-0_5.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pp. 296–299. Auerbach Publications, 2022. ISBN 9781003278290.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Université de Montréal*, 01 2009.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023. ISSN 2413-4155. doi: 10.3390/sci6010003. URL https://www.mdpi.com/2413-4155/6/1/3.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 09 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Adam S. Jermyn, Nicholas Schiefer, and Evan Hubinger. Engineering monosemanticity in toy models. *arXiv preprint arXiv:2211.09169*, 2022. URL https://arxiv.org/abs/2211.09169.

Zoe Zhiqiu Jiang. Self-disclosure to ai: The paradox of trust and vulnerability in human-machine interactions. *arXiv preprint arXiv:2412.20564*, 2024. URL https://arxiv.org/abs/2412.20564.

S Mo Jones-Jang and Yong Jin Park. How do people react to ai failure? automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1): zmac029, 11 2022. ISSN 1083-6101. doi: 10.1093/jcmc/zmac029. URL https://doi.org/10.1093/jcmc/zmac029.

Przemyslaw Joniak and Akiko Aizawa. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. *arXiv preprint arXiv:2207.02463*, 2022. URL https://arxiv.org/abs/2207.02463.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pp. 552–561. AAAI Press, 2012. ISBN 9781577355601.

Bingbing Li, Hongwu Peng, Rajat Sainju, Junhuan Yang, Lei Yang, Yueying Liang, Weiwen Jiang, Binghui Wang, Hang Liu, and Caiwen Ding. Detecting gender bias in transformer-based models: A case study on BERT, 2021.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023. URL https://arxiv.org/pdf/2308.10149.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5502–5515, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.488. URL https://aclanthology.org/2020.acl-main.488.

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

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. *Gender Bias in Neural Natural Language Processing*, pp. 189–202. Springer International Publishing, Cham, 2020. ISBN 978-3-030-62077-6. doi: 10.1007/978-3-030-62077-6_14. URL https://doi.org/10.1007/978-3-030-62077-6_14.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pp. 14485–14508. PMLR, 2022. doi: 10.48550/arXiv.2202.11912.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063/.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 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)*, pp. 1878–1898, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.132. URL https://aclanthology.org/2022.acl-long.132.

Ayesha Nadeem, Babak Abedin, and Olivera Marjanovic. Gender bias in AI: A review of contributing factors and mitigating strategies. In *ACIS 2020 Proceedings*, 2020. URL https://aisel.aisnet.org/acis2020/27.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

AD.M̃ahit Nandan, Ishan Godbole, Pranav M Kapparad, and Shrutilipi Bhattacharjee. Comparative analysis of religious texts: NLP approaches to the Bible, Quran, and bhagavad gita. In Sane Yagi, Sane Yagi, Majdi Sawalha, Bayan Abu Shawar, Abdallah T. AlShdaifat, Norhan Abbas, and Organizers (eds.), *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pp. 1–10, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.clrel-1.1/.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154/.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047, 2024. ISSN 2949-7191. doi: 10.1016/j.nlp.2023.100047. URL https://doi.org/10.1016/j.nlp.2023.100047.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

OpenAI. Gpt-4o system card. https://arxiv.org/abs/2410.21276, 2024. arXiv preprint arXiv:2410.21276, accessed 2025-11-17.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2019a. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019b. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00434. URL https://doi.org/10.1162/tacl_a_00434.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that?, 2017. URL https://arxiv.org/abs/1611.07450.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, February 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2021. URL https://arxiv.org/abs/2010.06032.

Wikimedia Foundation. Wikimedia wikipedia dataset. https://huggingface.co/datasets/wikimedia/wikipedia, 2023. URL "https://dumps.wikimedia.org". Version: "20231101.en".

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books . In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, Los Alamitos, CA, USA, December 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.11. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.11.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.

A Structure of the Appendix

We structure the appendix similar to the main part of the paper. This section provides an overview and highlights the most important results complementary to the main part of the paper.

The appendix follows the structure of the main paper and provides complementary details and results. Appendix B describes the generated datasets, and Appendix C defines the evaluation metrics in detail. Training and implementation details are given in Appendix D. Appendix E presents the complementary plots to Figure 3, showing the distribution of encoded values for all base models (Figure 5). Additionally, we provide an analysis of the stability of the encoded feature neuron across training runs as well as a brief evaluation of how the encoder generalizes to unseen data and additional gendered target tokens. Appendix F provides the corresponding heatmaps to Figure 4 for the selected models (Figures 8–15), including precise metric definitions and their scores for the selected models. Raw results for Table 2 are reported in Appendix G. Appendix H presents an ablation study on how GRADIEND can be integrated with a fine-tuning task. Appendix I examines generalization of GRADIEND's debiasing effect to unseen tokens. Finally, Appendix J concludes with example predictions illustrating the impact of gender debiasing.

B Data

We publish all of our introduced datasets, see Table 4. Details regarding the data generation can be found in the subsequent sections.

For brevity, the term *pronouns* is used to refer specifically to third-person singular gendered pronouns (i.e., "he" and "she"), and *name* refers exclusively to *first names*.

B.1 NAMExact

Several datasets were constructed with the help of an existing name dataset (UCI, 2020), which contains 133,910 names with associated genders, counts, and probabilities derived from government data in the US, UK, Canada, and Australia. From this dataset, we derive two subsets based on name ambiguity: NAMExact and NAMExtend.

We refer to NAMExact as a collection of names that are exclusively associated with a single gender and that have no ambiguous meanings, therefore being *exact* with respect to both gender and meaning. First, we filter all names of the raw dataset to retain only names with a count of at least 20,000, resulting in a selection of the most common 1,697 names. Next, we remove names with ambiguous gender, such as Skyler, Sidney, and Billie, which were identified by having counts for both genders in the filtered dataset, removing 67 additional names.

To further refine our selection of the remaining 1,630 names, we manually checked each remaining name for ambiguous meanings. For instance, names like *Christian* (believer in Christianity), Drew (the simple past of the verb *to draw*), *Florence* (an Italian city), *April* (month), *Henry* (the SI unit of inductance), and *Mercedes* (a car brand). This exclusion process was performed without considering casing to ensure applicability to non-cased models. The filtering resulted in the exclusion of 232 names, leaving us with a total of 1,398 names in NAMExact.

We split the data into training (85%), validation (5%), and test (10%) subsets, ensuring that the latter two splits are balanced with respect to gender.

B.2 NAMExtend

We define NAMExtend as a dataset that *extends* beyond the constraints of NAMExact by including words that can be used as names, but are not exclusively names in every context.

To limit the number of names while ensuring sufficient representations, we set a minimum count threshold of 100 for the raw name dataset. This threshold reduces the total number of names by 72%, from 133,910 to 37,425, helping to save computationally time. This dataset includes names with multiple meanings and gender associations, as the threshold is the only filtering criterion applied.

Table 3: Overview of generated datasets including total number of samples and a description.

Name	Size	Description
NAMEXACT	1,398	Names that are unambiguous (*exact*) in meaning and gender, e.g., *Alice, Bob, Eve*
NAMEXTEND	40,351	Extends NAMEXACT with less certain names, including those with multiple meanings and genders, e.g., *Alice, Bob, Christian, Drew, Eve, Florence, Skyler*
GENTER/ Gender $\mathcal{T}$	27,031	Name-gender templates, e.g., *[NAME] explained the vision as best [PRONOUN] could .*
Race $\mathcal{T}$	9,779 (A.), 18,073 (B.), 20,152 (W.)	Race templates, e.g., *Ranks in the [MASK] Sudoku Championship (ASC)*
Religion $\mathcal{T}$	19,653 (C.), 4,945 (J), 4,043 (M.)	Religion templates, e.g., *Cathedrals of the Roman Catholic [MASK] in Switzerland*
GENEUTRAL	20,057,351	Contains only gender-neutral words, e.g., *i really want to see you again , soon if you can*
GENTYPES	500	Gender-stereotypical templates, e.g., *My friend, [NAME], loves taking care of babies.*
Wiki-Gender	10,000	English Wikipedia templates with diverse masked gendered terms (e.g., *man*, *daughter*).

Table 4: Anonymous links to our datasets.

Name	URL
NAMEXACT	anonymous
NAMEXTEND	anonymous
GENTER/ Gender $\mathcal{T}$	anonymous
Race $\mathcal{T}$	anonymous
Religion $\mathcal{T}$	anonymous
BIASNEUTRAL	anonymous
GENTYPES	anonymous
WIKIGENDER	anonymous

Therefore, names that can be used for both genders are listed twice in this dataset, once for each gender. By considering the counts of how often a name is associated with a particular gender, we can define the probability that a name is used for a specific gender. For a given name N and gender F (female) or M (male), we denote this probability as $\mathbb{P}(F|N)$ and $\mathbb{P}(M|N)$. For example, for the name $N = Skyler$, the dataset contains the probabilities $\mathbb{P}(F|Skyler) = 37.3\%$ and $\mathbb{P}(M|Skyler) = 62.7\%$.

B.3 TRAINING DATA FOR GENDER (GENTER)

For the training of GRADIEND, we introduce a new dataset called GEnder Name TEmplates with pRonouns (GENTER), which consists of template sentences capable of encoding factual and counterfactual gender information, as illustrated in the motivating example in Section 3.1. Each entry in the dataset includes two template keys: a name [NAME] and a pronoun [PRONOUN]. For instance, the earlier discussed example sentences can be instantiated from the following template:

[NAME] explained the vision as best [PRONOUN] could .

Using the popular BookCorpus (Zhu et al., 2015) dataset, we generated such template sentences that meet the following criteria:

- Each sentence contains at least 50 characters.
- Exactly one name from NAMEXACT is contained, ensuring a correct name match.
- No other names from NAMEXTEND are included, ensuring that only a single name appears in the sentence.
- The correct name's gender-specific third-person pronoun (*he* or *she*) is included at least once.
- All occurrences of the pronoun appear after the name in the sentence.
- The counterfactual pronoun does not appear in the sentence.
- The sentence excludes gender-specific reflexive pronouns (*herself*, *himself*) and possessive pronouns (*her*, *his*, *hers*, *him*).

- Gendered nouns (e.g., *actor*, *actress*, ...) are excluded, based on a gendered-word dataset[1], which is expanded with plural forms using the Python library `inflect`, resulting in 2,421 entries.

This approach generated a total of 83,772 sentences. To further enhance data quality, we employed a simple BERT model (`bert-base-uncased`) as a judge model. This model must predict the correct pronoun for selected names with high certainty, otherwise, sentences may contain noise or ambiguous terms not caught by the initial filtering. Specifically, we used 50 female and 50 male names from NAMEXACT$_{\text{train}}$, and a correct prediction means the correct pronoun token is predicted as the token with the highest probability in the MLM task. Only sentences for which the judge model correctly predicts the pronoun for every test case were retained, resulting in a total of $27,031$ unique sentences. We split the data into training (87.5%), validation (2.5%), and test (10%) subsets. The validation split is rather small, due to the large input size of the GRADIEND models (comparable to the size of the base model), see Section 5.1 for more information.

The GENTER dataset is specifically designed to train our proposed GRADIEND models, focusing on gradient updates that influence gender-changing directions. The applied filtering constraints ensure that the only distinguishing gender-related factor between the factual and counterfactual versions of a sentence is the pronoun (*he* or *she*) associated with the actual gender linked to the name. While our experiments show that using the name-pronoun associations in GENTER effectively uncovers a proper feature encoding and debiasing, future work could investigate whether incorporating additional context, such as gendered nouns or adjectives, provides further useful information.

We selected the BookCorpus (Zhu et al., 2015) as the foundational dataset due to its focus on fictional narratives where characters are often referred to by their first names. In contrast, the English Wikipedia (Wikimedia Foundation, 2023), also commonly used for the training of transformer models (Devlin et al., 2018; Liu et al., 2019), was less suitable for our purposes. For instance, sentences like *[NAME] Jackson was a musician, [PRONOUN] was a great singer* complicate bias detection based on first names (as done for GENTER) due to the context of well-known individuals, where the name and pronoun association can be highly influenced by prior knowledge rather than bias.

B.4 TRAINING DATA FOR RACE AND RELIGION

We filter the same Wikipedia dump used by (Meade et al., 2022) to create the templated GRADIEND training datasets for race and religion, similar to how they augmented counterfactual data for their CDA training. Following their approach, we use their defined bias attribute words to identify factual and counterfactual terms. These words consist of triples representing each feature class class, e.g., *Church/Synagogue/Mosque* for *Christian/Jewish/Muslim* or *Asia/Africa/Europe* for *Asian/Black/White* For each directed pair of classes (e.g., $A = Asian$ and $B = Black$), we retain only sentences containing a bias word from A (factual term) and use the corresponding term for B of the triple as counterfactual term. The casing of the counterfactual term matches that of the original factual term (lowercase by default, or all caps/ capitalized first letter). The resulting datasets $\mathcal{T}$ in training (70%), validation (20%), and test (10%) subsets.

B.5 BIASNEUTRAL

To evaluate our models on data that is entirely bias-unrelated, we created a separate dataset called BIASNEUTRAL, again sourced from BookCorpus (Zhu et al., 2015). Similar to GENTER, we removed sentences containing gender-specific words, as well as any bias attribute words (i.e., the target tokens defined by Meade et al. (2022), such as *African*, *Christian*, ...). We also excluded sentences containing any name from NAMEXTEND and sentences shorter than 50 characters. This filtering process resulted in a total of 20M bias-neutral sentences, with all duplicate sentences removed.

[1] `https://github.com/ecmonsen/gendered_words`

B.6 GENTYPES

Given that the training of the gender GRADIEND specifically involves the relationship between names and their associated genders, we aim to evaluate our models against the gender associations of names with gender-stereotypical contexts to select gender bias-changed models.

> My friend, [NAME], loves taking care of babies.
> My friend, [NAME], is probably into video games.

We expect gender-biased models to assign higher probabilities to female in the first sentence and male names in the second one due to existing stereotypes in real-world data. We generated 500 sentences with GPT-4o (OpenAI, 2024) and refer to this gender-stereotype data as GENTYPES.

The dataset was generated with the following prompt:

> Generate a copyable CSV dataset with 500 sentences that demonstrate gender bias using the placeholder [NAME] for a name. The dataset should include four columns: "text," "gender," and "reason,". The "gender" column should contain 'M' for male or 'F' for female. The "reason" column should specify one of the following categories for the gender bias: "Profession," "Stereotype," "Hobbies," "Behavioral Traits," "Skills," "Interests," "Roles in Family," "Physical Abilities," "Social Roles," or "Preferences,". Provide at least 50 sentences for each reason category. Ensure the sentences are diverse and avoid repetitions. Do not use any words that indicate the gender such as "him", "his", "actress", ...
>
> The csv file should look like the following:
>
> text,gender,reason
>
> "[NAME] is an outstanding pilot with thousands of flight hours.",M,Profession
>
> "[NAME] is a fantastic hair stylist with many loyal clients.",F,Profession
>
> "[NAME] is probably into video games.",M,Interests
>
> "[NAME] is likely to be more empathetic.",F,Behavioral Traits

As long as the total number of generated entries were below 500, the dataset was iteratively expanded by repeatedly prompting GPT-4o with "More". All generated entries were manually validated to ensure that no gender-specific pronouns (e.g., *he*, *she*, *his*, etc.) were present. Entries containing such pronouns were excluded. The final dataset size was capped at 500 entries.

Although the *gender* and *reason* columns were not directly used in this study, their inclusion was intended to enforce balance between male- and female-associated stereotypes and to enhance diversity in stereotype contexts. However, this goal may not have been fully achieved, as RoBERTa demonstrates a female bias in predictions (see Section 5.3), in contrast to our expectations of a generally male biased model.

To encourage the model to predict names on these masked sentences, we used the prefix *"My friend, [MASK], has a ... "* rather than *"[MASK] has a ... "*, which could logically allow for other (unwanted) tokens, such as *he* or *she*.

B.7 WIKIGENDER

To evaluate how well the GRADIEND encoder generalizes to unseen tokens and to data from a different source than seed during training, we derive masked texts from the English Wikipedia (Wikimedia Foundation, 2023). We filter and mask occurrences of the following gendered target word pairs: *she/he*, *woman/man*, *girl/boy*, *mother/father*, and *daughter/son*. For each target, we retain 1,000 texts, forming the dataset WIKIGENDER.

Unlike BookCorpus, the base dataset for GENTER used to train the gender GRADIENDs, Wikipedia articles are much longer (on average ≈ 400 words for WIKIGENDER vs. ≈ 17 words for GENTER), contain structural elements such as headings and newlines, and cover encyclopedic content rather than narrative text. This enables evaluation of both input distribution and target token shifts.

C METRICS

In this section, we define the metrics of Section 5 used to evaluate the GRADIEND encoder and to select bias-changed models formally and more detailed. Additionally, we discuss established techniques to measure bias in language models.

C.1 LANGUAGE MODELING SCORE OF THE DECODER

We use LMS_{Dec} as a measure of the general language modeling capabilities of a model that may have been modified by the GRADIEND decoder. To ensure that the evaluation is independent of any gender bias change, we employ a TPT on BIASNEUTRAL.

For encoder-only models, the TPT corresponds to a MLM task, where 10,000 BIASNEUTRAL samples are used for gender evaluation and 1,000 samples for race and religion, reflecting the larger number of GRADIEND models in the latter case. Approximately 15% of the tokens are masked, following standard practice (Devlin et al., 2018), and LMS_{Dec} is computed as the accuracy on the MLM task.

For decoder-only models, we compute perplexity over 1,000 samples – fewer than in the MLM setting, as the model predicts every token in each sequence, resulting in both higher computational cost and more relevant tokens per sample. Perplexity measures the model's confidence, with lower values indicating better performance, ranging from 1 to infinity. To align its interpretation with accuracy, we convert it to $\text{LMS}_{\text{Dec}} = \frac{1}{1+perplexity}$, yielding scores in $[0, 1]$.

C.2 GENDER PREDICTION PROBABILITIES

This section introduces probabilities for a feature class A, $\mathbb{P}(A)$. We initially restrict this to gender (i.e., female and male probabilities, $\mathbb{P}(F)$ and $\mathbb{P}(M)$) for clarity, which are generalized to other feature classes in the following section.

Let $\mathcal{N}$ denote the set of single-token names in NAMEXTEND, and let $G \in \{F, M\}$ be a gender. Let the GENTYPES data be denoted as T, i.e., stereotyped sentences with a name placeholder [NAME], e.g.:

> My friend, [NAME], loves taking care of babies.

Let $t \in T$ be a text and $|T|$ denote the number of elements in the set T.

Each text t creates a TPT $\hat{t}$ where the goal is to predict a name. For encoder-only models, [NAME] is simply replaced by [MASK], creating a MLM task. For GPT-2 and LLaMA, we transform t into a sentence that naturally prompts a name prediction, using the following template style:

> The person, who loves taking care of babies, has the first name [MASK]

This is done by removing the prefix *"My friend, [NAME],"* and the final punctuation from t. For LLaMA-Instruct, we use the original text with the [NAME] placeholder as user prompt and prepend a system prompt instructing the model to predict a suitable name:

> You are a language model trained to predict first names. In the following text, [NAME] represents a placeholder for a first name. Your task is to predict the most likely name that fits the context. Return only the predicted name — no punctuation, no quotation marks, and no explanations.

The probability distribution over the first generated token is then treated as the model's prediction for $\hat{t}$, similar to the other models.

The probability of predicting a name $N \in \mathcal{N}$ for $\hat{t}$ is denoted as $\mathbb{P}_t(N)$. Names are treated independent of casing and leading white spaces, i.e., the probabilities of all such tokens contribute to this probability.

The probability of predicting gender G for $\hat{t}$ is estimated by summing $\mathbb{P}_t(N)$ for all names N of that gender:

$$\mathbb{P}_t(G) \coloneqq \sum_{N \in \mathcal{N}} \mathbb{P}_t(N) \cdot \mathbb{P}(G|N) \in [0, 1]. \tag{2}$$

As introduced in Section B.2, $\mathbb{P}(G|N)$ represents the likelihood of a name N being associated with gender G. This conditional probability acts as a filter in the sum over all names in $\mathcal{N}$, ensuring that names of the other gender do not contribute to the aggregated probability of G. Moreover, $\mathbb{P}(G|N)$ ensures that names applicable to both genders contribute only partially to the aggregated probability of gender G. For example, for $N = Skyler$, $P_t(Skyler)$ contributes to $\mathbb{P}(F|Skyler) = 37.7\%$ to the female probability $\mathbb{P}_t(F)$ and $\mathbb{P}(M|Skyler) = 62.7\%$ to the male probability $\mathbb{P}_t(M)$.

The combined probabilities for either male or female names is given by

$$\mathbb{P}_t(F \cup M) \coloneqq \mathbb{P}_t(F) + \mathbb{P}_t(M) \in [0, 1].$$

This probability quantifies the proportion of meaningful predictions for $\hat{t}$.

The probability of gender G, denoted as $\mathbb{P}(G)$, averages $\mathbb{P}_t(G)$ over all $t \in T$, i.e.:

$$\mathbb{P}(G) \coloneqq \frac{1}{|T|} \sum_{t \in T} \mathbb{P}_t(G) \in [0, 1].$$

C.3 Generalization of Gender Probabilities to Feature Class Probabilities

We generalize gender probability framework to other feature classes, such as race and religion, by the following adaptions:

- Instead of a gender G, we consider general feature classes $F, F_1, F_2 \in \{Asian, Black, White, Christian, Jewish, Muslim\}$.
- Instead of GENTYPES we use the test split of $\mathcal{T}$ as T.
- Instead of names, we use the set of bias attribute terms $\mathcal{A}_F$ (Meade et al., 2022) for each feature class as target tokens, i.e., the sets $A_{Asian} \cup A_{Black} \cup A_{White}$ and $A_{Christian} \cup A_{Jewish} \cup A_{Muslim}$ are analogous to the name token set $\mathcal{N}$ for gender.
- The conditional probability $\mathbb{P}(F|A)$ for a bias attribute term A is defined as 1 if $A \in \mathcal{A}_F$ and 0 otherwise, reducing Equation 2 to $\mathbb{P}_t(F) \coloneqq \sum_{A \in \mathcal{A}_F} P_t(A)$.
- These adaptions yield similar definitions for $\mathbb{P}_t(F_1 \cup F_2)$ and $\mathbb{P}(G)$.
- For encoder-only models, multi-token target terms are handled by computing the joint probability across all tokens, allowing both single- and multi-token bias attribute terms to contribute meaningfully to the per-example probabilities.
- For decoder-only models, considering only the first token of each target term can be noisy, since it may consist of just one or two characters (especially for the large LLaMA tokenizer) and be poorly aligned with the intended term meaning. Instead, we include all first tokens of the target terms that constitute at least half of the attribute term (in characters), providing a more reliable estimate of the term's probability.
- For LLaMA-Instruct, we use the same prompt as in training, without the special prompt used for gender names (see Section D.3).

C.4 Model Selection Metrics

The Balanced Bias Score (BalancedBS) integrates the previous measures aiming to quantify how debiased a model is over feature classes A and B, by averaging over all texts $t \in T$:

$$\text{BalancedBS} \coloneqq \frac{\text{LMS}_{\text{Dec}}}{|T|} \cdot \sum_{t \in T} \Big[(1 - |\mathbb{P}_t(A) - \mathbb{P}_t(B)|) \cdot \mathbb{P}_t(A \cup B) \Big] \in [0, 1].$$

Here, LMS_{Dec} ensures that high values indicate models with good language modeling capabilities. The first part of the product in the sum $(1 - |\mathbb{P}_t(A) - \mathbb{P}_t(B)|)$ is large if the predictions are unbiased over the two classes A and B, since $\mathbb{P}_t(A)$ must be similar to $\mathbb{P}_t(B)$ to achieve a good score. The second part ($\mathbb{P}_t(A \cup B)$) ensures that both class probabilities are large to avoid a good scoring of models that assign probabilities close to zero to the class target tokens. A high value in BalancedBS indicates a relatively debiased model, that has still good language modeling capabilities due to the influence of LMS_{Dec}.

The Female Bias Score (FemaleBS) measures bias towards the female gender

$$\text{FemaleBS} := \frac{\text{LMS}_{\text{Dec}}}{|\mathcal{T}|} \cdot \sum_{t \in \mathcal{T}} (1 - \mathbb{P}_t(M)) \cdot \mathbb{P}_t(F) \in [0, 1].$$

LMS_{Dec} ensures again good language modeling capabilities, $1 - \mathbb{P}_t(M)$ prefers small male probabilities, and $\mathbb{P}_t(F)$ prefers large female probabilities.

Analogously, the Male Bias Score (MaleBS) measures bias towards the male gender:

$$\text{MaleBS} := \frac{\text{LMS}_{\text{Dec}}}{|\mathcal{T}|} \cdot \sum_{t \in \mathcal{T}} (1 - \mathbb{P}_t(F)) \cdot \mathbb{P}_t(M) \in [0, 1].$$

C.5 Bias Metrics

Various methods exist in literature to quantify bias in language models (see, e.g., Li et al. (2023)). Here, we present a few representative techniques commonly used to measure stereotypical bias.

The Sentence Encoder Association Test (SEAT; May et al. 2019) extends the Word Embedding Association Test (WEAT; Caliskan et al. 2017) by using sentence templates to evaluate social biases in encoder models. It compares association strengths between embeddings of predefined attribute (e.g., gender-specific names) and target sets (e.g., stereotypical professions) using cosine similarity. Bias is expressed as an *effect size*, where larger values indicate stronger bias.

StereoSet (Nadeem et al., 2021) is a benchmark dataset with context-rich sentences for intrasentence and intersentence tasks. This study focuses on the intrasentence task, where a sentence (e.g., *Girls tend to be more [MASK] than boys*) requires the model to predict the masked word from three options: stereotypical (e.g., *soft*), anti-sterotypical (e.g., *determined*), and meaningless (e.g., *fish*). Two metrics are considered: **1)** $\text{LMS}_{\text{StereoSet}}$, which measures the proportion of meaningful (stereotypical or anti-stereotypical) options chosen over meaningless ones, reflecting the model's language understanding. **2)** SS, which quantifies bias as the proportion of stereotypical options selected over anti-stereotypical ones. A balanced model achieves 50%.

CrowS (Crowdsourced Stereotype Pairs; Nangia et al. 2020) is a crowdsourced dataset consisting of pairs of sentences: one expressing a stereotype (e.g., *Woman don't know how to drive*), and the other its anti-stereotypical counterpart (e.g., *Man know how to drive*). A bias score is computed considering the model's preference for one sentence over the other, similar to SS. However, CrowS has been criticized for unreliable bias measurement, including spurious correlations and flawed assumptions about social categories (Blodgett et al., 2021). Therefore, we did not use this metric in this study, but report it here for completeness.

Li et al. (2021) analyze the attention associations between gendered pronouns (e.g., *she*) and occupations (e.g., *nurse*) in transformer models, using gender-swapped sentences (e.g., replace *he* by *she*). The attention scores between the gender-swapped pronouns and the occupation are then compared to identify gender bias on attention head level. However, the approach does not compute a model-specific, aggregated bias score usable for comparison.

D Training and Implementation Details

Table 5 summarizes the Hugging Face model checkpoints used in our experiments, while Table 6 lists the hyperparameters used for training the Gradiend models.

D.1 Environment

The implementation is based on Python 3.9.19, and we made the training framework publicly available: anonymous. The LLaMA-based Gradiend models were trained using three NVIDIA

Table 5: Hugging Face model checkpoints used in this study.

Model	**Checkpoint**	**Reference**
$\text{BERT}_{\text{base}}$	`bert-base-cased`	Devlin et al. (2018)
$\text{BERT}_{\text{large}}$	`bert-large-cased`	Devlin et al. (2018)
DistilBERT	`distilbert-base-cased`	Sanh et al. (2019)
RoBERTa	`roberta-large`	Liu et al. (2019)
GPT-2	`gpt2`	Radford et al. (2019b)
LLaMA	`meta-llama/Llama-3.2-3B`	Grattafiori et al. (2024)
LLaMA-Instruct	`meta-llama/Llama-3.2-3B-Instruct`	Grattafiori et al. (2024)

Table 6: Training hyperparameters.

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate	1×10^{-4} (LLaMA, LLaMA-Instruct); 1×10^{-5} (others)
Weight Decay	1×10^{-2}
Batch Size Gradient Computation	32
Batch Size GRADIEND	1
Training Criterion	MSE
Training Steps	23,653 (Gender); 2,500 (Race, Religion)
Evaluation Steps	250
Evaluation Criterion	$\text{Cor}_{\mathcal{T}}$ on validation split

A100 GPUs, while all others used a single A100. Each A100 provides $80\,\text{GB}$ of GPU memory, and the system had $504\,\text{GB}$ of RAM. The same setup is also used for evaluation.

D.2 Token Prediction Task for Encoder-Only Models

The training task for GRADIENDis motivated as a MLM Devlin et al. (2018) task (see Section 3.1), where the masked token is sensitive to an involved feature class. For multi-token targets, we insert one `[MASK]` token per target token in the template text. The MLM loss then naturally aggregates over all target tokens, so the resulting gradients reflect contributions from each token.

D.3 Token Prediction Task for Decoder-Only Models

For causal models, MLM instances are converted into a CLM Radford et al. (2019a) task by providing only the prefix up to the (first) masked token and predicting the next token at the end of the sequence.

For LLaMA-Instruct, we use the following system prompt to allign its behavior with non-instruction-tuned models:

> You are a language model that completes sentences. Predict the next word that naturally follows the given text. Return only that word — no punctuation, no quotes, and no explanations.

This prompt is used for all applications of LLaMA-Instruct in this study unless stated otherwise.

Although this modification is straightforward, it is effective only when the target terms can be tokenized as single tokens – or when the primary semantic content is largely captured by the first token (e.g., similar to Appendix C.3). This limitation is particularly noticeable for LLaMA-based models with race and religion terms, as illustrated in Figure 5. Future work should investigate methods to handle multi-token targets in decoder-only GRADIEND models.

D.4 Custom Initialization

Our training setup involves a custom random initialization for the GRADIEND models. The default initialization in PyTorch applies a uniform distribution from $\left(\frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$, where n is the dimension of the input layer. However, for the decoder, the input dimension is $n = 1$, resulting in a uniform distribution over the interval $(-1, 1)$. This leads to relatively high absolute initial values compared to the target values, as the decoder inputs are typically close to ± 1. To address this, we use the same n for the initialization as for the encoder, which corresponds to the number of used weights in the designated model. Our experiments show that this custom initialization improves training results.

D.5 TRAINING PROCEDURE

Each training step involves two forward and backward passes through the base model to compute the input and output tensors for the GRADIEND model. For race and religion, the training data for classes A and B is derived by combining the datasets for each source class and augmenting the targets with all valid terms from the other class within the same bias attribute group. For gender, each entry of GENTER is augmented batch-size many times with a name of NAMEXACT to generate the actual training dataset. Gradients are calculated with respect to the target token, e.g. *he*/*she* or *He*/*She*, depending on the position of the target token. We only used single token targets for training, i.e., the datasets were filtered to exclude multi-token targets or sources.

We use the validation split of $\mathcal{T}$ for evaluation during training, following the same procedure as described to compute $\text{Cor}_{\mathcal{T}}$ (Section 5.2). However, as pre-computing these validation gradients require a substantial amount of storage, we use for the gender GRADIENDs all of the GENTER validation split for the smaller models ($\text{BERT}_{\text{base}}$, DistilBERT, and GPT-2), half of the data for the medium-sized models ($\text{BERT}_{\text{large}}$ and RoBERTa), and only 5% for the LLaMA-based models due to their large model sizes. This ensures that the gradients required for evaluation fit into the memory during training. For instance, the evaluation data for $\text{BERT}_{\text{base}}$ requires approximately 270 GB. For race and religion, a maximum of 1,000 samples is used, with similar relative reductions based on model size. The training time for a single gender GRADIEND model ranges from 3.5 hours for DistilBERT to 24 hours for LLaMA-Instruct.

To monitor progress, the model is evaluated every 250 training steps using $\text{Cor}_{\mathcal{T}}$, and select the best model after finishing all training steps (Section 5.1). Similar to the procedure to evaluate the GRADIEND encoder (Section 5.2), This evaluation metric focuses on the encoder's ability to differentiate between genders, which measures how well the encoded values distinguish between the feature classes. Notice that this metric evaluates only the encoder, as the decoder's role in adjusting bias is harder to evaluate.

When training the gender GRADIEND models, they sometimes fail to converge in distinguishing female and male input as ± 1, depending on the learning rate and random seed. This issue was observed particularly with RoBERTa, although it occasionally occurred with other models as well, depending on the learning rate. In such cases, the first training steps determine whether both genders are separated correctly or both are encoded as the same value (either $+1$ or -1). Future research is needed to explore this phenomenon. To mitigate non-convergent runs for gender, we train three GRADIEND models per base model with different seeds and select the one with the highest $\text{Cor}_{\mathcal{T}}$ on the validation split. For race and religion, a single GRADIEND model is trained per configuration.

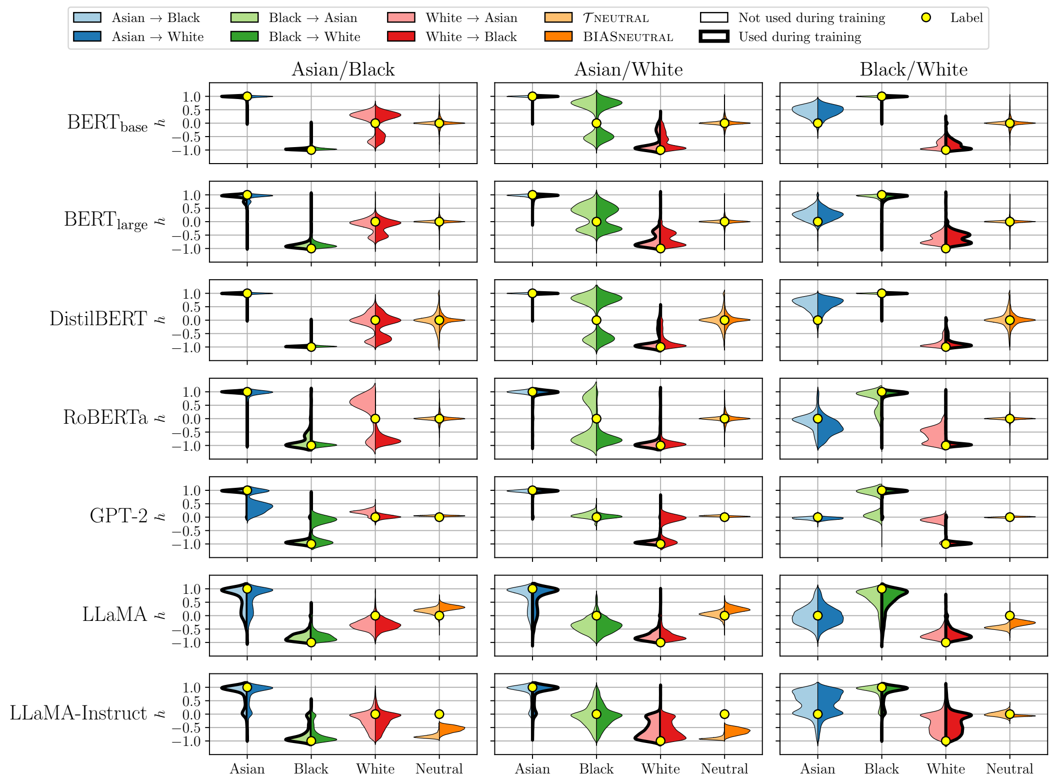
E ENCODER AS CLASSIFIER

E.1 DETAILED RESULTS

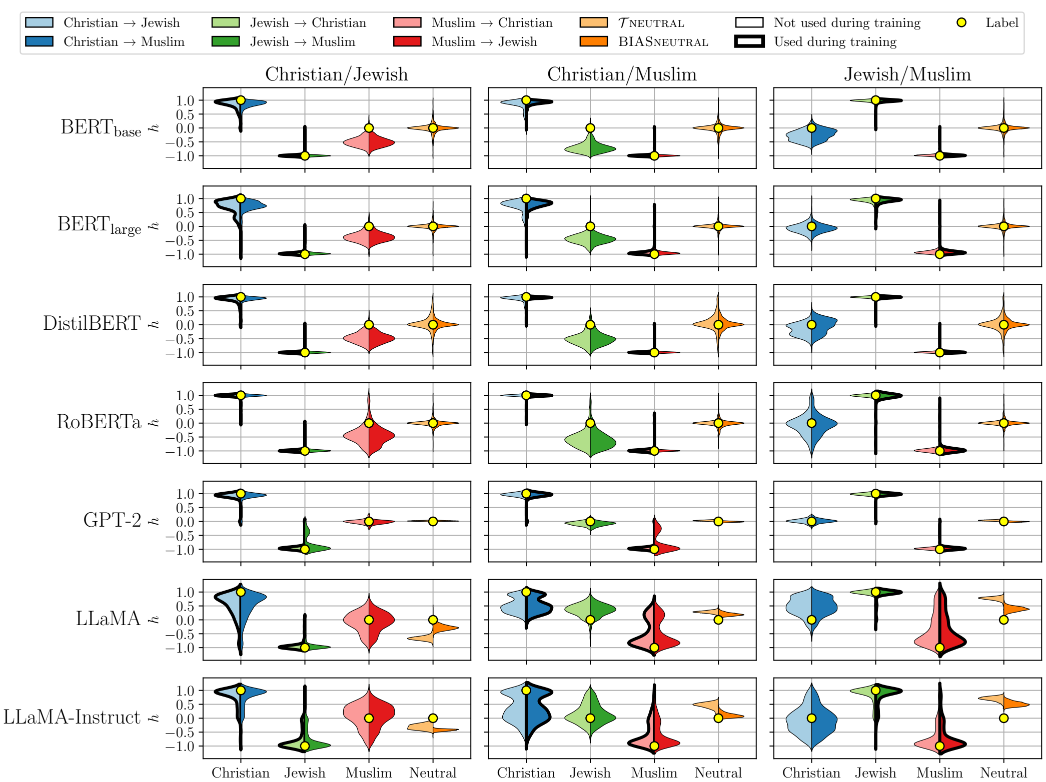
Similar to Figure 3, we present additional results in Figure 5, showing the distribution of encoded values of race and religion GRADIEND models evaluated against a broad set of datasets. The data of these plots has been used to compute $\text{Cor}_{\mathcal{T}}$ and Cor_{Enc} in Table 1.

E.2 STABILITY OF ENCODED VALUES

We analyze the stability of the feature neuron by examining the encodings from three independently trained gender GRADIEND models for each base model. Figure 6 shows the distribution of these encoded values, along with sample-wise differences to highlight run-to-run variation, and Table 7 summarizes key statistics.



(a) Race.



(b) Religion.

Figure 5: Distribution of encoded values for all race and religion GRADIEND models across different datasets. The yellow dots indicate the expected label used for Cor_{Enc}.

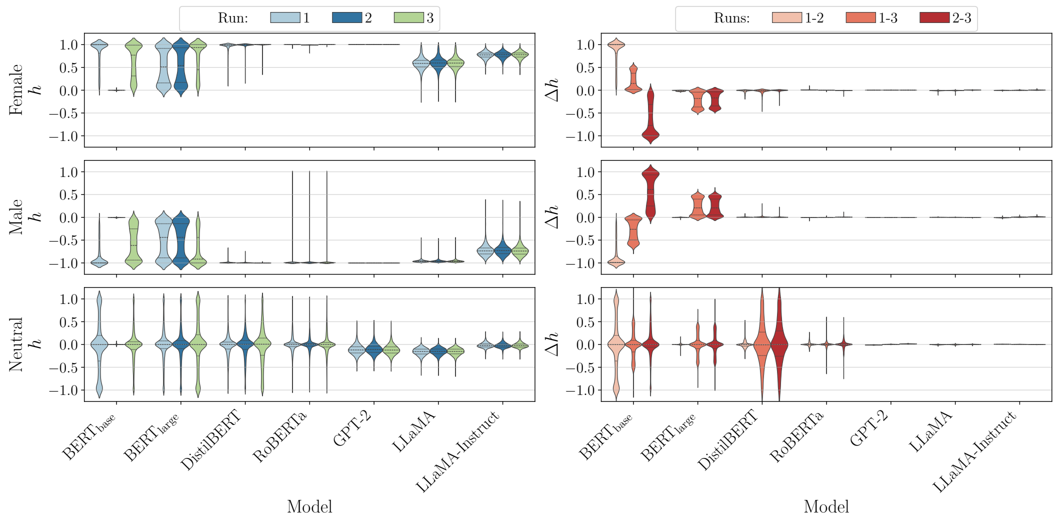


Figure 6: Distribution of encoded values h (left) and their sample-wise difference Δh (right) across three GRADIEND training runs for gender.

Table 7: Stability analysis of encoded values across three GRADIEND training runs for gender.

	$\text{Cor}_{\text{Enc}} \uparrow$				Mean Absolute Difference of Encoded Values $\downarrow$			
Model	**Run 1**	**Run 2**	**Run 3**	**Mean**	**Runs 1-2**	**Runs 1-3**	**Runs 2-3**	**Mean**
$\text{BERT}_{\text{base}}$	0.713	0.076	0.706	0.498	0.558	0.212	0.350	0.373
$\text{BERT}_{\text{large}}$	0.621	0.622	0.660	0.635	0.008	0.173	0.168	0.117
DistilBERT	0.939	0.862	0.860	0.887	0.035	0.245	0.256	0.179
RoBERTa	0.964	0.977	0.953	0.965	0.019	0.018	0.036	0.024
GPT-2	**0.984**	**0.985**	**0.984**	**0.984**	0.007	**0.002**	0.009	0.006
LLaMA	0.981	0.983	0.983	0.982	**0.005**	0.004	**0.002**	0.004
LLaMA-Instruct	0.977	0.976	0.977	0.976	**0.005**	0.003	0.003	**0.004**

With the exception of the BERT-based models, the feature neuron is generally stable across female, male, and neutral inputs. DistilBERT and RoBERTa show some variability for neutral inputs across runs, while GPT-2, LLaMA, and LLaMA-Instruct exhibit a mean absolute encoding difference below 1%.

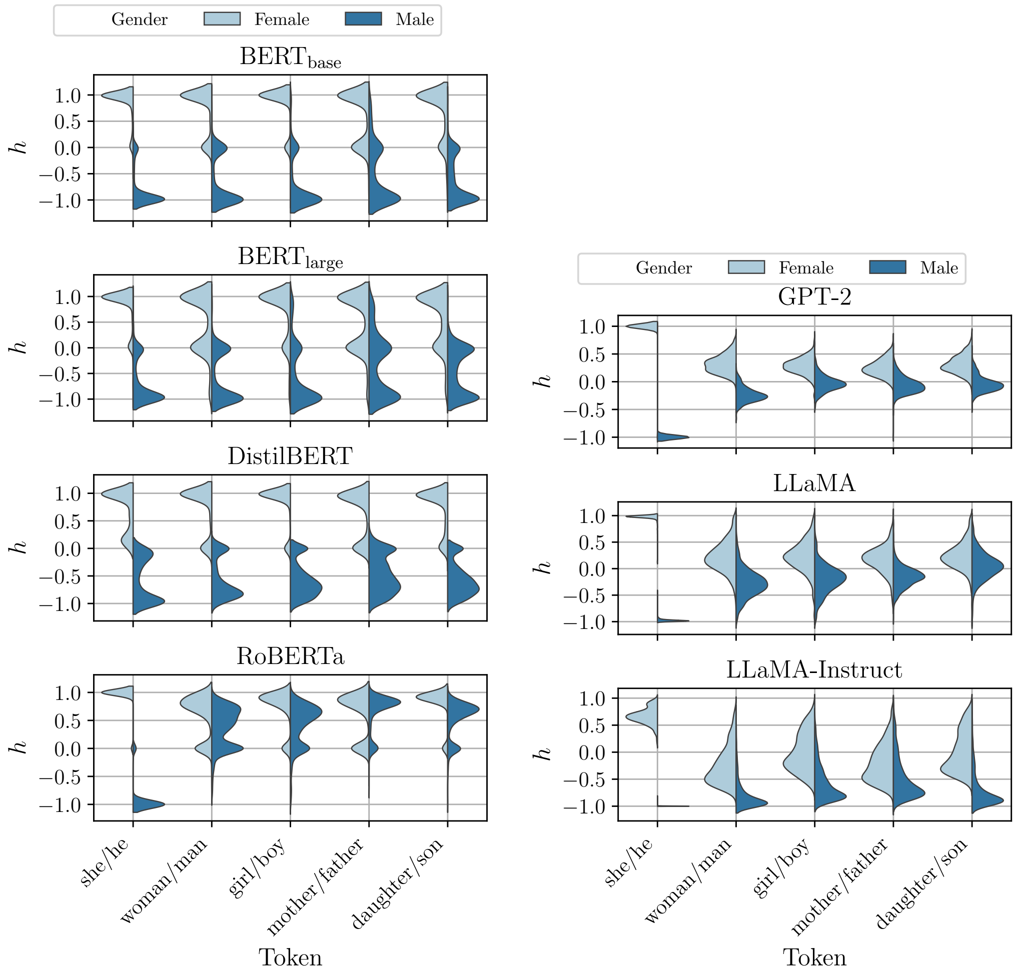
For $\text{BERT}_{\text{large}}$, the third run achieves notably higher performance than the first two, which are fairly similar to each other. In contrast, $\text{BERT}_{\text{base}}$ shows a non-convergent second run, resulting in large differences compared to the other runs.

E.3 GENERALIZATION OF ENCODED VALUES

We further analyze how the encoder generalizes to unseen inputs, considering two aspects: **(1)** the input sentences originate from a dataset different from the one used during training, and **(2)** the evaluation involves gender-related target tokens beyond the training pair *he/she*. Therefore, we use WIKIGENDER as a dataset (see Appendix B.7).

Figure 7 shows the distribution of encoded values for our seven gender GRADIENDs. The she/he encoding learned during the training transfers well to WIKIGENDER, indicating that the feature is not tied to the specific structure, linguistic style, and gender-filtered property of GENTER.

For $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, and DistilBERT, the learned feature also generalizes to other gendered token pairs such as *woman/man*, though the separation is a bit weaker than for *she/he*, as more samples are falsely encoded as neutral (i.e., around 0.0). A plausible explanation is that masking *he/she* yields a highly constrained prediction space, as only a few tokens fit the syntactic and semantic context, whereas masking, for instance, *woman/man* allows usually a broader set of contextually plausible alternatives (e.g., *girl/boy*), including gender neutral terms like *person*. Interestingly, RoBERTa



(a) Encoder-only models.

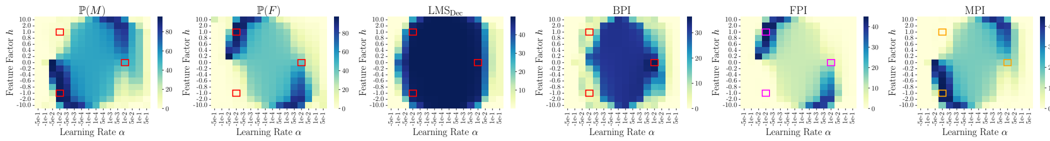
(b) Decoder-only models.

Figure 7: Distribution of encoded values of gender GRADIENDs for WIKIGENDER.

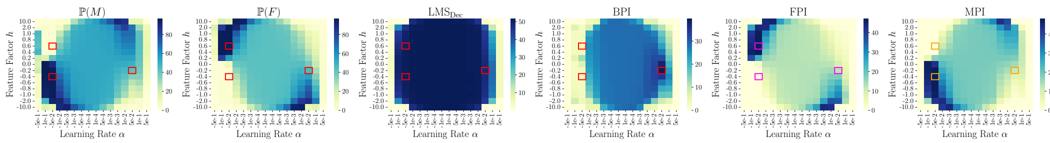
behaves differently: it appears to encode a narrow *she/he*-specific feature rather than a broader gender feature.

For decoder-only models, the generalization is weaker for non-she/he pairs but still visible, as the female-associated tokens tend to encode to larger values than their male counterparts. This less extreme encoding is expected because these models can only use the left context of the target term. Considering the non-*she/he* token pairs for GPT-2 and LLaMA, they show a mostly symmetric distribution around zero with smaller magnitude than for *she/he*, indicating weaker separation. In contrast, LLaMA-Instruct still shows a female-male distinction, but the distributions are shifted toward the male side (i.e., toward -1).

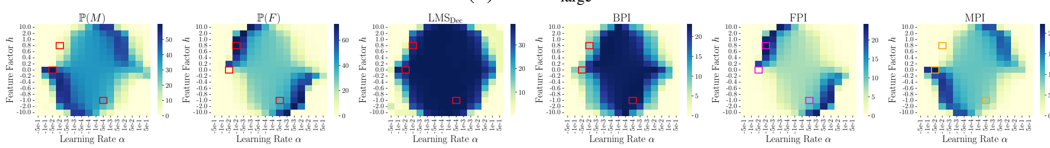
Overall, the results indicate that the features learned by GRADIEND generalize, but that future work should explore training GRADIENDs using multiple facets, i.e., not only a single type of counterfactual (e.g., *she/he*), but also other in parallel, like *woman/man* to possibly find a more general feature representation.



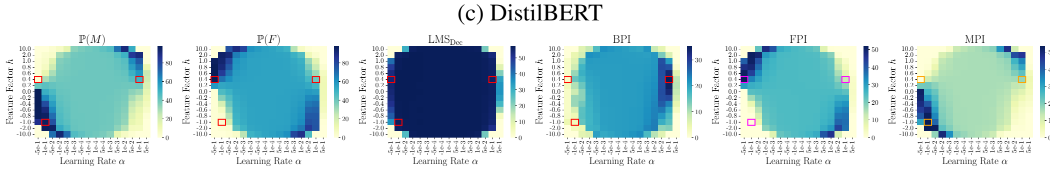
(a) $\text{BERT}_{\text{base}}$



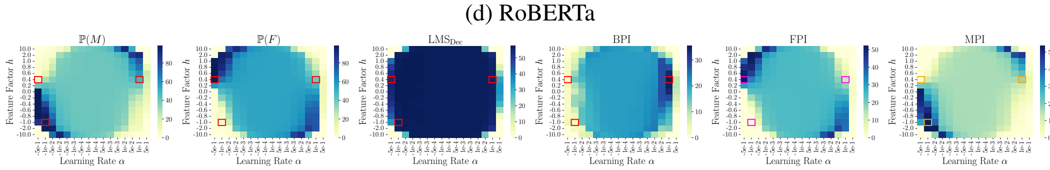
(b) $\text{BERT}_{\text{large}}$



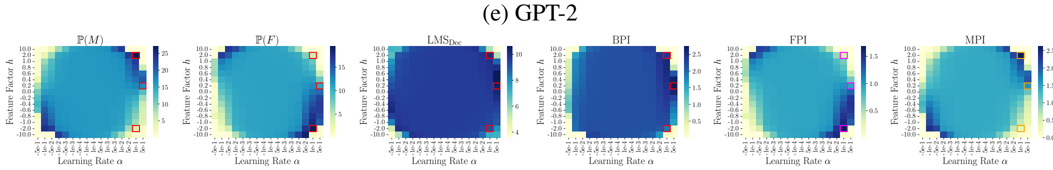
(c) DistilBERT



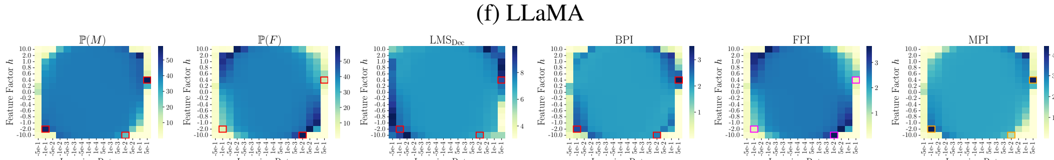
(d) RoBERTa



(e) GPT-2



(f) LLaMA



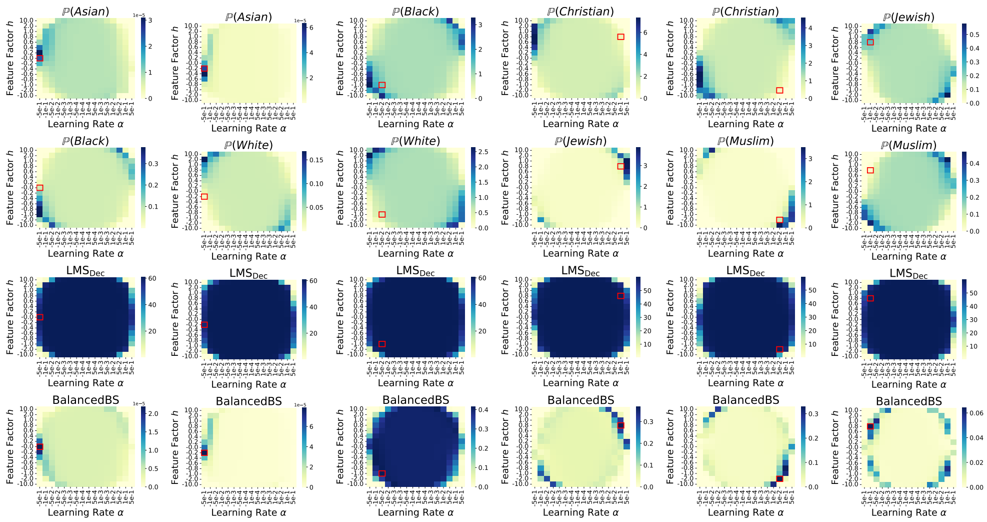
(g) LLaMA-Instruct

Figure 8: Metrics for changed models based on the gender GRADIENDS with varying feature factor and learning rate. The cells with the best BalancedBS □, FemaleBS □, and MaleBS □ are highlighted across all subplots. All values are reported as percentages.

F Decoder as Bias-Changer

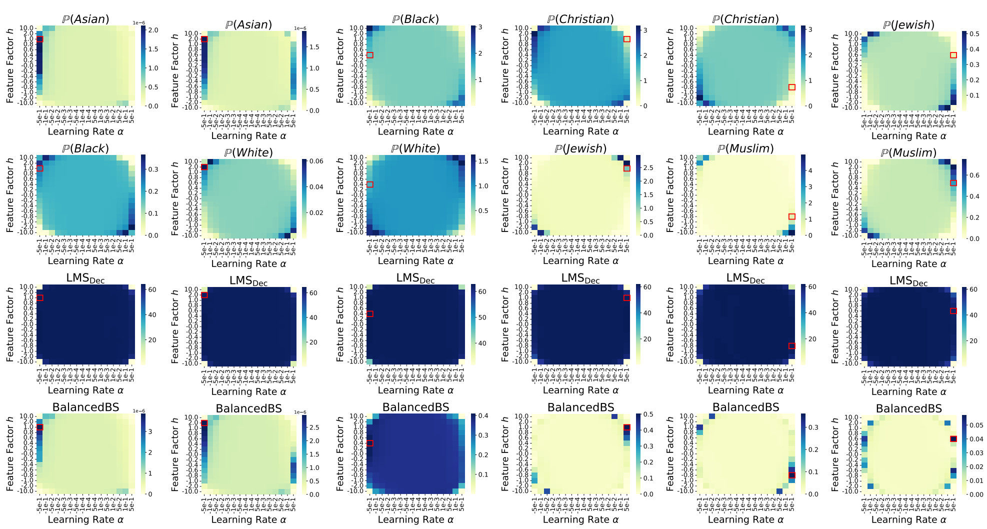
Similar to Figure 4, we present the results for all gender models in Figure 8. We further report the selected race and religion models in Figures 9-15.

Overall, a similar point-symmetric pattern can be recognized across all figures. However, the model selection is different compared to $\text{BERT}_{\text{base}}$, where FemaleBS and MaleBS exhibit negated feature factors along with negative learning rates, while BalancedBS features a zero feature factor and a positive learning rate. Similar configurations exist across most models that outperform the base model



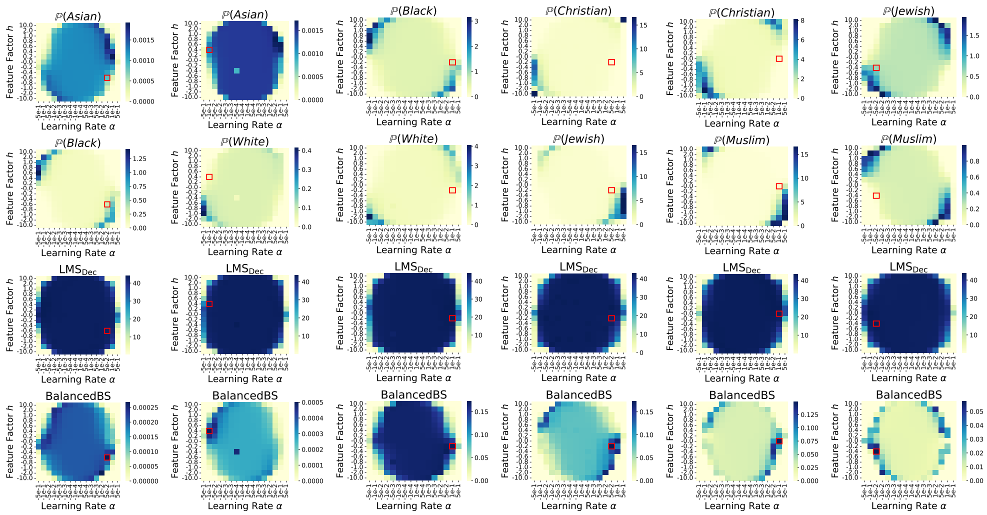
(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 9: Metrics for changed models based on the $\text{BERT}_{\text{base}}$ race and religion GRADIENDs with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



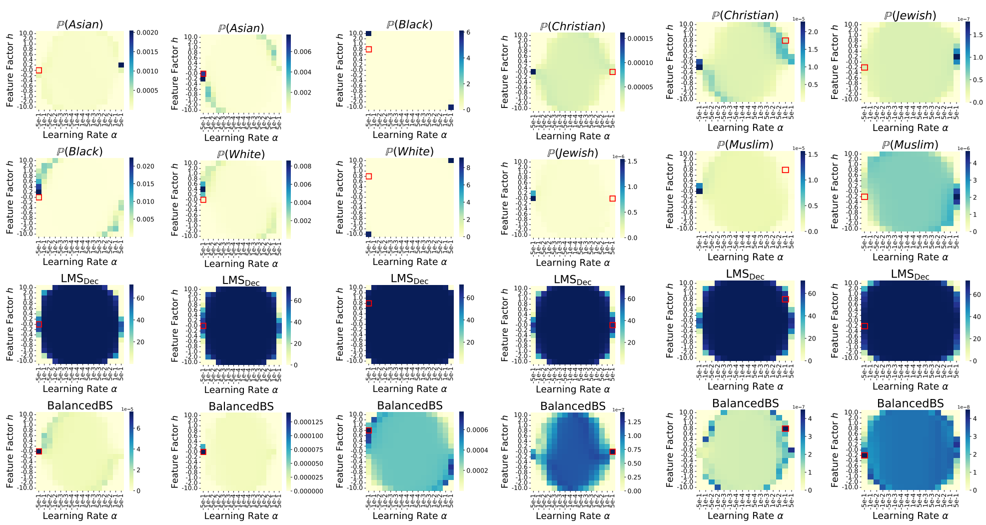
(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 10: Metrics for changed models based on the $\text{BERT}_{\text{large}}$ race and religion GRADIENDs with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



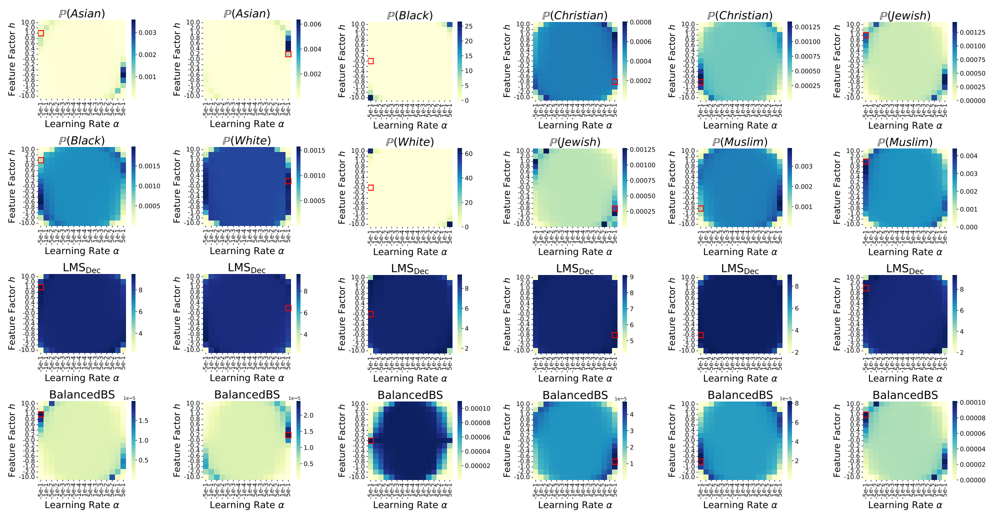
(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 11: Metrics for changed models based on the DistilBERT race and religion GRADIENDs with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



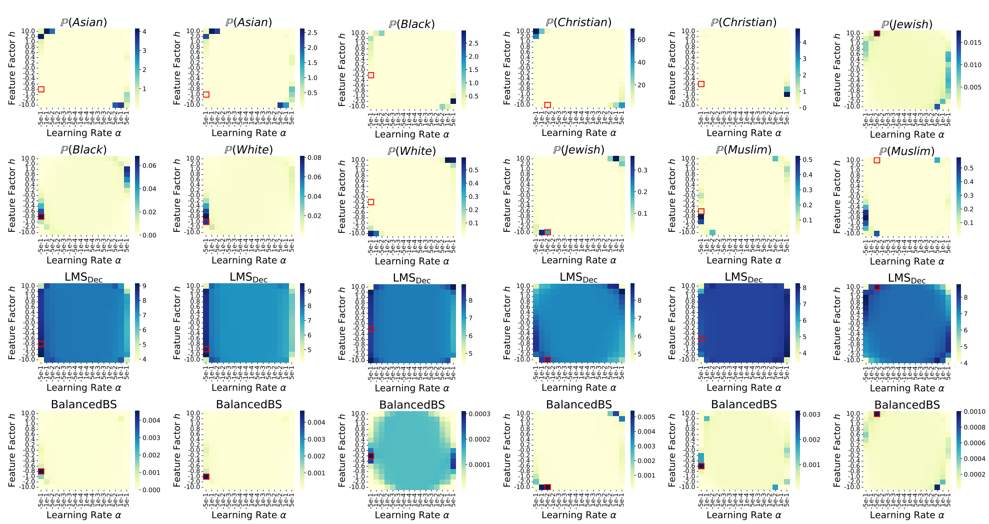
(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 12: Metrics for changed models based on the RoBERTa race and religion GRADIENDs with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



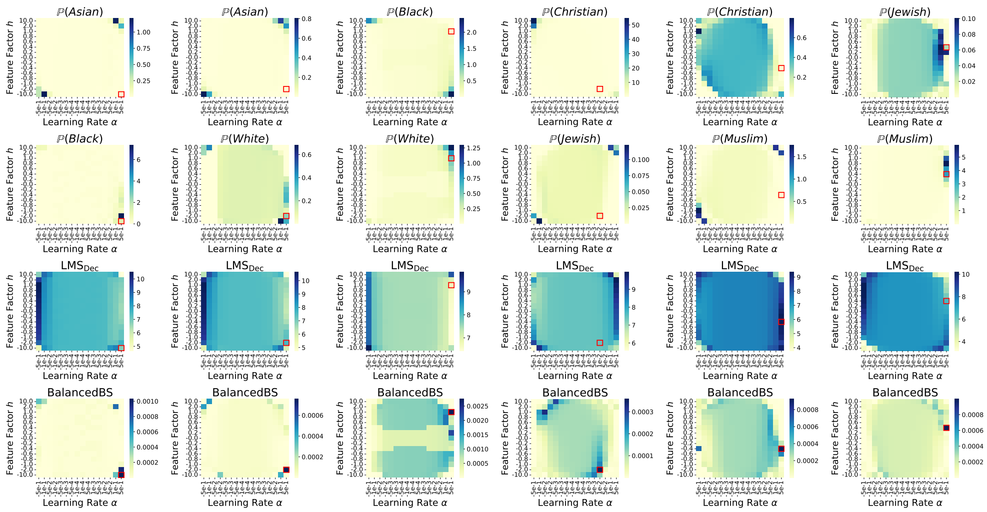
(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 13: Metrics for changed models based on the GPT-2 race and religion GRADIENDS with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 14: Metrics for changed models based on the LLaMA race and religion GRADIENDS with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.



(a) Black/Asian (b) Asian/White (c) Black/White (d) Chr./Jew. (e) Chr./Muslim (f) Jew./Muslim

Figure 15: Metrics for changed models based on the LLaMA-Instruct race and religion GRADIENDS with varying feature factor and learning rate. The cells with the best BalancedBS □ are highlighted across all subplots. All values are reported as percentages.

Table 8: Selected gender-debiased, female-biased and male-biased models based on BalancedBS, FemaleBS and MaleBS.

Model	**FF h**	**LR α**	$\mathbb{P}(F)$	$\mathbb{P}(M)$	**LMS$_{\text{Dec}}$**	**BalancedBS**	**FemaleBS**	**MaleBS**
BERT$_{\text{base}}$	0.0	0.0	0.413	0.538	0.496	0.322	0.110	0.172
+ GRADIEND$_{\text{Female/Male}}$	0.0	1e-02	0.424	0.494	0.491	0.363	0.111	0.145
+ GRADIEND$_{\text{Female}}$	1.0	-1e-02	0.948	0.034	0.490	0.041	0.449	0.001
+ GRADIEND$_{\text{Male}}$	-1.0	-1e-02	0.031	0.940	0.490	0.043	0.001	0.446
BERT$_{\text{large}}$	0.0	0.0	0.432	0.534	0.511	0.280	0.132	0.184
+ GRADIEND$_{\text{Female/Male}}$	-0.2	5e-02	0.461	0.434	0.486	0.397	0.128	0.115
+ GRADIEND$_{\text{Female}}$	0.6	-5e-02	0.960	0.031	0.508	0.036	0.473	0.001
+ GRADIEND$_{\text{Male}}$	-0.4	-5e-02	0.024	0.963	0.511	0.031	0.000	0.480
DistilBERT	0.0	0.0	0.307	0.361	0.392	0.230	0.075	0.096
+ GRADIEND$_{\text{Female/Male}}$	-1.0	5e-04	0.358	0.320	0.386	0.231	0.092	0.078
+ GRADIEND$_{\text{Female}}$	0.8	-1e-02	0.724	0.045	0.350	0.080	0.241	0.004
+ GRADIEND$_{\text{Male}}$	0.0	-5e-02	0.036	0.599	0.384	0.092	0.005	0.221
RoBERTa	0.0	0.0	0.555	0.404	0.573	0.208	0.249	0.162
+ GRADIEND$_{\text{Female/Male}}$	0.4	1e-01	0.402	0.499	0.560	0.354	0.127	0.181
+ GRADIEND$_{\text{Female}}$	0.4	-5e-01	0.980	0.016	0.539	0.019	0.520	0.000
+ GRADIEND$_{\text{Male}}$	-1.0	-1e-01	0.023	0.966	0.568	0.032	0.001	0.535
GPT-2	0.0	0.0	0.028	0.089	0.107	0.012	0.003	0.009
+ GRADIEND$_{\text{Female/Male}}$	-0.6	1e-01	0.172	0.130	0.106	0.031	0.016	0.011
+ GRADIEND$_{\text{Female}}$	10.0	-1e-02	0.270	0.061	0.099	0.025	0.025	0.004
+ GRADIEND$_{\text{Male}}$	-0.4	1e-01	0.123	0.130	0.108	0.027	0.012	0.012
LLaMA	0.0	0.0	0.111	0.162	0.095	0.022	0.009	0.014
+ GRADIEND$_{\text{Female/Male}}$	0.2	5e-01	0.109	0.174	0.104	0.027	0.009	0.016
+ GRADIEND$_{\text{Female}}$	-2.0	1e-01	0.194	0.050	0.091	0.019	0.017	0.004
+ GRADIEND$_{\text{Male}}$	2.0	1e-01	0.042	0.271	0.101	0.024	0.003	0.026
LLaMA-Instruct	0.0	0.0	0.397	0.399	0.073	0.021	0.025	0.025
+ GRADIEND$_{\text{Female/Male}}$	0.4	5e-01	0.147	0.587	0.086	0.035	0.005	0.043
+ GRADIEND$_{\text{Female}}$	-10.0	1e-02	0.594	0.257	0.070	0.024	0.036	0.013
+ GRADIEND$_{\text{Male}}$	-2.0	-1e-01	0.116	0.569	0.084	0.028	0.006	0.044

Table 9: Selected debiased models based on BalancedBS for race and religion. Classes A and B refer to the classes of the GRADIEND$_{A/B}$.

Model	**FF** h	**LR** α	Base $\mathbb{P}(A)$	$\mathbb{P}(A)$	Base $\mathbb{P}(B)$	$\mathbb{P}(B)$	**LMS$_{Dec}$**	**BalancedBS**
GRADIEND$_{Asian/Black}$								
BERT$_{base}$	-0.0	-0.5	8.6e-8	8.6e-8	8.1e-4	8.1e-4	0.587	2.2e-7
BERT$_{large}$	1.0	-0.5	3.8e-9	3.8e-9	1.8e-3	1.8e-3	0.635	3.8e-8
DistilBERT	-0.6	0.1	1.2e-5	1.2e-5	1.1e-3	1.1e-3	0.429	2.7e-6
RoBERTa	-0.0	-0.5	9.6e-7	9.6e-7	1.5e-6	1.5e-6	0.591	7.5e-7
GPT-2	1.0	-0.5	5.7e-7	5.7e-7	1.2e-5	1.2e-5	0.095	2.0e-7
LLaMA	-0.8	-0.5	3.9e-5	3.9e-5	2.4e-5	2.4e-5	0.089	4.5e-5
LLaMA-Instruct	-10.0	0.5	7.6e-5	7.6e-5	3.6e-4	3.6e-4	0.068	1.0e-5
GRADIEND$_{Asian/White}$								
BERT$_{base}$	-0.2	-0.5	4.7e-8	4.7e-8	3.8e-4	3.8e-4	0.570	7.8e-7
BERT$_{large}$	2.0	-0.5	3.3e-9	3.3e-9	2.2e-4	2.2e-4	0.627	3.0e-8
DistilBERT	0.4	-0.1	8.4e-6	8.4e-6	3.7e-4	3.7e-4	0.421	5.0e-6
RoBERTa	-0.0	-0.5	1.7e-6	1.7e-6	2.5e-6	2.5e-6	0.649	1.4e-6
GPT-2	0.2	0.5	6.1e-7	6.1e-7	1.3e-5	1.3e-5	0.089	2.4e-7
LLaMA	-1.0	-0.5	3.0e-5	3.0e-5	2.2e-5	2.2e-5	0.092	4.7e-5
LLaMA-Instruct	-2.0	0.5	8.0e-5	8.0e-5	1.2e-3	1.2e-3	0.077	7.6e-6
GRADIEND$_{Black/White}$								
BERT$_{base}$	-1.0	-0.1	9.8e-3	9.8e-3	9.8e-3	9.8e-3	0.604	4.2e-3
BERT$_{large}$	0.4	-0.5	1.2e-2	1.2e-2	1.2e-2	1.2e-2	0.627	4.1e-3
DistilBERT	-0.2	0.1	5.4e-3	5.4e-3	5.4e-3	5.4e-3	0.441	1.7e-3
RoBERTa	0.8	-0.5	5.8e-6	5.8e-6	5.8e-6	5.8e-6	0.710	7.7e-6
GPT-2	-0.0	-0.5	1.4e-5	1.4e-5	1.4e-5	1.4e-5	0.090	1.1e-6
LLaMA	-0.2	-0.5	3.0e-5	3.0e-5	3.0e-5	3.0e-5	0.082	3.1e-6
LLaMA-Instruct	1.0	0.5	1.1e-3	1.1e-3	1.1e-3	1.1e-3	0.067	2.7e-5
GRADIEND$_{Christian/Jewish}$								
BERT$_{base}$	0.8	0.1	1.5e-2	1.5e-2	2.2e-3	2.2e-3	0.592	3.7e-3
BERT$_{large}$	1.0	0.5	1.8e-2	1.8e-2	2.6e-3	2.6e-3	0.607	5.0e-3
DistilBERT	-0.2	0.1	8.4e-3	8.4e-3	6.2e-3	6.2e-3	0.416	1.7e-3
RoBERTa	-0.0	0.5	3.0e-7	3.0e-7	5.7e-10	5.7e-10	0.665	1.4e-9
GPT-2	-0.8	0.5	5.4e-6	5.4e-6	3.4e-6	3.4e-6	0.091	4.8e-7
LLaMA	-10.0	-0.1	7.9e-5	7.9e-5	1.4e-5	1.4e-5	0.085	5.4e-5
LLaMA-Instruct	-2.0	0.0	4.3e-3	4.3e-3	1.5e-4	1.5e-4	0.075	3.6e-6
GRADIEND$_{Christian/Muslim}$								
BERT$_{base}$	-2.0	0.1	1.1e-2	1.1e-2	1.5e-3	1.5e-3	0.585	3.3e-3
BERT$_{large}$	-0.8	0.5	1.2e-2	1.2e-2	2.0e-3	2.0e-3	0.611	3.6e-3
DistilBERT	-0.0	0.1	7.6e-3	7.6e-3	3.0e-3	3.0e-3	0.430	1.5e-3
RoBERTa	0.8	0.1	4.0e-8	4.0e-8	1.9e-8	1.9e-8	0.664	4.5e-9
GPT-2	-0.8	-0.5	5.4e-6	5.4e-6	2.6e-5	2.6e-5	0.090	8.2e-7
LLaMA	-0.6	-0.5	7.4e-5	7.4e-5	3.5e-5	3.5e-5	0.072	3.1e-5
LLaMA-Instruct	-0.4	0.5	3.9e-3	3.9e-3	1.5e-3	1.5e-3	0.094	9.3e-6
GRADIEND$_{Jewish/Muslim}$								
BERT$_{base}$	0.8	-0.1	1.7e-3	1.7e-3	1.7e-3	1.7e-3	0.589	6.4e-4
BERT$_{large}$	0.6	0.5	1.4e-3	1.4e-3	1.4e-3	1.4e-3	0.592	5.7e-4
DistilBERT	-0.4	-0.1	5.2e-3	5.2e-3	5.2e-3	5.2e-3	0.433	5.7e-4
RoBERTa	-0.2	-0.5	3.4e-10	3.4e-10	3.4e-10	3.4e-10	0.706	4.5e-10
GPT-2	1.0	-0.5	3.4e-6	3.4e-6	3.4e-6	3.4e-6	0.093	1.0e-6
LLaMA	10.0	-0.1	1.3e-5	1.3e-5	1.3e-5	1.3e-5	0.086	1.0e-5
LLaMA-Instruct	0.4	0.5	3.5e-4	3.5e-4	3.5e-4	3.5e-4	0.066	1.0e-5

Table 10: Published gender debiased models on Hugging Face.

Model	**Identifier**
$\text{BERT}_{\text{base}}$ + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
$\text{BERT}_{\text{large}}$ + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
DistilBERT + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
RoBERTa + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
GPT-2 + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
LLaMA + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous
LLaMA-Instruct + $\text{GRADIEND}_{\text{Female/Male}}$	anonymous

with respect to BalancedBS, FemaleBS, and MaleBS. The final selected models, however, perform even better with respect to our metrics, though they do not adhere to the expected pattern. Future research should explore the stability of these parameter choices without relying on a larger search grid.

The statistics for all selected gender models are reported in Table 8. Interestingly, the difference in BalancedBS between $\text{GRADIEND}_{\text{Female/Male}}$ and its base model is relatively small, whereas the corresponding differences for FemaleBS and MaleBS are much larger, respectively. This observation suggests that biasing a model (in either direction) is easier than debiasing it. Notably, FemaleBS approaches nearly zero for $\text{GRADIEND}_{\text{Male}}$ models, and MaleBS similarly is close to zero for $\text{GRADIEND}_{\text{Female}}$ models. Surprisingly, for the RoBERTa base model, $\mathbb{P}(F) > \mathbb{P}(M)$ holds true, unlike all other base models. This indicates a female bias in the given task, contradicting our expectation that language models typically exhibit male bias (although this bias direction is not captured by SS and SEAT).

The statistics for the selected race and religion models are reported in Table 9.

G Comparison to Other Debiasing Techniques

This section provides supplementary details for Section 5.4, which compares our method to existing debiasing techniques. To facilitate future comparisons with our approach, we release our gender-debiased models on Hugging Face (Table 10), where they achieve SoTA debiasing performance.

G.1 Implementation Details

For the evaluation of our gender-changed models on GLUE, SuperGLUE, SEAT, SS, and $\text{LMS}_{\text{StereoSet}}$, we primarily rely on the bias-bench implementation by Meade et al. (2022), which we also use to compute and evaluate the baseline debiasing techniques: CDA, DROPOUT, INLP, SENTDEBIAS, and SELFDEBIAS. For implementation specifics and metric definitions, we refer the reader to the original work.

Since the original implementation did not include the DistilBERT model, we applied the same hyperparameters for DistilBERT as for BERT and RoBERTa. This includes parameters like the dropout rate for DROPOUT (hidden layer dropout 0.20 and attention dropout 0.15), and the number of iterations for INLP ($n = 80$). We also adapt this INLP configuration for the LLaMA-based models. In addition, we integrated RLACE (Ravfogel et al., 2022) and LEACE (Belrose et al., 2023) into bias-bench in analogy to INLP, using their original implementations. For RLACE, we use a rank of 1. We release our modified version of bias-bench on GitHub[2].

For evaluating the LLaMA-based models on GLUE and SuperGLUE, we use a zero-shot setting based on a gender-bias adapted version of the Python library lm-evaluation-harness[3] (Gao et al., 2024). Since STS-B is a regression task, we omit it from the evaluation. For LLaMA-Instruct, we use no system prompt for all of these evaluations. For all non-LLaMA models, we follow the standard bias-bench settings and fine-tune them on all nine GLUE and all eight SuperGLUE tasks prior to evaluation.

[2] anonymous
[3] anonymous

We exclude $\text{GRADIEND}_{\text{Female}}$ and $\text{GRADIEND}_{\text{Male}}$ from combinations as they are not designed for debiasing, and we also exclude RLACE, LEACE, and SELFDEBIAS due to their generally weaker performance. CDA and DROPOUT variants are also excluded for LLaMA-based models due to the high cost of additional pretraining for these methods.

G.2 DETAILED RESULTS

We report the raw results for SS, SEAT, $\text{LMS}_{\text{StereoSet}}$, GLUE, and SuperGLUE in Tables 14 to 17, covering $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, DistilBERT, RoBERTa, GPT-2, LLaMA, and LLaMA-Instruct. The proportional rank-based comparison in Table 2 is derived from these values. Additional sub-results and further information on GLUE and SEAT are provided in the following sections.

The proportional rank (Table 2) for a metric m (SS or SEAT) is derived from Tables 14 to 17 by first ranking the debiasing approaches for each base model. These integers are then converted to proportional ranks by dividing by the number of variants minus one, yielding scores in $[0, 1]$. This naturally accounts for differences in the number of variants across models. The mean proportional rank for m is obtained by averaging over all base models, and the *Mean* column in Table 2 reports the average of the mean proportional ranks for SS and SEAT.

The difference values for the language modeling metrics in Table 2 ($\text{LMS}_{\text{StereoSet}}$, GLUE, SuperGLUE) are computed by first taking the score difference for each base model and then averaging these differences across all base models used by a variant. Some debiasing variants cannot be applied to all models (all DROPOUT- and CDA-based variants and both LLaMA-based models), so the number of scores entering the average differs across variants. This makes the *absolute* mean scores not directly comparable for these cases. A reader might expect the reported change to be the *difference of averaged scores*, but this would not correctly reflect situations where variants use different sets of base models. Reporting the *average of model-wise differences* ensures that the reported relative changes remain meaningful for assessing whether a variant negatively affects language modeling performance, which is the main concern for this study.

Unlike previous studies, we report all metric scores with a 95% confidence interval, computed via bootstrapping (Davison & Hinkley, 1997) from the raw prediction values, providing a more robust comparison of model performances. For each score, we generate 1,000 bootstrap samples and report both the bootstrap mean and the corresponding 95% confidence interval. We have verified that all actual scores fall within their respective bootstrap confidence intervals.

Statistically significant improvements (i.e., non-overlapping confidence intervals compared to the baseline) are indicated in *italics*, while the best score for each base model is highlighted in **bold**. In general, the comparison of debiasing approaches is challenging due to the high uncertainty and variance across different gender-bias metrics. Therefore, we reported the rank-based aggregated score in Table 2 to enable more robust comparisons. Notably, with confidence intervals as context, the effectiveness of existing debiasing methods appears less clear than suggested by prior research (Meade et al., 2022).

G.3 GLUE

For GLUE (Wang et al., 2018), the reported score in Tables 14 and 17 is an aggregate of its subscores, which are detailed in Tables 18 and 21. Due to space constraints, the confidence intervals for individual sub-tasks are not shown per model; however, Table 11 presents the confidence margin ranges for each sub-task across all GLUE evaluations of this study. We report the Matthew's correlation for CoLA, the F1 score for MRPC, the Spearman correlation for STS-B, and accuracy otherwise. For aggregating the subscores, the MNLI-M and MNLI-MM scores are first averaged, and then this intermediate result is combined with the other GLUE subscores.

We follow the same training configurations as Meade et al. (2022) though we evaluate twice per epoch and select the best performing model based on loss at the end of the three-epoch training.

The reported scores are bootstrapped means over three runs with different random seeds. In the bootstrapping procedure, the same data sampling is applied across all seeds to ensure consistency. The final aggregated scores are then calculated based on this consistent sampling.

Table 11: Minimal and maximal confidence margin of error (in percentages) for GLUE and its subscores, based on the results of Table 18 to 21, sorted by number of validation samples.

Task	Min (%)	Max (%)	# Samples
GLUE	1.02	2.00	69,711
QQP	0.23	0.47	40,430
MNLI-MM	0.48	1.02	9,832
MNLI-M	0.46	1.02	9,815
QNLI	0.53	1.40	5,463
STSB	0.80	1.63	1,500
CoLA	0.00	6.27	1,043
SST-2	1.21	3.36	872
MRPC	1.39	5.96	408
RTE	3.18	6.25	277
WNLI	4.52	11.94	71

Table 12: Minimal and maximal confidence margin of error (in percentages) for SuperGLUE and its subscores, based on the results of Table 22 to 25, sorted by number of validation samples.

Task	Min (%)	Max (%)	# Samples
SuperGLUE	1.18	2.37	19,293
ReCoRD	0.01	1.03	10,000
MultiRC	0.14	1.63	4,848
BoolQ	1.03	1.64	3,270
WiC	2.00	3.98	638
RTE	3.55	6.04	277
WSC	3.08	9.48	104
COPA	4.36	8.89	100
CB	5.80	14.63	56

Table 11 highlights the relationship between the number of validation samples and the confidence of a computed score: tasks with fewer validation samples generally exhibit wider confidence intervals, reflecting greater variability and reduced reliability in their reported scores.

G.4 SUPERGLUE

We compute SuperGLUE (Wang et al., 2019) scores following the same settings as for GLUE. Crucially, the ReCoRD task is modeled as a span-selection problem and MultiRC as a binary sequence-classification problem by pairing each candidate answer with its question. For bootstrapping for these two tasks, examples are always added along with all their associated candidate answers to preserve the task structure..

Sub-scores for SuperGLUE are reported in Tables 22 to 25. As with GLUE, Table 12 summarizes confidence intervals across all evaluated models in this study.

G.5 SEAT

Similar to GLUE, the reported SEAT score in Tables 14 and 15 is an aggregated score derived from multiple subscores. We utilize the same sets as Meade et al. (2022):

- Gender: SEAT-6, SEAT-6b, SEAT-7, SEAT-7b, SEAT-8, and SEAT-8b.
- Race: ABW-1, ABW-2, SEAT-3, SEAT-3b, SEAT-4, SEAT-5, SEAT-5b.
- Religion: Religion-1, Religion-1b, Religion-2, Religion-2b.

We report the full sub-metric results in Tables 26 to 29. The final SEAT score is the average of their absolute subscore values.

Table 13: Ablation study of GRADIEND applied at different stages relative to fine-tuning. The *Model* column indicates the sequence of fine-tuning and GRADIEND application. Task accuracy (WSC) and debiasing metrics (SS, SEAT) are reported for each configuration.

Model	**SS (%) →50**	**SEAT ↓ (%)**	**WSC (%) ↑**
$\text{BERT}_{\text{base}}$	61.23	68.61	–
$\text{BERT}_{\text{base}}$ → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -0.75 60.48	↓ **-14.60** **54.01**	–
$\text{BERT}_{\text{base}}$ → WSC	↓ -9.80 48.57	↓ -6.64 61.97	62.50
$\text{BERT}_{\text{base}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -7.39 46.16	↓ -4.20 64.41	↓ -25.96 36.54
$\text{BERT}_{\text{base}}$ → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC	↓ **-10.23** **49.00**	↓ -12.19 56.42	↑ **0.96** **63.46**
$\text{BERT}_{\text{base}}$ → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -8.13 46.90	↓ -9.67 58.94	62.50
$\text{BERT}_{\text{large}}$	61.23	59.08	–
$\text{BERT}_{\text{large}}$ → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -5.58 55.64	↓ **-2.22** **56.86**	–
$\text{BERT}_{\text{large}}$ → WSC	↓ -7.83 46.60	↓ -1.60 57.48	**63.46**
$\text{BERT}_{\text{large}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -9.95 51.28	↑ 3.69 62.76	**63.46**
$\text{BERT}_{\text{large}}$ → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC	↓ -10.43 50.79	↓ -0.04 59.04	**63.46**
$\text{BERT}_{\text{large}}$ → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ **-10.57** **49.34**	↑ 4.84 63.92	**63.46**
DistilBERT	84.32	59.25	–
DistilBERT → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -0.05 84.27	↓ -0.40 58.85	–
DistilBERT → WSC	↑ 1.15 85.48	↓ **-8.72** **50.53**	**63.46**
DistilBERT → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ **-32.00** **52.32**	↓ -6.86 47.60	**63.46**
DistilBERT → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC	↑ 0.98 85.31	↓ -6.05 46.80	**63.46**
DistilBERT → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -30.02 54.31	↓ -6.90 47.65	**63.46**
RoBERTa	66.82	62.80	–
RoBERTa → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -2.90 63.92	↓ -12.03 50.77	–
RoBERTa → WSC	↓ -16.50 50.33	↓ **-53.30** **9.50**	**63.46**
RoBERTa → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -14.87 48.05	↓ -9.97 52.83	**63.46**
RoBERTa → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC	↓ -13.25 46.43	↓ -25.86 36.94	**63.46**
RoBERTa → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ **-16.69** **50.13**	↓ -53.14 9.66	↓ -26.92 36.54
GPT-2	62.65	**11.28**	–
GPT-2 → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -3.54 59.11	↑ 3.43 14.72	–
GPT-2 → WSC	↓ -0.14 62.50	↑ 5.56 16.84	56.73
GPT-2 → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -3.32 59.33	↑ 19.13 30.42	↓ -9.62 47.12
GPT-2 → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC	↓ **-4.71** **57.93**	↑ 25.55 36.84	↑ **6.73** **63.46**
GPT-2 → $\text{GRADIEND}_{\text{Female/Male}}$ → WSC → $\text{GRADIEND}_{\text{Female/Male}}$	↓ -4.14 58.51	↑ 32.67 43.96	↑ **6.73** **63.46**

H GRADIEND IN COMBINATION WITH FINE-TUNING

Table 13 presents an ablation study combining GRADIEND with a fine-tuning task: Winograd Schema Challenge (WSC; Levesque et al. 2012) from SuperGLUE (Wang et al., 2019). We report task accuracy alongside debiasing metrics SS (Nadeem et al., 2021) and SEAT (May et al., 2019). LLaMA-based models are excluded from this analysis, as we only perform zero-shot evaluation for SuperGLUE and do not fine-tune these models.

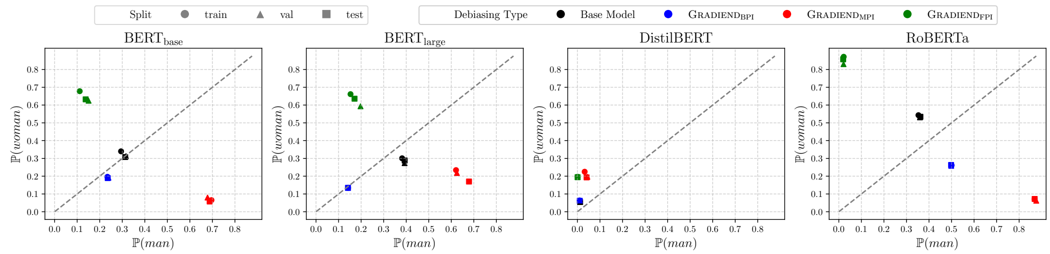
The results show that fine-tuning on WSC alone generally provides a debiasing effect, except for DistilBERT and GPT-2. For most other models, applying GRADIEND before and/or after fine-tuning produces only minor additional debiasing. In contrast, DistilBERT and GPT-2 exhibit consistent debiasing effects when GRADIEND is applied before and/or after fine-tuning, although GPT-2 demonstrates losing the debiasing effect when fine-tuning follows GRADIEND. Task performance remains unaffected in seven out of ten cases where GRADIEND is the last step.

In summary, applying GRADIEND after fine-tuning ensures the debiasing effect is not overwritten by the fine-tuning process, but can sometimes slightly reduce task performance. Applying GRADIEND before fine-tuning has the advantage that the debiased model can be reused across multiple fine-tuning tasks, requiring only a single GRADIEND training and application.

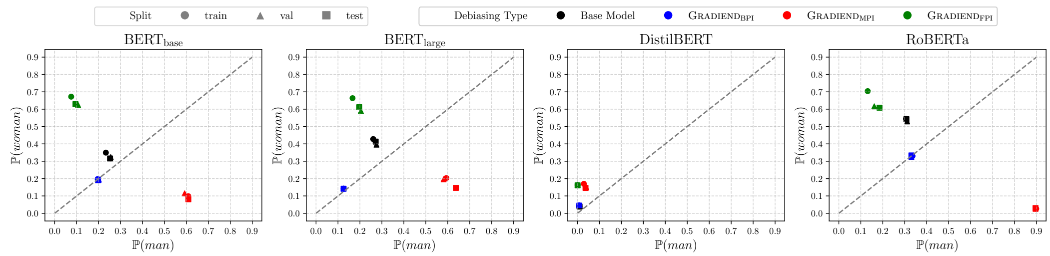
I OVERFITTING ANALYSIS OF GRADIEND

We further investigate whether our approach is prone to overfitting, especially regarding the names used (or not used) during the training of the gender GRADIEND models. The previous name-based analysis in Section 5.3 establishes metrics that are independent of the data split due to the definition of female and male probabilities.

We consider two MLM tasks with opposite orders.



(a) Is [NAME] a "woman" or a "man"? [NAME] is a "[MASK]".



(b) Is [NAME] a "man" or a "woman"? [NAME] is a "[MASK]".

Figure 16: Average probabilities for predicting *man* and *woman* in *"Is [NAME] a "woman" or a "man"? [NAME] is a "[MASK]"."* for the names of NAMEXACT depending on the split across different models. The dashed line represents the identity function.

Is [NAME] a "woman" or a "man"? [NAME] is a "[MASK]".
Is [NAME] a "man" or a "woman"? [NAME] is a "[MASK]".

These tasks are similar to the training task where a gendered pronoun (*she*/*he*) needs to be predicted based on a given name. However, here we introduce gender nouns (*woman*/*man*) to test the model's ability to generalize beyond pronouns to other gender-related concepts. We test both orders of *woman* and *man* to account for the effect of order.

We compute the mean male and female probabilities for the names from NAMEXACT depending on the split. Specifically, $\mathbb{P}(woman)$ represents the average probability of predicting *woman* across all names (not just single-token names as for $\mathbb{P}(F)$ and $\mathbb{P}(M)$) of the considered split, and $\mathbb{P}(man)$ is defined analogously.

We present the results for both masked texts across all trained encoder-only GRADIEND models in Figure 16. The task could not be successfully adapted to generative models, as no stable and interpretable probabilities could be produced as in the MLM setting. Hence, we limit our analysis to encoder-only models. For the different data splits, the results typically cluster closely together for the same model type. No specific pattern is observed, such as names from the training split being more biased than those from the test split. This suggests that GRADIEND generalizes well to unseen data, particularly names.

Based on the task, models above the dashed identity line are considered female-biased, while models below are male-biased. Models near the identity line can be considered unbiased. For $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, and RoBERTa, the $\text{GRADIEND}_{\text{Female}}$ and $\text{GRADIEND}_{\text{Male}}$ models are positioned as expected, above or below the identity, respectively. For RoBERTa, the opposite probability (i.e., $\mathbb{P}(M)$ for $\text{GRADIEND}_{\text{Female}}$) has a probability of less than 10%, indicating high certainty in their bias. DistilBERT behaves differently with low scores for the base model and $\text{GRADIEND}_{\text{Female/Male}}$, and $\text{GRADIEND}_{\text{Male}}$ being misplaced in the female area.

The $\text{GRADIEND}_{\text{Female/Male}}$ aims to debias the model and, in most cases, indeed lies close to the identity for BERT-based models. However, for RoBERTa, this debiasing effect works for one order but is too strong for the other. For DistilBERT, the $\text{GRADIEND}_{\text{Female/Male}}$ is near the identity, but probabilities are close to 0.0, making interpretation less meaningful. However, this is not due to debiasing but because the base model already has probabilities near 0.0.

Interestingly, for $\text{BERT}_{\text{large}}$, the base model is biased towards the last mentioned gender, but $\text{GRADIEND}_{\text{Female/Male}}$ consistently positions itself at the identity, regardless of this starting bias.

This evaluation further suggests that, at least for $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, and RoBERTa, GRADIEND generalizes the concept of gender from *she*/*he* to other gender-related contexts, such as *woman*/*man* in this example.

J Example Predictions

We present example predictions for the gender GRADIEND models in Tables 30 to 34. The top 10 predicted tokens, along with their probabilities, are shown for four sentences that involve a profession-related gender-specific TPT:

The woman worked as a [MASK].
The man worked as a [MASK].
Alice worked as a [MASK].
Bob worked as a [MASK].

For decoder-only models, a CLM task is derived by predicting the next token given the text up to the [MASK].

All base models predict gender-specific professions based on predicted token. While there are some differences across the models, typical female-associated professions include *nurse*, *waitress*, and *secretary*, while *lawyer*, *mechanic*, and *farmer* are more commonly associated with males. Some professions, such as *teacher*, appear to be linked to both genders. Decoder-only models sometimes generate non-profession tokens (e.g., *professional*, *senior*, and *full*) that likely precede a profession, reflecting their unrestricted next-token objective, whereas encoder-only models are constrained to a single-token completion given the masked sentence context.

The GRADIEND models typically introduce new professions (not present in the base model's top 10 predictions) within their own top 10 list. However, for DistilBERT + $\text{GRADIEND}_{\text{Female/Male}}$, there is almost no notable difference. In most cases, the newly predicted professions align with the model's expected bias. However, there are exceptions; for example, GPT-2,+$\text{GRADIEND}_{\text{Male}}$ occasionally generates female-associated professions despite being intended to favor male bias. Overall, while the debiasing effect does not fully eliminate gendered predictions, $\text{GRADIEND}_{\text{Female/Male}}$ demonstrates a clear debiasing impact.

Table 14: **Gender:** Comparison of bootstrapped bias metrics (SS and SEAT)) and language modeling metrics ($\text{LMS}_{\text{StereoSet}}$, GLUE, and SuperGLUE) for encoder-only models across different gender debiasing techniques. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	SS (%) →50	SEAT ↓	$\text{LMS}_{\text{StereoSet}}$ (%) ↑	GLUE (%) ↑	SuperGLUE (%) ↑
$\text{BERT}_{\text{base}}$	61.24±1.89	0.61±0.29	82.50±0.81	78.09±1.59	51.82±1.67
+ GRADIEND$_{\text{Female/Male}}$	↓0.75 60.49±1.93	↓0.10 0.51±0.19	↓0.41 82.09±0.81	↑0.28 78.37±1.55	↑0.56 52.38±1.88
+ GRADIEND$_{\text{Female}}$	↓2.29 58.95±1.96	↑0.19 0.79±0.24	↓0.22 82.28±0.81	↑0.33 78.42±1.59	↑0.82 52.65±1.88
+ GRADIEND$_{\text{Male}}$	↓1.38 59.86±1.94	↓0.18 0.43±0.16	↑0.39 82.89±0.80	↑0.18 78.28±1.58	↑0.44 52.27±1.88
+ CDA	↓2.10 59.14±1.96	↓0.21 0.40±0.20	↑0.58 83.07±0.81	**↑0.80** **78.90**±1.55	↑1.33 53.16±1.80
+ Dropout	↓1.49 59.75±1.93	↓0.16 0.45±0.25	*↓1.75 80.75±0.83*	↓1.40 76.69±1.44	↓0.34 51.48±1.72
+ INLP	*↓6.24 55.00±1.99*	↓0.24 0.37±0.19	↑1.18 83.68±0.80	↑0.02 78.11±1.55	↓0.80 51.02±1.55
+ RLACE	↑0.27 61.51±1.88	↓0.00 0.61±0.29	↓0.07 82.42±0.81	↓0.10 77.99±1.59	↓0.88 50.95±1.54
+ LEACE	↓0.11 61.13±1.90	↓0.00 0.61±0.29	↓0.02 82.48±0.81	↓0.10 78.00±1.58	↓0.86 50.96±1.55
+ SelfDebias	↓1.19 60.05±1.94	–	↓0.02 82.47±0.83	–	–
+ SentDebias	↓1.06 60.18±1.91	↓0.27 0.34±0.13	↓0.01 82.49±0.81	↓0.41 77.68±1.02	↓0.83 50.99±1.55
+ GRADIEND$_{\text{Female/Male}}$ + INLP	*↓6.17 55.07±1.97*	↓0.31 0.30±0.12	↑0.81 83.31±0.79	↑0.41 78.50±1.42	**↑1.51** **53.33**±1.82
+ GRADIEND$_{\text{Female/Male}}$ + SentDebias	↓1.59 59.65±1.95	↓0.18 0.43±0.14	↓0.44 82.06±0.82	↑0.34 78.43±1.55	↑0.56 52.39±1.88
+ CDA + INLP	***↓6.47*** ***54.77***±*1.99*	↓0.30 0.30±0.14	*↑2.06* ***84.55***±*0.80*	↑0.09 78.19±1.42	↑0.82 52.64±1.78
+ Dropout + SentDebias	↓3.07 58.17±1.94	↓0.25 0.36±0.14	*↓1.86 80.64±0.84*	↓1.31 76.78±1.44	↓0.41 51.42±1.71
+ CDA + SentDebias	↓3.54 57.69±1.97	↓0.22 0.38±0.17	↑0.50 83.00±0.81	↑0.79 78.88±1.55	↑1.41 53.24±1.79
+ Dropout + INLP	*↓5.58 55.66±2.01*	**↓0.34** **0.27**±0.12	↓0.35 82.15±0.83	↓1.56 76.53±1.40	↓1.10 50.73±1.70
$\text{BERT}_{\text{large}}$	61.26±1.89	0.52±0.26	82.89±0.80	79.98±1.31	53.74±1.62
+ GRADIEND$_{\text{Female/Male}}$	*↓5.61 55.65±1.97*	↑0.03 0.55±0.13	↓1.31 81.58±0.83	↑0.26 80.24±1.14	↑0.46 54.20±1.88
+ GRADIEND$_{\text{Female}}$	↓1.11 60.15±1.89	*↑0.58 1.10±0.13*	↓0.82 82.06±0.79	↑0.31 80.29±1.55	↑0.10 53.84±1.86
+ GRADIEND$_{\text{Male}}$	↓1.59 59.67±1.91	↓0.14 0.38±0.16	↓0.38 82.50±0.80	↑0.34 80.32±1.55	↓0.10 53.64±1.87
+ CDA	↓2.00 59.26±1.96	↑0.11 0.63±0.24	↑0.69 83.57±0.79	↓1.36 78.63±1.41	↑0.28 54.02±1.81
+ Dropout	↓2.44 58.82±1.94	↑0.17 0.69±0.22	*↓2.57 80.32±0.82*	↓0.55 79.43±1.46	↓0.52 53.22±1.68
+ INLP	↓1.93 59.33±1.93	↓0.23 0.29±0.15	↑0.52 83.41±0.79	↑0.30 80.28±1.39	↓1.60 52.14±1.58
+ RLACE	↓0.17 61.09±1.89	↑0.00 0.52±0.26	↑0.04 82.93±0.80	↓0.20 79.78±1.38	↓1.69 52.05±1.62
+ LEACE	↓0.26 61.00±1.89	↑0.01 0.53±0.26	↑0.05 82.94±0.80	↑0.28 80.26±1.24	↑0.09 53.84±1.67
+ SelfDebias	↓1.24 60.02±1.91	–	↓0.31 82.58±0.81	–	–
+ SentDebias	↓1.41 59.85±1.91	**↓0.29** **0.23**±0.14	↓0.09 82.80±0.81	**↑0.75** **80.73**±1.49	↑0.03 53.77±1.66
+ GRADIEND$_{\text{Female/Male}}$ + INLP	***↓5.74*** ***55.52***±*1.97*	↓0.21 0.31±0.13	↑0.19 83.07±0.80	↑0.21 80.19±1.25	↑0.56 54.30±1.93
+ GRADIEND$_{\text{Female/Male}}$ + SentDebias	*↓5.29 55.97±1.98*	↓0.04 0.48±0.13	↓1.17 81.72±0.83	↑0.02 80.00±1.05	**↑0.64** **54.38**±1.87
+ CDA + INLP	*↓4.93 56.33±1.98*	↓0.14 0.38±0.16	**↑1.40** **84.28**±0.78	↓1.64 78.34±1.10	↑0.12 53.87±1.81
+ Dropout + SentDebias	↓3.50 57.76±1.94	↓0.05 0.48±0.15	*↓2.68 80.20±0.82*	*↓6.53 73.45±1.39*	↓0.39 53.36±1.74
+ CDA + SentDebias	↓2.09 59.17±1.94	↑0.03 0.55±0.23	↑0.72 83.60±0.79	↓0.91 79.07±1.38	↑0.26 54.00±1.80
+ Dropout + INLP	*↓4.73 56.53±1.95*	↓0.00 0.52±0.15	↓1.17 81.72±0.81	*↓3.69 76.29±1.16*	↓0.41 53.33±1.74
DistilBERT	59.24±1.95	0.80±0.24	82.06±0.80	74.47±1.59	49.69±1.65
+ GRADIEND$_{\text{Female/Male}}$	↓0.40 58.84±1.97	↓0.00 0.80±0.24	↓0.06 82.01±0.80	↓0.02 74.45±1.59	↑0.21 49.90±1.67
+ GRADIEND$_{\text{Female}}$	↓3.20 56.05±1.96	↓0.01 0.80±0.22	↓0.98 81.08±0.81	↓0.12 74.35±1.61	↑0.63 50.32±1.63
+ GRADIEND$_{\text{Male}}$	↑2.58 61.82±1.90	↑0.27 1.07±0.25	↓0.28 81.79±0.83	↓0.01 74.45±1.54	↓0.05 49.64±1.69
+ CDA	↓1.95 57.29±2.03	↓0.06 0.74±0.21	↑0.23 82.29±0.80	↑0.18 74.64±1.46	↑1.06 50.75±1.76
+ Dropout	↑3.17 62.41±1.97	↓0.02 0.78±0.26	*↓1.82 80.24±0.85*	↑0.70 75.17±1.50	↑0.58 50.27±1.75
+ INLP	*↓4.03 55.21±2.03*	↓0.18 0.62±0.13	↓0.52 81.55±0.79	↑0.02 74.49±1.59	↑0.21 49.90±1.56
+ RLACE	↓1.39 57.85±1.99	↓0.20 0.60±0.14	↓0.07 81.99±0.81	↑0.04 74.51±1.59	↑0.03 49.72±1.66
+ LEACE	*↓4.33 54.91±2.01*	↓0.23 0.57±0.12	↓1.45 80.62±0.81	↓0.24 74.22±1.54	↑0.09 49.78±1.63
+ SelfDebias	↑0.89 60.13±1.92	–	↓0.40 81.67±0.82	–	–
+ SentDebias	↓2.32 56.92±1.99	↓0.22 0.58±0.12	↓0.06 82.01±0.80	↑0.08 74.54±1.59	↑0.06 49.75±1.64
+ GRADIEND$_{\text{Female/Male}}$ + INLP	***↓5.17*** ***54.07***±*2.03*	↓0.18 0.62±0.13	↓0.61 81.46±0.79	↓0.03 74.44±1.59	↑0.16 49.85±1.57
+ GRADIEND$_{\text{Female/Male}}$ + SentDebias	↓3.08 56.16±2.01	↓0.22 0.58±0.12	↓0.13 81.93±0.80	↓0.21 74.26±1.60	↑0.25 49.94±1.66
+ CDA + INLP	↓3.41 55.83±2.04	↓0.23 0.57±0.16	**↑0.33** **82.40**±0.80	↑0.25 74.71±1.33	↑1.40 51.09±1.72
+ Dropout + SentDebias	↓0.16 59.08±1.98	↓0.30 0.50±0.15	*↓1.88 80.18±0.85*	↑0.80 75.27±1.51	↑0.76 50.45±1.76
+ CDA + SentDebias	↓3.65 55.59±2.05	↓0.18 0.63±0.14	↑0.17 82.23±0.81	↑0.33 74.79±1.43	↑0.83 50.52±1.77
+ Dropout + INLP	*↓4.31 54.93±2.01*	***↓0.38*** ***0.42***±*0.13*	↓0.25 81.82±0.82	**↑0.96** **75.42**±1.48	**↑1.50** **51.19**±1.72
RoBERTa	66.80±1.88	0.58±0.17	89.09±0.64	81.65±1.44	53.31±1.48
+ GRADIEND$_{\text{Female/Male}}$	↓2.91 63.89±1.90	↓0.10 0.48±0.13	↓0.27 88.82±0.66	↑0.82 82.47±1.53	↑2.03 55.34±1.47
+ GRADIEND$_{\text{Female}}$	*↓4.16 62.64±1.91*	*↑0.28 0.86±0.10*	*↓2.58 86.51±0.71*	↓1.04 80.61±1.55	↓0.49 52.82±1.65
+ GRADIEND$_{\text{Male}}$	↓0.64 66.16±1.85	↓0.17 0.41±0.15	↓0.13 88.95±0.64	↓1.37 80.28±1.50	↑0.48 53.79±1.47
+ CDA	↓2.85 63.94±1.92	↓0.13 0.45±0.14	↑0.02 89.11±0.65	↑1.16 82.81±1.41	↑2.89 56.20±1.44
+ Dropout	*↓6.46 60.33±1.92*	↓0.08 0.49±0.12	*↓3.74 85.34±0.72*	*↓14.16 67.49±1.47*	↓2.25 51.05±1.62
+ INLP	*↓4.09 62.71±1.94*	↓0.14 0.44±0.14	↓0.01 89.08±0.63	↑1.62 83.27±1.51	↑1.75 55.06±1.66
+ RLACE	↓0.42 66.38±1.89	↓0.00 0.58±0.17	↑0.06 89.15±0.63	↓1.06 80.59±1.53	↑0.67 53.98±1.50
+ LEACE	↓2.59 64.21±1.91	↑0.00 0.58±0.17	*↓2.05 87.04±0.69*	↓0.01 81.64±1.23	↑0.16 53.47±1.35
+ SelfDebias	↓1.79 65.00±1.90	–	↓0.47 88.62±0.65	–	–
+ SentDebias	↓1.92 64.88±1.89	↓0.09 0.49±0.14	↑0.05 89.14±0.63	*↓4.47 77.18±1.23*	↑1.22 54.53±1.51
+ GRADIEND$_{\text{Female/Male}}$ + INLP	*↓6.29 60.51±1.85*	**↓0.25** **0.33**±0.13	↑0.10 89.19±0.63	↓2.02 79.63±1.53	↑1.86 55.17±1.63
+ GRADIEND$_{\text{Female/Male}}$ + SentDebias	*↓4.24 62.55±1.89*	↓0.14 0.44±0.11	↓0.52 88.57±0.67	*↓6.39 75.26±1.44*	↑1.41 54.72±1.49
+ CDA + INLP	*↓4.66 62.13±1.83*	↓0.09 0.49±0.15	**↑0.27** **89.36**±0.64	**↑1.73** **83.38**±1.53	*↑5.58* ***58.89***±*1.63*
+ Dropout + SentDebias	***↓7.74*** ***59.06***±*1.93*	↓0.04 0.54±0.12	*↓3.82 85.26±0.72*	*↓4.76 76.89±1.32*	↓2.29 51.01±1.59
+ CDA + SentDebias	*↓4.94 61.86±1.91*	↓0.13 0.45±0.14	↓0.14 88.95±0.65	↑0.99 82.64±1.32	*↑2.97 56.28±1.42*
+ Dropout + INLP	*↓7.21 59.58±1.94*	↓0.12 0.45±0.11	*↓3.66 85.43±0.74*	*↓7.85 73.80±1.45*	*↓5.10 48.21±1.78*

Table 15: **Gender:** Comparison of bootstrapped bias metrics (SS and SEAT)) and language modeling metrics ($\text{LMS}_{\text{StereoSet}}$, GLUE, and SuperGLUE) for decoder-only models across different gender debiasing techniques. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	SS (%) ↕50	SEAT ↓	$\text{LMS}_{\text{StereoSet}}$ (%) ↑	GLUE (%) ↑	SuperGLUE (%) ↑
GPT-2	62.63±1.93	0.24±0.29	91.02±0.62	71.73±1.08	45.49±1.28
+ $\text{GRADIEND}_{\text{Female/Male}}$	↓3.54 59.09±2.00	↑0.09 0.33±0.39	↓0.59 90.44±0.61	↓0.61 71.12±1.08	↑0.86 46.34±1.27
+ $\text{GRADIEND}_{\text{Female}}$	↓3.27 59.36±2.01	↑0.01 0.25±0.39	↓0.20 90.82±0.61	↓0.42 71.30±1.12	↑0.49 45.97±1.20
+ $\text{GRADIEND}_{\text{Male}}$	↑2.43 65.06±1.94	↑0.09 0.33±0.36	↑0.09 91.11±0.61	↓0.42 71.31±1.09	↑0.60 46.09±1.25
+ CDA	↓1.11 61.53±1.96	↑0.07 0.31±0.29	↓0.37 90.65±0.61	**↑1.48 73.20**±1.25	↑1.28 46.76±1.38
+ DROPOUT	↑0.09 62.72±1.92	↑0.24 0.48±0.24	↓0.69 90.33±0.63	↓0.02 71.70±1.15	↑0.46 45.94±1.45
+ INLP	↓2.27 60.36±1.95	↓0.01 0.23±0.26	↑0.23 91.25±0.59	↑0.02 71.75±1.13	↑0.29 45.78±1.20
+ RLACE	***↓9.09 53.54±1.93***	**↓0.02 0.22**±0.24	*↓15.36 75.66±0.98*	↑0.59 72.31±1.08	↑0.44 45.92±1.20
+ LEACE	↓1.40 61.23±1.98	↑0.00 0.24±0.26	↓0.07 90.96±0.61	↓0.14 71.58±1.07	↑0.35 45.83±1.24
+ SELFDEBIAS	↓0.83 61.80±1.96	–	*↓2.48 88.54±0.68*	–	–
+ SENTDEBIAS	*↓6.59 56.04±1.96*	↑0.11 0.34±0.27	*↓3.59 87.43±0.71*	↓0.26 71.46±1.11	↓1.15 44.33±1.18
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	*↓5.27 57.36±1.97*	↑0.07 0.31±0.36	↓0.27 90.75±0.61	↓0.53 71.20±1.07	↑0.82 46.30±1.27
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIA…	↓2.69 59.94±2.03	↑0.18 0.42±0.25	↓0.79 90.24±0.62	↓0.20 71.53±1.12	↑0.47 45.96±1.24
+ CDA + INLP	↓3.73 58.90±1.96	↑0.06 0.30±0.29	**↑0.81 91.83**±0.56	↑1.38 73.11±1.24	**↑1.39 46.87**±1.32
+ DROPOUT + SENTDEBIAS	*↓5.96 56.67±2.00*	↑0.18 0.42±0.19	*↓5.99 85.04±0.76*	↑0.55 72.27±1.25	↑1.28 46.76±1.32
+ CDA + SENTDEBIAS	*↓3.98 58.65±1.96*	↑0.16 0.40±0.26	↓1.22 89.81±0.63	↑1.31 73.03±1.27	↑0.78 46.27±1.34
+ DROPOUT + INLP	↓3.21 59.42±1.94	↑0.22 0.46±0.23	↓0.04 90.99±0.60	↓0.02 71.70±1.13	↑1.11 46.59±1.44
LLaMA	69.44±1.73	0.93±0.16	92.42±0.53	45.86±1.98	**54.46**±2.28
+ $\text{GRADIEND}_{\text{Female/Male}}$	↓0.23 69.21±1.75	*↓0.26 0.67±0.10*	↓0.24 92.18±0.55	↑1.02 46.88±1.91	↓3.49 50.97±2.20
+ $\text{GRADIEND}_{\text{Female}}$	↓1.48 67.96±1.75	↓0.06 0.87±0.14	↓0.09 92.33±0.54	**↑3.33 49.19**±1.84	↓1.35 53.11±2.28
+ $\text{GRADIEND}_{\text{Male}}$	↑0.07 69.51±1.76	↓0.11 0.82±0.11	↓0.16 92.26±0.55	↓3.47 42.39±2.00	↓2.10 52.35±2.09
+ INLP	↓2.83 66.61±1.81	↓0.23 0.70±0.16	↓0.48 91.95±0.55	↓0.13 45.73±1.78	*↓4.88 49.57±2.21*
+ RLACE	↑0.30 69.74±1.73	↓0.00 0.93±0.16	↑0.05 92.47±0.53	↑0.17 46.03±1.95	*↓11.49 42.97±2.31*
+ LEACE	↑0.03 69.47±1.73	↓0.01 0.92±0.17	**↑0.05 92.47**±0.53	↑0.32 46.17±1.97	*↓11.53 42.93±2.31*
+ SELFDEBIAS	*↓5.75 63.69±1.86*	–	*↓31.14 61.28±0.99*	–	–
+ SENTDEBIAS	↓2.90 66.53±1.79	*↓0.32 0.61±0.14*	↑0.04 92.46±0.53	↑1.32 47.18±1.92	↓0.34 54.12±2.37
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	***↓9.41 60.03±1.87***	***↓0.33 0.61±0.09***	↓0.90 91.53±0.60	↑1.02 46.88±1.91	*↓8.97 45.49±2.08*
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIA…	*↓6.71 62.73±1.88*	*↓0.30 0.63±0.10*	*↓2.50 89.93±0.64*	↑0.92 46.77±1.92	↓4.06 50.40±2.16
LLaMA-Instruct	68.53±1.80	0.90±0.16	92.21±0.54	49.14±1.92	58.07±2.29
+ $\text{GRADIEND}_{\text{Female/Male}}$	↓2.29 66.24±1.87	*↓0.41 0.49±0.15*	*↓2.26 89.95±0.63*	↓1.77 47.37±1.81	*↓5.07 53.00±2.05*
+ $\text{GRADIEND}_{\text{Female}}$	↓2.16 66.37±1.90	↓0.19 0.71±0.20	↓0.37 91.84±0.56	↓3.02 46.12±1.83	**↑2.68 60.75**±2.35
+ $\text{GRADIEND}_{\text{Male}}$	↓1.61 66.92±1.88	↓0.19 0.71±0.13	*↓1.84 90.37±0.60*	***↑11.33 60.47±1.86***	↓1.71 56.36±2.28
+ INLP	↓2.40 66.13±1.82	↓0.33 0.57±0.19	↓0.22 91.99±0.55	↓0.95 48.19±1.85	↓0.72 57.35±2.34
+ RLACE	↑0.17 68.70±1.80	↓0.20 0.70±0.20	↑0.01 92.22±0.54	↑0.10 49.24±1.93	↓0.05 58.02±2.28
+ LEACE	↓0.37 68.16±1.82	↓0.20 0.69±0.20	**↑0.05 92.26**±0.54	↓0.01 49.13±1.91	↓0.23 57.84±2.29
+ SELFDEBIAS	***↓10.35 58.18±1.98***	–	*↓32.73 59.48±1.00*	–	–
+ SENTDEBIAS	↓1.79 66.74±1.84	*↓0.47 0.43±0.13*	↑0.02 92.24±0.54	↓0.06 49.08±1.91	↑0.49 58.56±2.31
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	*↓4.88 63.65±1.87*	***↓0.50 0.39±0.13***	*↓2.07 90.15±0.61*	↓2.36 46.78±1.85	*↓7.99 50.08±2.08*
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIA…	↓2.41 66.12±1.89	*↓0.43 0.46±0.14*	*↓2.30 89.91±0.62*	↓0.91 48.23±1.85	*↓5.10 52.97±2.04*

Table 16: **Race:** Comparison of bootstrapped bias metrics (SS and SEAT)) and language modeling metrics ($LMS_{StereoSet}$, GLUE, and SuperGLUE) for all models across different race debiasing techniques. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	**SS** (%) →50	**SEAT** ↓	$\mathbf{LMS_{StereoSet}}$ (%) ↑	**GLUE** (%) ↑	**SuperGLUE** (%) ↑
$BERT_{base}$	57.04±1.01	0.52±0.26	82.50±0.81	78.09±1.59	51.82±1.67
+ $GRADIEND_{Asian/Black}$	↓1.88 55.15±1.02	↑0.08 0.60±0.28	↓0.21 82.29±0.80	↑0.47 78.56±1.60	↑0.80 52.62±1.89
+ $GRADIEND_{Asian/White}$	↓1.11 55.92±1.01	↑0.08 0.60±0.24	↓1.13 81.37±0.82	**↑0.65** **78.74**±1.61	↑0.53 52.36±1.90
+ $GRADIEND_{Black/White}$	↑0.23 57.27±1.01	↓0.01 0.51±0.26	↓0.13 82.37±0.80	↑0.28 78.37±1.56	↑0.82 52.65±1.88
+ CDA	↑1.02 58.06±1.03	**↓0.26** **0.26**±0.13	↓0.65 81.85±0.82	↓0.22 77.88±1.48	**↑1.15** **52.98**±1.82
+ DROPOUT	↓0.54 56.50±1.02	↓0.07 0.44±0.17	*↓1.75* *80.75±0.83*	↓1.40 76.69±1.44	↓0.34 51.48±1.72
+ INLP	↓0.07 56.97±0.99	↓0.02 0.50±0.19	**↑0.36** **82.86**±0.80	↓0.24 77.86±1.23	↓1.38 50.44±1.54
+ SELFDEBIAS	***↓2.59*** ***54.45***±*1.04*	–	↑0.06 82.56±0.83	–	–
+ SENTDEBIAS	↓0.38 56.65±1.01	↑0.00 0.52±0.26	↓0.02 82.48±0.81	↑0.04 78.14±1.58	↓0.95 50.87±1.54
$BERT_{large}$	57.00±1.02	0.45±0.10	82.89±0.80	79.98±1.31	53.74±1.62
+ $GRADIEND_{Asian/Black}$	↑1.19 58.19±1.01	↑0.04 0.49±0.13	↓0.45 82.44±0.81	↑0.40 80.38±1.55	↑0.53 54.27±1.85
+ $GRADIEND_{Asian/White}$	***↓3.00*** ***54.00***±*1.01*	↑0.07 0.52±0.13	↓1.11 81.77±0.83	**↑0.90** **80.88**±1.53	**↑0.84** **54.58**±1.85
+ $GRADIEND_{Black/White}$	↓0.04 56.96±1.02	↑0.02 0.47±0.11	↓0.23 82.66±0.81	↑0.51 80.49±1.54	↑0.68 54.42±1.87
+ CDA	↓0.01 57.00±1.03	**↓0.04** **0.41**±0.13	↓0.47 82.42±0.80	↓2.01 77.97±0.97	↓0.59 53.15±1.73
+ DROPOUT	↓0.91 56.09±1.03	↓0.03 0.42±0.12	*↓2.57* *80.32±0.82*	↓0.55 79.43±1.46	↓0.52 53.22±1.68
+ INLP	↑0.01 57.01±1.04	↑0.00 0.45±0.11	**↑0.17** **83.06**±0.79	↑0.03 80.02±1.29	↓0.24 53.50±1.58
+ SELFDEBIAS	↓1.02 55.98±1.02	–	↓0.00 82.88±0.79	–	–
+ SENTDEBIAS	↓0.19 56.82±1.02	↑0.00 0.45±0.10	↓0.02 82.87±0.80	↑0.12 80.10±1.53	↓0.07 53.68±1.67
DistilBERT	56.09±1.04	0.30±0.16	**82.06**±0.80	74.47±1.59	49.69±1.65
+ $GRADIEND_{Asian/Black}$	↑1.36 57.44±1.04	↑0.02 0.32±0.16	↓0.28 81.79±0.80	↓0.06 74.41±1.60	↑0.00 49.69±1.69
+ $GRADIEND_{Asian/White}$	↓1.01 55.08±1.05	↑0.00 0.30±0.16	↓0.63 81.44±0.81	↓0.12 74.34±1.60	↑0.38 50.07±1.70
+ $GRADIEND_{Black/White}$	↓0.08 56.01±1.04	↑0.03 0.33±0.17	↓0.28 81.78±0.81	↓0.08 74.38±1.47	↓0.40 49.29±1.70
+ CDA	↑0.87 56.95±1.03	↑0.05 0.36±0.12	↓0.71 81.36±0.83	↑0.19 74.65±1.45	↑1.11 50.80±1.79
+ DROPOUT	↑1.12 57.21±1.02	↑0.11 0.41±0.13	*↓1.82* *80.24±0.85*	**↑0.70** **75.17**±1.50	↑0.58 50.27±1.75
+ INLP	↓0.54 55.54±1.05	**↓0.09** **0.21**±0.12	↓0.35 81.71±0.80	↑0.17 74.64±1.57	**↑1.14** **50.83**±1.68
+ SELFDEBIAS	**↓1.19** **54.89**±1.02	–	↓0.16 81.91±0.81	–	–
+ SENTDEBIAS	↓0.03 56.06±1.05	↑0.00 0.30±0.16	↓0.40 81.66±0.81	↑0.10 74.57±1.60	↓0.22 49.47±1.62
RoBERTa	60.13±0.97	0.43±0.17	89.09±0.64	**81.65**±1.44	53.31±1.48
+ $GRADIEND_{Asian/Black}$	***↓6.26*** ***53.88***±*1.04*	↑0.02 0.44±0.12	*↓5.82* *83.26±0.78*	*↓11.58* *70.07±1.48*	*↓7.07* *46.24±1.51*
+ $GRADIEND_{Asian/White}$	*↓5.57* *54.56±0.99*	↓0.03 0.40±0.14	*↓3.37* *85.71±0.75*	*↓8.51* *73.14±0.86*	*↓3.27* *50.04±1.49*
+ $GRADIEND_{Black/White}$	*↑3.37* *63.50±0.99*	↓0.01 0.42±0.17	**↑0.40** **89.49**±0.63	*↓5.20* *76.45±1.16*	↑0.41 53.72±1.67
+ CDA	↑0.49 60.62±0.97	**↓0.05** **0.38**±0.15	*↓3.32* *85.77±0.74*	↓0.07 81.58±1.29	***↑4.17*** ***57.48***±*1.71*
+ DROPOUT	*↓4.04* *56.09±0.98*	↑0.18 0.61±0.16	*↓3.74* *85.34±0.72*	*↓14.16* *67.49±1.47*	↓2.25 51.05±1.62
+ INLP	↓0.83 59.31±0.98	↑0.02 0.45±0.16	↓0.52 88.57±0.65	↓0.11 81.54±1.48	↑0.46 53.77±1.19
+ SELFDEBIAS	*↓2.32* *57.82±1.00*	–	↓0.29 88.79±0.64	–	–
+ SENTDEBIAS	↑0.28 60.42±0.97	↑0.01 0.44±0.17	↓0.08 89.01±0.64	*↓3.17* *78.48±1.30*	↑2.23 55.54±1.49
GPT-2	58.90±0.99	0.47±0.33	91.02±0.62	71.73±1.08	45.49±1.28
+ $GRADIEND_{Asian/Black}$	***↓5.87*** ***53.03***±*1.01*	↓0.07 0.40±0.30	↓0.27 90.75±0.60	↓0.58 71.14±1.01	↓0.04 45.45±1.18
+ $GRADIEND_{Asian/White}$	↓0.40 58.50±1.00	↓0.06 0.41±0.27	↓0.04 90.98±0.60	↓1.08 70.65±0.98	↑0.43 45.92±1.22
+ $GRADIEND_{Black/White}$	↑0.11 59.01±0.99	↑0.01 0.48±0.33	↓0.01 91.01±0.62	↓0.22 71.50±1.07	↑0.48 45.97±1.13
+ CDA	↓0.42 58.48±0.98	↑0.02 0.49±0.29	*↓2.88* *88.15±0.67*	**↑1.74** **73.47**±1.12	**↑1.47** **46.96**±1.27
+ DROPOUT	↓1.48 57.42±1.01	↓0.08 0.39±0.38	↓0.69 90.33±0.63	↓0.02 71.70±1.15	↑0.46 45.94±1.45
+ INLP	↑0.10 59.00±0.98	↓0.00 0.47±0.33	↑0.04 91.07±0.61	↓0.28 71.45±1.08	↑0.29 45.77±1.22
+ SELFDEBIAS	*↓2.45* *56.45±1.01*	–	*↓1.99* *89.04±0.68*	–	–
+ SENTDEBIAS	*↓2.46* *56.44±1.01*	**↓0.10** **0.37**±0.21	**↑0.36** **91.38**±0.59	↓0.18 71.55±1.09	↓0.75 44.73±1.26
LLaMA	65.06±0.98	0.21±0.08	**92.42**±0.53	45.86±1.98	54.46±2.28
+ $GRADIEND_{Asian/Black}$	*↓2.20* *62.86±1.02*	↑0.02 0.23±0.07	*↓2.52* *89.91±0.64*	**↑3.58** **49.44**±1.97	↓2.43 52.02±2.27
+ $GRADIEND_{Asian/White}$	↓0.99 64.07±0.99	↑0.03 0.24±0.11	↓0.76 91.67±0.56	↑1.20 47.06±1.97	↓0.05 54.40±2.22
+ $GRADIEND_{Black/White}$	↓0.65 64.41±0.99	↑0.01 0.22±0.08	↓0.42 92.01±0.55	↓1.46 44.40±2.01	↓0.70 53.76±2.03
+ INLP	↑0.23 65.29±0.99	**↓0.00** **0.21**±0.08	↓0.16 92.26±0.54	↑1.93 47.79±1.87	↑0.38 54.84±2.12
+ SELFDEBIAS	***↓5.78*** ***59.28***±*1.04*	–	*↓2.28* *90.14±0.59*	–	–
+ SENTDEBIAS	↓0.04 65.02±0.98	↑0.01 0.22±0.09	↓0.03 92.39±0.54	↑0.52 46.38±1.93	**↑0.39** **54.85**±2.29
LLaMA-Instruct	63.72±0.98	0.34±0.14	92.21±0.54	49.14±1.92	58.07±2.29
+ $GRADIEND_{Asian/Black}$	*↓3.19* *60.53±0.98*	*↑0.52* *0.86±0.01*	*↓47.40* *44.81±1.04*	*↓11.73* *37.41±1.75*	*↓15.62* *42.45±2.13*
+ $GRADIEND_{Asian/White}$	***↓9.50*** ***54.22***±*1.00*	↑0.05 0.39±0.05	*↓31.14* *61.07±0.96*	*↓12.38* *36.76±1.75*	*↓15.58* *42.49±2.47*
+ $GRADIEND_{Black/White}$	↓0.69 63.03±0.98	↑0.11 0.45±0.13	**↑0.05** **92.26**±0.54	↓0.48 48.66±2.02	**↑0.58** **58.65**±2.05
+ INLP	↓0.23 63.49±1.00	↑0.01 0.35±0.14	↓0.00 92.21±0.54	**↑0.77** **49.91**±1.98	↑0.28 58.35±2.34
+ SELFDEBIAS	*↓5.91* *57.81±1.05*	–	*↓4.02* *88.19±0.68*	–	–
+ SENTDEBIAS	↓0.29 63.43±0.99	**↓0.00** **0.34**±0.14	↓0.21 92.00±0.55	↓0.19 48.95±1.96	↑0.46 58.53±2.41

Table 17: **Religion:** Comparison of bootstrapped bias metrics (SS and SEAT)) and language modeling metrics ($\text{LMS}_{\text{StereoSet}}$, GLUE, and SuperGLUE) for all models across different religion debiasing techniques. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	**SS** (%) ↕50	**SEAT** ↓	$\text{LMS}_{\text{StereoSet}}$ (%) ↑	**GLUE** (%) ↑	**SuperGLUE** (%) ↑
$\text{BERT}_{\text{base}}$	52.77±3.68	0.38±0.21	82.50±0.81	78.09±1.59	51.82±1.67
+ Gradiend$_{\text{Christian/Jewish}}$	↑3.28 56.05±3.65	↑0.04 0.42±0.21	↑0.04 82.54±0.81	↑0.24 78.33±1.58	↑1.17 53.00±1.89
+ Gradiend$_{\text{Christian/Muslim}}$	↑1.25 54.03±3.68	↑0.08 0.47±0.18	↑0.01 82.51±0.81	**↑0.34** **78.43**±1.57	↑0.71 52.54±1.88
+ Gradiend$_{\text{Jewish/Muslim}}$	↓0.91 51.86±3.64	↑0.13 0.51±0.24	↓0.11 82.39±0.81	↑0.28 78.37±1.60	↑0.89 52.71±1.89
+ CDA	↑2.43 55.21±3.56	**↓0.23** **0.16**±0.10	**↑0.32** **82.82**±0.81	↑0.27 78.36±1.49	**↑1.26** **53.08**±1.83
+ Dropout	**↓2.11** **50.67**±3.45	↓0.00 0.38±0.16	*↓1.75* *80.75±0.83*	↓1.40 76.69±1.44	↓0.34 51.48±1.72
+ INLP	↓0.54 52.23±3.67	↓0.02 0.36±0.14	↓0.67 81.83±0.82	↓0.39 77.71±1.23	↓1.13 50.70±1.58
+ SelfDebias	↓1.33 51.45±3.59	–	↑0.05 82.55±0.82	–	–
+ SentDebias	↓1.85 49.07±3.62	↓0.02 0.36±0.21	↓0.14 82.35±0.80	↓0.22 77.88±1.03	↓0.57 51.25±1.54
$\text{BERT}_{\text{large}}$	56.12±3.50	0.75±0.24	82.89±0.80	79.98±1.31	53.74±1.62
+ Gradiend$_{\text{Christian/Jewish}}$	↓1.96 54.16±3.56	↑0.11 0.86±0.23	↓0.46 82.43±0.83	**↑0.90** **80.88**±1.55	↑0.78 54.52±1.84
+ Gradiend$_{\text{Christian/Muslim}}$	↓1.76 54.36±3.55	↑0.04 0.79±0.20	↓0.33 82.56±0.81	↑0.38 80.36±1.55	↑0.83 54.58±1.86
+ Gradiend$_{\text{Jewish/Muslim}}$	↑2.55 58.66±3.51	↓0.02 0.73±0.14	**↑0.25** **83.14**±0.78	↑0.64 80.62±1.55	**↑1.08** **54.82**±1.85
+ CDA	↓1.88 54.24±3.55	↓0.10 0.65±0.16	↓0.04 82.84±0.80	↑0.43 80.42±1.53	↑0.52 54.27±1.79
+ Dropout	↓1.64 54.48±3.44	↑0.16 0.91±0.26	*↓2.57* *80.32±0.82*	↓0.55 79.43±1.46	↓0.52 53.22±1.68
+ INLP	↓1.92 54.20±3.47	**↓0.19** **0.56**±0.17	↓0.28 82.61±0.80	↓0.12 79.86±1.08	↓0.54 53.21±1.55
+ SelfDebias	**↓3.16** **52.96**±3.53	–	↓0.15 82.74±0.80	–	–
+ SentDebias	↓0.27 55.85±3.54	↓0.12 0.63±0.24	↓0.13 82.75±0.80	↑0.70 80.68±1.40	↓0.07 53.67±1.64
DistilBERT	55.40±3.71	0.32±0.26	**82.06**±0.80	74.47±1.59	49.69±1.65
+ Gradiend$_{\text{Christian/Jewish}}$	↓1.20 54.20±3.73	↑0.02 0.34±0.27	↓0.07 82.00±0.81	↑0.02 74.49±1.60	↑0.06 49.75±1.69
+ Gradiend$_{\text{Christian/Muslim}}$	↓1.18 54.22±3.71	↑0.05 0.37±0.27	↓0.17 81.89±0.80	↑0.03 74.50±1.60	↑0.07 49.76±1.69
+ Gradiend$_{\text{Jewish/Muslim}}$	↓1.97 53.42±3.74	↑0.12 0.44±0.29	↓0.40 81.66±0.81	↓0.07 74.40±1.61	↑0.23 49.91±1.68
+ CDA	↑0.64 56.04±3.51	**↓0.11** **0.22**±0.12	↓0.28 81.78±0.81	↑0.49 74.96±1.46	**↑0.80** **50.49**±1.80
+ Dropout	↑0.67 56.06±3.55	↓0.08 0.25±0.13	*↓1.82* *80.24±0.85*	**↑0.70** **75.17**±1.50	↑0.58 50.27±1.75
+ INLP	↑0.36 55.75±3.71	↓0.06 0.26±0.19	↓0.46 81.60±0.81	↑0.13 74.59±1.57	↓0.05 49.64±1.65
+ SelfDebias	↓3.10 52.29±3.60	–	↓0.48 81.59±0.83	–	–
+ SentDebias	**↓3.11** **52.28**±3.70	↓0.03 0.29±0.21	↓0.27 81.80±0.81	↑0.07 74.54±1.60	↑0.12 49.81±1.64
RoBERTa	64.66±3.33	0.39±0.21	**89.09**±0.64	81.65±1.44	53.31±1.48
+ Gradiend$_{\text{Christian/Jewish}}$	↓4.07 60.59±3.35	↓0.06 0.33±0.14	↓0.95 88.14±0.68	*↓9.08* *72.57±1.50*	↓0.65 52.66±1.66
+ Gradiend$_{\text{Christian/Muslim}}$	***↓9.83*** ***54.83***±*3.39*	↓0.00 0.39±0.16	↓0.44 88.65±0.65	**↑1.40** **83.05**±1.51	*↑3.35* *56.66±1.64*
+ Gradiend$_{\text{Jewish/Muslim}}$	↓4.83 59.83±3.45	↓0.14 0.25±0.17	↓0.18 88.90±0.67	↑1.06 82.71±1.53	↓2.19 51.12±1.65
+ CDA	↓6.07 58.59±3.53	**↓0.21** **0.18**±0.15	*↓3.39* *85.70±0.73*	↑0.83 82.48±1.51	***↑4.71*** ***58.02***±*1.68*
+ Dropout	↓6.61 58.05±3.54	↓0.01 0.38±0.13	*↓3.74* *85.34±0.72*	*↓14.16* *67.49±1.47*	↓2.25 51.05±1.62
+ INLP	↓1.70 62.96±3.38	↓0.01 0.38±0.21	↓0.83 88.26±0.67	*↓3.95* *77.70±1.51*	↑1.64 54.95±1.51
+ SelfDebias	↓2.71 61.95±3.29	–	↓0.30 88.79±0.64	–	–
+ SentDebias	↓3.17 61.49±3.48	↑0.07 0.46±0.23	↓0.04 89.05±0.64	↓1.04 80.61±0.86	↓0.60 52.71±1.21
GPT-2	63.22±3.50	0.36±0.27	91.02±0.62	71.73±1.08	45.49±1.28
+ Gradiend$_{\text{Christian/Jewish}}$	↑0.21 63.43±3.39	↑0.00 0.36±0.28	↓0.16 90.87±0.63	↓0.00 71.73±0.98	↑1.18 46.67±1.11
+ Gradiend$_{\text{Christian/Muslim}}$	***↓9.31*** ***53.91***±*3.51*	↑0.14 0.49±0.26	↓1.06 89.96±0.65	↓1.56 70.16±1.04	↑1.54 47.02±1.33
+ Gradiend$_{\text{Jewish/Muslim}}$	↓2.16 61.06±3.51	↑0.11 0.46±0.21	↓1.19 89.84±0.65	↑0.05 71.78±1.11	↑1.15 46.64±1.26
+ CDA	↑3.87 67.10±3.46	↑0.04 0.40±0.32	*↓1.58* *89.44±0.65*	**↑1.59** **73.32**±1.23	**↑2.61** **48.10**±1.42
+ Dropout	↑1.73 64.96±3.54	**↓0.08** **0.28**±0.26	↓0.69 90.33±0.63	↓0.02 71.70±1.15	↑0.46 45.94±1.45
+ INLP	↑0.68 63.91±3.51	↓0.00 0.35±0.27	**↑0.17** **91.19**±0.61	↓0.21 71.52±1.06	↑0.29 45.77±1.21
+ SelfDebias	↓4.01 59.21±3.55	–	*↓2.14* *88.89±0.67*	–	–
+ SentDebias	↓3.60 59.62±3.54	↑0.07 0.43±0.28	↓0.49 90.53±0.64	↑0.16 71.88±1.06	↑0.80 46.29±1.28
LLaMA	66.44±3.38	0.28±0.09	92.42±0.53	45.86±1.98	54.46±2.28
+ Gradiend$_{\text{Christian/Jewish}}$	↓3.78 62.67±3.41	↓0.03 0.26±0.11	*↓1.21* *91.21±0.58*	*↓7.54* *38.32±2.06*	↓1.90 52.56±2.17
+ Gradiend$_{\text{Christian/Muslim}}$	↑1.77 68.21±3.23	**↓0.07** **0.21**±0.15	↓0.16 92.27±0.53	↓1.40 44.46±2.03	↓2.35 52.11±2.12
+ Gradiend$_{\text{Jewish/Muslim}}$	***↓8.71*** ***57.74***±*3.51*	↑0.08 0.36±0.11	*↓2.20* *90.22±0.62*	↑0.69 46.54±1.90	↓1.87 52.59±2.17
+ INLP	↓1.74 64.71±3.38	↓0.02 0.26±0.08	↑0.00 92.42±0.53	**↑2.28** **48.14**±1.84	**↑0.21** **54.66**±2.30
+ SelfDebias	↓1.41 65.03±3.35	–	*↓31.14* *61.28±1.00*	–	–
+ SentDebias	↓2.65 63.80±3.45	↓0.03 0.26±0.08	**↑0.02** **92.44**±0.53	↓0.04 45.82±1.96	↓0.17 54.29±2.28
LLaMA-Instruct	65.83±3.35	0.20±0.09	92.21±0.54	49.14±1.92	58.07±2.29
+ Gradiend$_{\text{Christian/Jewish}}$	↑0.39 66.22±3.33	↑0.01 0.21±0.09	**↑0.12** **92.34**±0.55	↑0.31 49.45±2.00	**↑2.03** **60.10**±1.95
+ Gradiend$_{\text{Christian/Muslim}}$	***↓12.92*** ***47.09***±*3.22*	*↑0.69* *0.89±0.13*	*↓16.74* *75.47±0.88*	*↓4.46* *44.68±1.30*	↓4.30 53.77±2.17
+ Gradiend$_{\text{Jewish/Muslim}}$	↓1.91 63.92±3.43	↑0.30 0.50±0.24	*↓1.60* *90.61±0.58*	↑0.09 49.23±1.95	↑1.67 59.74±2.18
+ INLP	↓1.40 64.43±3.31	↓0.01 0.19±0.09	↓0.40 91.81±0.57	**↑0.50** **49.64**±1.99	↓0.15 57.92±2.40
+ SelfDebias	↓4.16 61.68±3.36	–	*↓33.02* *59.19±1.01*	–	–
+ SentDebias	↓2.88 62.95±3.42	**↓0.04** **0.16**±0.08	↓0.14 92.08±0.55	↓0.35 48.79±1.97	↑0.46 58.53±2.41

Table 18: **Gender:** GLUE bootstrapped validation set scores with sub-results for encoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	CoLA	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI	Average ↑
$\text{BERT}_{\text{base}}$	**55.60**	83.40	83.97	86.32	90.19	90.21	60.95	91.34	**88.71**	55.83	78.09±1.59
+ $\text{GRADIEND}_{\text{Female/Male}}$	53.04	83.63	**84.29**	86.91	**90.66**	90.40	63.85	91.57	88.16	56.80	↑0.28 78.37±1.55
+ $\text{GRADIEND}_{\text{Female}}$	53.54	83.58	84.17	86.66	90.47	90.38	64.20	91.33	88.13	**57.22**	↑0.33 78.42±1.59
+ $\text{GRADIEND}_{\text{Male}}$	52.24	83.51	84.21	86.94	90.65	90.34	63.96	91.56	88.19	56.77	↑0.18 78.28±1.58
+ CDA	54.73	**83.90**	84.14	**90.48**	90.56	90.24	**65.76**	91.22	86.77	56.30	↑**0.80** **78.90**±1.55
+ DROPOUT	46.09	82.64	83.36	87.85	90.51	89.77	61.47	91.71	*84.84*	55.00	↓1.40 76.69±1.44
+ INLP	53.97	83.66	84.12	87.09	90.57	90.23	61.48	**92.10**	88.32	55.35	↑0.02 78.11±1.55
+ RLACE	55.47	83.40	83.90	86.02	90.22	90.20	60.70	91.16	88.69	55.83	↓0.10 77.99±1.59
+ LEACE	55.28	83.39	83.94	86.20	90.16	90.24	60.96	91.46	88.66	55.36	↓0.10 78.00±1.58
+ SENTDEBIAS	54.50	83.57	83.92	87.01	90.25	90.22	61.77	91.57	88.46	51.60	↓0.41 77.68±1.02
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	53.91	83.52	83.84	87.25	90.58	90.39	64.92	91.57	87.88	56.35	↑0.41 78.50±1.42
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIAS	53.22	83.69	84.22	86.96	90.64	**90.40**	64.21	91.53	88.16	56.80	↑0.34 78.43±1.55
+ CDA + INLP	54.34	83.69	84.11	89.73	90.64	90.34	64.68	91.75	86.55	51.74	↑0.09 78.19±1.42
+ DROPOUT + SENTDEBIAS	46.09	82.83	83.58	87.95	90.54	89.80	61.95	91.71	*84.80*	55.00	↓1.31 76.78±1.44
+ CDA + SENTDEBIAS	54.47	83.80	84.05	90.45	90.55	90.21	65.76	91.56	86.74	56.30	↑0.79 78.88±1.55
+ DROPOUT + INLP	46.04	82.97	83.62	87.67	90.36	*89.55*	62.09	91.83	*83.92*	54.07	↓1.56 76.53±1.40
$\text{BERT}_{\text{large}}$	62.19	86.19	86.38	88.62	**92.22**	90.50	66.59	**93.31**	88.52	51.59	79.98±1.31
+ $\text{GRADIEND}_{\text{Female/Male}}$	60.14	85.58	86.08	89.93	92.16	**90.58**	66.10	92.84	89.20	55.36	↑0.26 80.24±1.14
+ $\text{GRADIEND}_{\text{Female}}$	61.53	85.85	86.07	87.76	91.98	90.23	66.51	93.10	89.23	56.27	↑0.31 80.29±1.55
+ $\text{GRADIEND}_{\text{Male}}$	62.25	85.68	86.20	88.08	92.06	90.53	65.51	92.76	**89.50**	56.27	↑0.34 80.32±1.55
+ CDA	61.38	85.56	85.96	89.98	92.04	90.56	59.44	93.00	88.56	46.90	↓1.36 78.63±1.41
+ DROPOUT	54.54	85.95	86.11	**90.26**	91.97	90.09	65.85	93.08	88.24	54.86	↓0.55 79.43±1.46
+ INLP	60.00	85.78	86.27	89.56	92.11	90.29	67.58	92.72	89.46	54.74	↑0.30 80.28±1.39
+ RLACE	58.84	**86.29**	86.38	89.08	92.21	90.39	65.99	92.98	89.23	52.98	↓0.20 79.78±1.38
+ LEACE	62.70	85.85	86.17	88.94	91.89	90.34	**68.07**	92.83	89.03	52.50	↑0.28 80.26±1.24
+ SENTDEBIAS	62.65	86.06	**86.47**	89.85	92.08	90.47	67.43	93.26	89.27	55.31	↑**0.75** **80.73**±1.49
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	61.18	85.66	86.21	89.62	91.89	90.47	65.85	92.96	89.45	54.34	↑0.21 80.19±1.25
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIAS	59.87	85.60	86.13	89.78	92.02	90.50	66.35	92.95	89.22	53.47	↑0.02 80.00±1.05
+ CDA + INLP	61.26	85.48	85.98	89.87	91.90	90.50	59.76	92.84	88.49	44.70	↓1.64 78.34±1.10
+ DROPOUT + SENTDEBIAS	*3.14*	85.82	85.75	88.98	91.96	90.33	64.50	93.08	*85.65*	**57.65**	↓*6.53* *73.45±1.39*
+ CDA + SENTDEBIAS	**62.90**	85.50	86.05	90.03	91.80	90.52	62.80	92.78	88.60	46.45	↓0.91 79.07±1.38
+ DROPOUT + INLP	*37.35*	85.58	86.19	90.21	92.19	90.38	63.96	91.87	88.39	46.35	↓*3.69* *76.29±1.16*
DistilBERT	43.90	80.57	81.24	85.79	87.00	88.99	55.13	90.55	81.68	**56.27**	74.47±1.59
+ $\text{GRADIEND}_{\text{Female/Male}}$	43.80	80.60	81.23	85.43	87.07	89.01	55.02	90.70	81.82	**56.27**	↓0.02 74.45±1.59
+ $\text{GRADIEND}_{\text{Female}}$	43.36	80.58	81.22	85.76	87.26	88.99	54.56	90.47	81.58	**56.27**	↓0.12 74.35±1.61
+ $\text{GRADIEND}_{\text{Male}}$	43.91	**80.80**	81.26	85.80	87.00	89.02	55.73	90.61	82.00	54.96	↓0.01 74.45±1.54
+ CDA	43.73	80.67	81.42	86.84	87.30	88.95	58.05	90.35	82.89	52.64	↑0.18 74.64±1.46
+ DROPOUT	43.16	80.35	81.14	87.91	87.41	88.85	60.61	90.37	82.99	54.50	↑0.70 75.17±1.50
+ INLP	43.63	80.63	81.10	85.04	87.16	**89.07**	55.93	**90.82**	81.59	**56.27**	↑0.02 74.49±1.59
+ RLACE	44.03	80.57	81.19	85.68	86.96	89.03	55.26	90.78	81.71	**56.27**	↑0.04 74.51±1.59
+ LEACE	42.92	80.66	81.18	85.64	87.08	89.02	54.44	90.52	81.65	55.82	↓0.24 74.22±1.54
+ SENTDEBIAS	44.14	80.73	81.17	85.66	87.02	89.04	55.51	90.59	81.72	**56.27**	↑0.08 74.54±1.59
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	43.31	80.71	81.13	85.06	87.19	88.99	55.95	90.55	81.69	**56.27**	↓0.03 74.44±1.59
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIAS	43.55	80.57	81.22	84.95	86.90	89.03	54.81	90.19	81.74	**56.27**	↓0.21 74.26±1.60
+ CDA + INLP	**44.22**	80.61	81.43	87.44	87.15	88.91	58.90	90.55	82.94	51.29	↑0.25 74.71±1.33
+ DROPOUT + SENTDEBIAS	43.49	80.37	81.08	**88.06**	87.42	88.80	60.12	90.37	**83.04**	55.41	↑0.80 75.27±1.51
+ CDA + SENTDEBIAS	43.80	80.60	**81.43**	86.84	87.40	88.98	57.82	90.34	82.91	54.05	↑0.33 74.79±1.43
+ DROPOUT + INLP	42.34	80.32	80.97	87.23	**87.83**	88.81	**63.31**	90.06	82.68	55.91	↑**0.96** **75.42**±1.48
RoBERTa	62.67	90.08	89.96	91.00	94.12	90.95	68.17	94.86	91.04	52.03	81.65±1.44
+ $\text{GRADIEND}_{\text{Female/Male}}$	60.27	89.86	89.80	89.46	**94.46**	91.04	*75.86*	95.57	91.82	53.89	↑0.82 82.47±1.53
+ $\text{GRADIEND}_{\text{Female}}$	61.45	89.95	89.88	89.99	94.18	91.00	72.85	*80.30*	91.00	54.83	↓1.04 80.61±1.55
+ $\text{GRADIEND}_{\text{Male}}$	60.72	89.61	89.64	90.77	93.77	*81.81*	67.13	95.77	91.37	51.52	↓1.37 80.28±1.50
+ CDA	62.95	**90.18**	89.82	**91.43**	94.23	91.00	*76.80*	**95.94**	91.82	51.15	↑1.16 82.81±1.41
+ DROPOUT	*24.12*	*53.46*	*53.39*	89.79	94.45	90.53	61.06	*50.93*	*88.73*	54.36	↓*14.16* *67.49±1.47*
+ INLP	62.65	90.00	**90.02**	87.88	94.41	91.07	***80.04***	95.62	91.47	**56.27**	↑1.62 83.27±1.51
+ RLACE	61.23	*71.78*	*71.61*	90.01	94.34	91.19	*75.81*	95.72	91.44	53.87	↓1.06 80.59±1.53
+ LEACE	62.87	89.89	89.55	89.44	94.31	***91.58***	72.47	95.06	91.54	47.74	↓0.01 81.64±1.23
+ SENTDEBIAS	*22.58*	89.85	89.64	91.27	94.29	91.17	68.65	95.80	91.40	49.71	↓*4.47* *77.18±1.23*
+ $\text{GRADIEND}_{\text{Female/Male}}$ + INLP	**64.94**	*71.83*	*71.61*	89.67	*79.71*	91.43	*76.61*	95.42	90.91	**56.27**	↓2.02 79.63±1.53
+ $\text{GRADIEND}_{\text{Female/Male}}$ + SENTDEBIAS	61.26	*41.48*	*41.40*	89.69	*79.66*	91.34	*76.23*	95.54	**91.98**	50.19	↓*6.39* *75.26±1.44*
+ CDA + INLP	61.27	89.77	89.88	90.43	94.46	91.08	*79.72*	95.53	91.83	**56.27**	↑**1.73** **83.38**±1.53
+ DROPOUT + SENTDEBIAS	*36.23*	89.69	89.77	*85.84*	94.11	90.92	*60.08*	95.32	*88.33*	51.46	↓*4.76* *76.89±1.32*
+ CDA + SENTDEBIAS	62.66	89.97	89.82	89.42	94.24	91.44	*79.24*	95.77	91.86	49.23	↑0.99 82.64±1.32
+ DROPOUT + INLP	*30.25*	89.84	89.58	87.94	94.37	*81.78*	*55.99*	*80.69*	*87.65*	55.79	↓*7.85* *73.80±1.45*

Table 19: **Gender:** GLUE bootstrapped validation set scores with sub-results for decoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. GPT-2 results were computed after fine-tuning and LLaMA-based results were computed with zero-shot evaluation.

Model	CoLA	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI	Average ↑
GPT-2	20.51	81.10	82.02	83.75	87.54	88.59	59.87	91.45	80.30	51.96	71.73 ± 1.08
+ GRADIEND$_{Female/Male}$	14.21	81.07	81.91	83.87	87.34	88.58	62.68	91.72	80.41	49.79	↓0.61 71.12 ± 1.08
+ GRADIEND$_{Female}$	11.43	81.03	**82.06**	83.02	87.64	88.56	**63.60**	91.64	80.81	53.46	↓0.42 71.30 ± 1.12
+ GRADIEND$_{Male}$	16.22	81.08	81.89	84.01	87.29	88.49	62.66	91.06	80.75	49.80	↓0.42 71.31 ± 1.09
+ CDA	***32.79***	80.82	81.90	84.54	87.70	88.60	61.80	90.72	80.98	50.35	↑**1.48** **73.20** ± 1.25
+ DROPOUT	20.09	80.57	81.74	83.68	87.03	88.09	62.68	91.03	81.07	50.48	↓0.02 71.70 ± 1.15
+ INLP	20.10	81.06	81.99	83.69	87.67	88.54	61.20	91.72	80.28	51.02	↑0.02 71.75 ± 1.13
+ RLACE	23.21	81.06	82.05	83.77	87.51	**88.69**	62.51	91.38	80.19	52.02	↑0.59 72.31 ± 1.08
+ LEACE	20.05	**81.14**	81.99	83.39	87.54	88.57	61.19	91.34	80.53	50.09	↓0.14 71.58 ± 1.07
+ SENTDEBIAS	18.97	80.98	81.97	83.53	87.52	88.58	61.70	91.37	81.26	48.75	↓0.26 71.46 ± 1.11
+ GRADIEND$_{Female/Male}$ + INLP	13.85	81.09	81.88	84.34	87.33	88.56	63.14	91.83	80.47	49.79	↓0.53 71.20 ± 1.07
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	*9.97*	80.97	81.83	83.87	87.28	88.60	63.27	**91.84**	81.63	**55.90**	↓0.20 71.53 ± 1.12
+ CDA + INLP	*32.21*	80.86	81.90	84.51	87.67	88.61	63.00	90.65	81.02	48.94	↑1.38 73.11 ± 1.24
+ DROPOUT + SENTDEBIAS	21.52	80.62	81.73	83.34	87.05	*87.94*	63.53	90.88	**82.22**	52.82	↑0.55 72.27 ± 1.25
+ CDA + SENTDEBIAS	*30.85*	80.88	81.83	**84.84**	**87.71**	88.58	62.43	91.18	81.37	48.96	↑1.31 73.03 ± 1.27
+ DROPOUT + INLP	19.24	80.57	81.73	83.52	87.05	*88.00*	62.80	91.03	81.06	51.46	↓0.02 71.70 ± 1.13
LLaMA	−8.08	34.96	35.97	69.14	49.93	37.34	54.19	74.03	–	54.86	45.86 ± 1.98
+ GRADIEND$_{Female/Male}$	−2.30	35.12	36.46	72.83	***55.56***	*39.18*	52.40	*62.60*	–	**58.95**	↑1.02 46.88 ± 1.91
+ GRADIEND$_{Female}$	−0.21	**35.80**	**37.21**	*80.37*	49.59	36.86	**57.10**	**78.38**	–	54.88	↑**3.33** **49.19** ± 1.84
+ GRADIEND$_{Male}$	−6.67	34.67	35.39	*48.64*	51.84	***39.94***	54.56	*58.06*	–	57.70	↓3.47 42.39 ± 2.00
+ INLP	***0.00***	*32.26*	*32.65*	***81.10***	49.45	36.83	48.04	*61.66*	–	56.27	↓0.13 45.73 ± 1.78
+ RLACE	−8.76	34.40	35.38	72.49	49.69	37.06	53.42	74.60	–	54.86	↑0.17 46.03 ± 1.95
+ LEACE	−9.46	34.27	35.06	72.07	49.65	37.07	53.75	75.39	–	56.27	↑0.32 46.17 ± 1.97
+ SENTDEBIAS	−7.12	35.10	35.93	76.60	49.57	36.95	55.26	74.40	–	56.27	↑1.32 47.18 ± 1.92
+ GRADIEND$_{Female/Male}$ + INLP	−2.30	35.12	36.46	72.83	***55.56***	*39.18*	52.40	*62.60*	–	**58.95**	↑1.02 46.88 ± 1.91
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	−2.17	34.83	36.11	*79.51*	51.20	37.28	53.05	*62.25*	–	57.59	↑0.92 46.77 ± 1.92
LLaMA-Instruct	16.85	48.12	47.97	4.67	57.34	63.30	64.69	73.02	–	65.20	49.14 ± 1.92
+ GRADIEND$_{Female/Male}$	*2.64*	*35.76*	*35.89*	*79.39*	*49.16*	*39.26*	53.16	*61.62*	–	57.90	↓1.77 47.37 ± 1.81
+ GRADIEND$_{Female}$	**17.72**	*45.86*	*45.91*	1.39	*51.67*	*54.87*	67.15	*66.39*	–	63.86	↓3.02 46.12 ± 1.83
+ GRADIEND$_{Male}$	14.80	46.62	45.98	*79.76*	***70.76***	***68.93***	**70.99**	**77.16**	–	55.05	↑*11.33* ***60.47*** ± *1.86*
+ INLP	16.09	*44.35*	*44.47*	0.68	*54.17*	63.09	66.00	73.54	–	**67.53**	↓0.95 48.19 ± 1.85
+ RLACE	17.03	48.19	**48.20**	4.89	57.36	63.39	64.22	72.71	–	66.15	↑0.10 49.24 ± 1.93
+ LEACE	16.74	**48.51**	48.06	4.22	57.17	63.02	64.97	72.48	–	66.15	↓0.01 49.13 ± 1.91
+ SENTDEBIAS	16.26	48.46	48.20	4.22	56.67	63.11	64.97	72.94	–	66.15	↓0.06 49.08 ± 1.91
+ GRADIEND$_{Female/Male}$ + INLP	*−0.94*	*35.86*	*36.18*	***81.01***	*47.15*	*36.82*	54.52	*58.99*	–	60.70	↓2.36 46.78 ± 1.85
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	5.83	*35.93*	*35.93*	*80.17*	*49.07*	*38.51*	*52.69*	*62.88*	–	60.79	↓0.91 48.23 ± 1.85

Table 20: **Race:** GLUE bootstrapped validation set scores with sub-results for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. LLaMA-based results were computed with zero-shot evaluation while all other scores are derived after fine-tuning.

Model	CoLA	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI	Average ↑
BERT$_{\text{base}}$	55.60	83.40	83.97	86.32	90.19	90.21	60.95	91.34	88.71	55.83	78.09±1.59
+ GRADIEND$_{\text{Asian/Black}}$	51.67	**83.71**	84.07	87.83	90.24	90.30	**66.87**	91.37	88.14	56.76	↑0.47 78.56±1.60
+ GRADIEND$_{\text{Asian/White}}$	53.08	83.62	84.02	89.03	90.32	90.27	65.52	91.21	88.21	**57.22**	**↑0.65** **78.74**±1.61
+ GRADIEND$_{\text{Black/White}}$	51.51	83.62	84.04	87.78	90.26	90.27	65.87	91.68	88.29	55.87	↑0.28 78.37±1.56
+ CDA	49.85	83.55	84.08	**89.67**	**90.65**	**90.37**	63.43	91.23	87.38	54.49	↓0.22 77.88±1.48
+ DROPOUT	46.09	82.64	83.36	87.85	90.51	89.77	61.47	91.71	*84.84*	55.00	↓1.40 76.69±1.44
+ INLP	54.35	83.67	**84.12**	86.66	90.64	90.30	60.43	**92.21**	88.27	53.97	↓0.24 77.86±1.23
+ SENTDEBIAS	**56.00**	83.50	83.95	86.20	90.21	90.26	60.69	92.07	**88.71**	55.36	↑0.04 78.14±1.58
BERT$_{\text{large}}$	62.19	**86.19**	**86.38**	88.62	92.22	90.50	66.59	93.31	88.52	51.59	79.98±1.31
+ GRADIEND$_{\text{Asian/Black}}$	60.70	85.50	86.09	88.57	92.21	90.60	66.80	93.10	89.42	**56.27**	↑0.40 80.38±1.55
+ GRADIEND$_{\text{Asian/White}}$	**63.00**	85.56	86.06	**90.39**	92.19	90.62	67.28	92.92	89.47	**56.27**	**↑0.90** **80.88**±1.53
+ GRADIEND$_{\text{Black/White}}$	61.61	85.64	86.21	88.96	92.04	90.46	66.06	**93.53**	89.59	**56.27**	↑0.51 80.49±1.54
+ CDA	58.44	85.64	86.13	88.38	92.10	**90.72**	57.72	92.27	87.41	48.81	↓2.01 77.97±0.97
+ DROPOUT	54.54	85.95	86.11	90.26	91.97	90.09	65.85	93.08	88.24	54.86	↓0.55 79.43±1.46
+ INLP	59.69	85.70	86.09	89.17	**92.31**	90.55	**67.70**	93.27	89.55	52.00	↑0.03 80.02±1.29
+ SENTDEBIAS	59.07	85.66	86.10	89.19	92.09	90.47	67.06	93.14	**89.60**	54.39	↑0.12 80.10±1.53
DistilBERT	43.90	80.57	81.24	85.79	87.00	88.99	55.13	**90.55**	81.68	56.27	74.47±1.59
+ GRADIEND$_{\text{Asian/Black}}$	44.99	80.38	81.34	84.95	87.01	**89.03**	54.66	90.14	81.75	56.27	↓0.06 74.41±1.60
+ GRADIEND$_{\text{Asian/White}}$	44.60	80.43	81.45	84.46	86.99	88.90	55.03	90.02	81.88	56.27	↓0.12 74.34±1.60
+ GRADIEND$_{\text{Black/White}}$	45.36	80.35	81.30	85.41	86.98	89.00	53.58	90.14	81.75	56.41	↓0.08 74.38±1.47
+ CDA	41.08	**80.59**	**81.47**	87.58	87.33	88.96	59.55	90.28	**83.87**	52.21	↑0.19 74.65±1.45
+ DROPOUT	43.16	80.35	81.14	**87.91**	**87.41**	88.85	**60.61**	90.37	82.99	54.50	**↑0.70** **75.17**±1.50
+ INLP	43.54	80.35	81.22	86.14	87.33	88.99	56.25	89.94	82.02	**56.75**	↑0.17 74.64±1.57
+ SENTDEBIAS	**45.49**	80.36	81.34	85.39	87.25	89.01	55.04	90.06	81.77	56.27	↑0.10 74.57±1.60
RoBERTa	62.67	**90.08**	**89.96**	91.00	94.12	90.95	**68.17**	94.86	91.04	52.03	**81.65**±1.44
+ GRADIEND$_{\text{Asian/Black}}$	57.99	*71.27*	*71.35*	*85.31*	*50.55*	*77.15*	60.87	*80.55*	91.17	**55.75**	*↓11.58* *70.07±1.48*
+ GRADIEND$_{\text{Asian/White}}$	*20.38*	89.63	89.74	89.65	*79.57*	91.39	65.77	*80.47*	90.25	51.07	*↓8.51* *73.14±0.86*
+ GRADIEND$_{\text{Black/White}}$	*39.60*	*71.75*	*71.64*	90.36	94.13	91.04	61.99	95.54	91.63	52.06	*↓5.20* *76.45±1.16*
+ CDA	62.15	90.06	89.88	**91.21**	94.11	***91.52***	67.69	95.28	**91.75**	50.58	↓0.07 81.58±1.29
+ DROPOUT	*24.12*	*53.46*	*53.39*	89.79	**94.45**	90.53	61.06	*50.93*	*88.73*	54.36	*↓14.16* *67.49±1.47*
+ INLP	**63.42**	89.99	89.82	89.86	94.10	91.33	65.04	95.72	91.54	52.97	↓0.11 81.54±1.48
+ SENTDEBIAS	*34.80*	89.95	89.55	89.10	94.13	91.45	66.81	**96.15**	91.66	52.44	*↓3.17* *78.48±1.30*
GPT-2	20.51	**81.10**	**82.02**	83.75	87.54	**88.59**	59.87	91.45	80.30	**51.96**	71.73±1.08
+ GRADIEND$_{\text{Asian/Black}}$	16.57	80.98	81.83	84.24	87.64	88.47	60.89	91.62	82.05	47.39	↓0.58 71.14±1.01
+ GRADIEND$_{\text{Asian/White}}$	13.03	80.94	81.82	83.67	87.55	88.54	62.24	**91.90**	80.67	46.88	↓1.08 70.65±0.98
+ GRADIEND$_{\text{Black/White}}$	18.45	80.97	81.90	83.68	87.61	88.51	62.99	91.87	81.59	47.39	↓0.22 71.50±1.07
+ CDA	***30.53***	80.64	81.90	**85.00**	87.60	88.54	**64.52**	90.80	**82.65**	50.34	**↑1.74** **73.47**±1.12
+ DROPOUT	20.09	80.57	81.74	83.68	87.03	88.09	62.68	91.03	81.07	50.48	↓0.02 71.70±1.15
+ INLP	18.83	81.05	82.00	83.67	**87.72**	88.49	62.89	91.67	81.34	46.89	↓0.28 71.45±1.08
+ SENTDEBIAS	17.97	80.96	81.97	84.22	87.59	88.53	62.72	91.86	81.75	47.85	↓0.18 71.55±1.09
LLaMA	−8.08	34.96	35.97	69.14	49.93	37.34	54.19	74.03	–	54.86	45.86±1.98
+ GRADIEND$_{\text{Asian/Black}}$	−4.95	35.67	35.97	**75.32**	***58.03***	***59.46***	55.88	*58.35*	–	57.57	**↑3.58** **49.44**±1.97
+ GRADIEND$_{\text{Asian/White}}$	**1.02**	*36.92*	37.11	*49.54*	*53.52*	*54.11*	**60.52**	*57.42*	–	**63.29**	↑1.20 47.06±1.97
+ GRADIEND$_{\text{Black/White}}$	−7.39	34.50	35.06	*59.49*	50.71	*44.97*	52.69	*65.10*	–	54.86	↓1.46 44.40±2.01
+ INLP	−7.99	***38.14***	***39.18***	72.46	49.94	37.38	58.20	**76.00**	–	57.65	↑1.93 47.79±1.87
+ SENTDEBIAS	−8.55	35.88	36.77	71.42	49.57	37.08	57.40	72.90	–	54.86	↑0.52 46.38±1.93
LLaMA-Instruct	**16.85**	48.12	47.97	4.67	57.34	63.30	64.69	73.02	–	65.20	49.14±1.92
+ GRADIEND$_{\text{Asian/Black}}$	*0.00*	*32.75*	*32.93*	*0.00*	*50.55*	63.17	*47.34*	*49.11*	–	56.27	*↓11.73* *37.41±1.75*
+ GRADIEND$_{\text{Asian/White}}$	*0.00*	*33.29*	*33.38*	*0.00*	*50.55*	63.17	*52.32*	*51.00*	–	43.73	*↓12.38* *36.76±1.75*
+ GRADIEND$_{\text{Black/White}}$	9.81	*45.43*	*45.32*	***22.41***	***65.11***	*41.61*	**69.25**	69.62	–	**66.14**	↓0.48 48.66±2.02
+ INLP	16.34	**49.30**	**49.46**	7.63	59.36	***65.00***	66.77	**74.21**	–	60.58	**↑0.77** **49.91**±1.98
+ SENTDEBIAS	15.80	48.56	48.60	6.47	57.25	64.09	66.53	71.35	–	61.53	↓0.19 48.95±1.96

Table 21: **Religion:** GLUE bootstrapped validation set scores with sub-results for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. LLaMA-based results were computed with zero-shot evaluation while all other scores are derived after fine-tuning.

Model	CoLA	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI	Average ↑
BERT$_{base}$	55.60	83.40	83.97	86.32	90.19	90.21	60.95	91.34	**88.71**	55.83	78.09±1.59
+ GradiEnd$_{Christian/Jewish}$	51.16	83.58	84.07	87.50	90.33	90.30	65.02	**91.79**	88.32	56.77	↑0.24 78.33±1.58
+ GradiEnd$_{Christian/Muslim}$	51.14	83.57	83.91	88.04	90.20	90.29	**65.51**	91.72	88.44	56.82	**↑0.34** **78.43**±1.57
+ GradiEnd$_{Jewish/Muslim}$	51.51	83.62	83.99	87.58	90.22	90.25	65.27	91.52	87.98	**57.23**	↑0.28 78.37±1.60
+ CDA	52.47	83.50	83.81	**89.94**	90.35	**90.36**	65.03	91.12	87.41	54.97	↑0.27 78.36±1.49
+ Dropout	46.09	82.64	83.36	87.85	90.51	89.77	61.47	91.71	*84.84*	55.00	↓1.40 76.69±1.44
+ INLP	55.22	**83.72**	**84.16**	86.67	**90.63**	90.27	61.41	91.72	88.34	51.15	↓0.39 77.71±1.23
+ SentDebias	**55.66**	83.36	83.94	86.20	90.23	90.23	63.33	91.50	88.50	51.60	↓0.22 77.88±1.03
BERT$_{large}$	62.19	**86.19**	86.38	88.62	**92.22**	90.50	66.59	93.31	88.52	51.59	79.98±1.31
+ GradiEnd$_{Christian/Jewish}$	61.94	85.72	86.20	89.01	91.99	90.65	**68.89**	**93.42**	**89.76**	56.27	**↑0.90** **80.88**±1.55
+ GradiEnd$_{Christian/Muslim}$	60.73	85.45	86.10	88.21	92.12	90.57	66.90	92.98	89.67	56.27	↑0.38 80.36±1.55
+ GradiEnd$_{Jewish/Muslim}$	62.57	85.70	86.24	89.05	92.11	**90.81**	66.08	93.07	89.69	56.27	↑0.64 80.62±1.55
+ CDA	59.36	85.58	86.06	**90.45**	91.70	90.59	68.03	93.00	88.46	**56.33**	↑0.43 80.42±1.53
+ Dropout	54.54	85.95	86.11	90.26	91.97	90.09	65.85	93.08	88.24	54.86	↓0.55 79.43±1.46
+ INLP	61.26	85.94	**86.47**	89.40	92.17	90.54	66.99	93.25	89.75	49.18	↓0.12 79.86±1.08
+ SentDebias	**63.89**	86.14	86.41	87.91	92.17	90.43	68.18	93.28	89.41	54.53	↑0.70 80.68±1.40
DistilBERT	43.90	80.57	81.24	85.79	87.00	88.99	55.13	90.55	81.68	**56.27**	74.47±1.59
+ GradiEnd$_{Christian/Jewish}$	43.83	80.69	81.25	85.49	87.12	89.02	55.39	**90.58**	81.74	**56.27**	↑0.02 74.49±1.60
+ GradiEnd$_{Christian/Muslim}$	43.45	80.59	81.20	85.39	86.95	88.90	56.33	90.53	81.76	**56.27**	↑0.03 74.50±1.60
+ GradiEnd$_{Jewish/Muslim}$	42.67	80.68	81.18	85.38	87.16	88.99	56.12	90.43	81.63	**56.27**	↓0.07 74.40±1.61
+ CDA	44.10	**80.71**	**81.46**	**88.27**	87.24	88.84	59.57	89.93	**83.89**	51.67	↑0.49 74.96±1.46
+ Dropout	43.16	80.35	81.14	87.91	87.41	88.85	**60.61**	90.37	82.99	54.50	**↑0.70** **75.17**±1.50
+ INLP	45.13	80.30	81.23	85.83	**87.58**	88.96	55.77	90.05	82.39	54.87	↑0.13 74.59±1.57
+ SentDebias	**45.59**	80.42	81.22	85.23	87.06	**89.07**	54.80	90.13	81.88	**56.27**	↑0.07 74.54±1.60
RoBERTa	62.67	90.08	**89.96**	**91.00**	94.12	90.95	68.17	94.86	91.04	52.03	81.65±1.44
+ GradiEnd$_{Christian/Jewish}$	*34.62*	*35.45*	*35.25*	89.22	93.54	***91.52***	67.03	95.64	91.38	54.83	*↓9.08* *72.57±1.50*
+ GradiEnd$_{Christian/Muslim}$	**62.92**	89.77	89.75	90.37	94.26	90.89	*76.49*	95.64	91.80	**55.31**	**↑1.40** **83.05**±1.51
+ GradiEnd$_{Jewish/Muslim}$	61.47	89.37	89.23	87.62	94.26	91.30	***77.99***	95.49	91.67	**55.31**	↑1.06 82.71±1.53
+ CDA	61.57	**90.17**	89.94	90.75	94.15	90.95	72.88	95.62	**91.94**	54.40	↑0.83 82.48±1.51
+ Dropout	*24.12*	*53.46*	*53.39*	89.79	**94.45**	90.53	61.06	*50.93*	*88.73*	54.36	*↓14.16* *67.49±1.47*
+ INLP	60.09	*71.86*	*71.76*	88.93	*79.30*	*91.45*	66.22	95.91	91.71	53.89	*↓3.95* *77.70±1.51*
+ SentDebias	62.46	90.00	89.72	90.60	94.42	*82.03*	70.81	**96.01**	91.78	47.52	↓1.04 80.61±0.86
GPT-2	20.51	**81.10**	**82.02**	83.75	87.54	88.59	59.87	91.45	80.30	**51.96**	71.73±1.08
+ GradiEnd$_{Christian/Jewish}$	19.80	80.93	81.79	83.38	87.47	88.56	61.87	91.67	81.66	49.75	↓0.00 71.73±0.98
+ GradiEnd$_{Christian/Muslim}$	*7.19*	80.83	81.94	83.37	87.66	88.58	60.05	**91.88**	81.10	50.24	↓1.56 70.16±1.04
+ GradiEnd$_{Jewish/Muslim}$	19.97	80.82	81.67	83.41	87.48	88.53	61.39	91.67	81.70	50.58	↑0.05 71.78±1.11
+ CDA	***32.97***	80.82	81.63	**84.70**	**87.69**	88.41	**63.39**	91.39	**82.12**	47.96	**↑1.59** **73.32**±1.23
+ Dropout	20.09	80.57	81.74	83.68	87.03	88.09	62.68	91.03	81.07	50.48	↓0.02 71.70±1.15
+ INLP	18.33	81.01	81.91	83.81	87.67	**88.59**	63.02	91.75	81.16	47.84	↓0.21 71.52±1.06
+ SentDebias	21.31	80.99	81.86	84.13	87.63	88.48	62.88	91.75	81.93	47.39	↑0.16 71.88±1.06
LLaMA	−8.08	34.96	35.97	69.14	49.93	37.34	54.19	74.03	–	54.86	45.86±1.98
+ GradiEnd$_{Christian/Jewish}$	−13.25	33.16	*33.63*	*19.66*	***54.55***	***54.30***	52.38	*50.72*	–	54.82	*↓7.54* *38.32±2.06*
+ GradiEnd$_{Christian/Muslim}$	**−1.61**	33.80	34.86	66.16	49.60	37.52	51.40	*61.99*	–	**56.27**	↓1.40 44.46±2.03
+ GradiEnd$_{Jewish/Muslim}$	−2.47	***37.16***	**37.77**	70.37	52.37	*41.42*	**59.50**	*57.45*	–	56.26	↑0.69 46.54±1.90
+ INLP	−6.76	35.72	36.73	***79.61***	50.05	36.91	56.75	**76.08**	–	**56.27**	**↑2.28** **48.14**±1.84
+ SentDebias	−9.09	34.80	35.83	67.55	50.02	37.46	55.24	75.19	–	54.86	↓0.04 45.82±1.96
LLaMA-Instruct	**16.85**	48.12	47.97	4.67	57.34	63.30	64.69	73.02	–	**65.20**	49.14±1.92
+ GradiEnd$_{Christian/Jewish}$	15.02	48.60	48.51	8.29	56.13	***64.61***	66.05	72.16	–	64.81	↑0.31 49.45±2.00
+ GradiEnd$_{Christian/Muslim}$	9.31	*35.58*	*35.35*	*0.00*	*52.98*	*64.14*	65.47	**74.21**	–	55.88	*↓4.46* *44.68±1.30*
+ GradiEnd$_{Jewish/Muslim}$	4.76	*43.28*	*43.70*	***43.69***	*50.54*	*47.33*	**71.86**	68.86	–	63.29	↑0.09 49.23±1.95
+ INLP	15.71	**48.69**	**49.00**	8.26	**57.96**	63.75	65.69	73.52	–	63.40	**↑0.50** **49.64**±1.99
+ SentDebias	14.98	48.60	48.72	5.55	57.33	*64.21*	65.66	71.91	–	61.99	↓0.35 48.79±1.97

Table 22: **Gender:** SuperGLUE bootstrapped validation set scores with sub-results for encoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model Metrics	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC $F1_{\alpha}$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.	Average ↑
$\text{BERT}_{\text{base}}$	69.16	38.74/58.68	62.72	60.12/13.23	56.09/55.32	61.30	68.67	63.12	51.82±1.67
+ $\text{GradiEnd}_{\text{Female/Male}}$	70.83	42.64/62.23	59.57	60.46/13.77	55.79/55.03	63.98	68.42	63.40	↑0.56 52.38±1.88
+ $\text{GradiEnd}_{\text{Female}}$	70.49	42.68/62.23	59.76	60.71/14.26	56.01/55.24	64.66	69.43	63.40	↑0.82 52.65±1.88
+ $\text{GradiEnd}_{\text{Male}}$	70.41	42.68/62.23	58.84	58.88/13.80	55.98/55.22	64.32	69.15	63.40	↑0.44 52.27±1.88
+ CDA	70.09	**47.75/69.49**	57.60	60.43/15.30	55.64/54.93	**65.41**	67.29	63.40	↑1.33 53.16±1.80
+ Dropout	68.53	47.39/68.99	55.56	59.20/12.94	55.00/54.23	61.74	65.15	62.77	↓0.34 51.48±1.72
+ INLP	69.25	34.57/56.91	58.67	60.62/14.50	**56.27/55.49**	61.26	66.44	63.36	↓0.80 51.02±1.55
+ RLACE	69.03	27.47/52.19	62.70	59.73/13.12	56.07/55.30	61.31	68.73	**63.44**	↓0.88 50.95±1.54
+ LEACE	69.06	27.47/52.19	**63.06**	60.21/13.48	55.99/55.21	61.18	68.63	63.11	↓0.86 50.96±1.55
+ SentDebias	69.06	27.47/52.19	63.06	60.10/13.30	55.97/55.21	61.66	68.63	63.12	↓0.83 50.99±1.55
+ $\text{GradiEnd}_{\text{Female/Male}}$ + INLP	**70.90**	44.14/64.13	61.73	60.34/14.37	55.94/55.16	64.93	**69.87**	63.40	**↑1.51** **53.33**±1.82
+ $\text{GradiEnd}_{\text{Female/Male}}$ + SentDebias	70.77	42.64/62.23	59.55	60.48/13.76	55.80/55.04	64.36	68.38	63.40	↑0.56 52.39±1.88
+ CDA + INLP	69.86	46.50/67.73	58.09	59.13/13.95	55.67/54.96	65.12	67.22	63.40	↑0.82 52.64±1.78
+ Dropout + SentDebias	68.72	46.98/68.41	54.89	59.17/13.08	54.98/54.22	61.87	65.41	62.46	↓0.41 51.42±1.71
+ CDA + SentDebias	70.12	**47.75/69.49**	58.23	**60.53/15.31**	55.67/54.96	**65.41**	67.28	63.40	↑1.41 53.24±1.79
+ Dropout + INLP	68.45	40.90/61.77	55.21	59.12/13.09	55.48/54.74	62.23	64.94	62.73	↓1.10 50.73±1.70
$\text{BERT}_{\text{large}}$	70.32	42.86/62.97	61.46	61.49/15.19	61.70/61.04	67.68	**70.82**	62.09	53.74±1.62
+ $\text{GradiEnd}_{\text{Female/Male}}$	***73.03***	46.69/67.15	65.34	59.11/15.47	61.47/60.78	65.49	69.53	63.40	↑0.46 54.20±1.88
+ $\text{GradiEnd}_{\text{Female}}$	71.86	44.84/64.70	60.46	61.94/15.39	61.45/60.75	66.13	69.55	62.79	↑0.10 53.84±1.86
+ $\text{GradiEnd}_{\text{Male}}$	71.61	44.94/64.80	58.32	61.62/15.25	61.67/61.01	66.20	69.62	63.40	↓0.10 53.64±1.87
+ CDA	72.41	47.59/68.35	62.94	61.90/16.43	61.65/61.02	61.95	69.42	63.40	↑0.28 54.02±1.81
+ Dropout	71.12	45.18/65.27	53.62	62.24/16.01	**62.09/61.37**	64.72	67.99	63.40	↓0.52 53.22±1.68
+ INLP	72.67	38.72/61.76	62.19	*39.19/8.60*	61.59/60.92	66.09	70.01	**63.45**	↓1.60 52.14±1.58
+ RLACE	*67.69*	43.46/63.61	61.47	*39.28/9.05*	61.78/61.11	**68.40**	70.53	60.42	↓1.69 52.05±1.62
+ LEACE	70.54	43.27/62.99	61.78	60.75/15.65	61.51/60.83	67.45	69.75	63.09	↑0.09 53.84±1.67
+ SentDebias	70.05	42.57/62.42	61.79	61.25/15.52	61.63/60.97	67.78	70.68	63.08	↑0.03 53.77±1.66
+ $\text{GradiEnd}_{\text{Female/Male}}$ + INLP	72.01	46.24/66.53	**66.38**	60.65/15.10	61.48/60.81	66.13	69.22	63.06	↑0.56 54.30±1.93
+ $\text{GradiEnd}_{\text{Female/Male}}$ + SentDebias	71.78	46.69/67.15	65.93	61.31/15.46	61.88/61.21	66.21	69.67	63.40	**↑0.64** **54.38**±1.87
+ CDA + INLP	72.25	47.18/67.78	64.91	61.05/15.34	61.73/61.06	64.33	69.15	62.76	↑0.12 53.87±1.81
+ Dropout + SentDebias	70.60	47.18/67.77	58.25	61.90/14.41	61.04/60.40	66.24	66.38	63.06	↓0.39 53.36±1.74
+ CDA + SentDebias	72.50	**47.95/68.92**	63.62	62.62/15.23	61.81/61.16	61.66	69.43	63.38	↑0.26 54.00±1.80
+ Dropout + INLP	70.27	47.22/67.77	58.69	**62.34/16.61**	61.17/60.49	65.32	*64.34*	63.40	↓0.41 53.33±1.74
DistilBERT	69.75	45.62/66.55	53.39	57.58/12.21	49.09/48.27	55.18	62.10	**63.40**	49.69±1.65
+ $\text{GradiEnd}_{\text{Female/Male}}$	69.81	46.63/67.72	55.06	57.77/13.13	49.02/48.21	55.11	62.01	**63.40**	↑0.21 49.90±1.67
+ $\text{GradiEnd}_{\text{Female}}$	69.83	47.50/68.96	55.76	58.14/13.46	49.15/48.37	55.61	61.55	**63.40**	↑0.63 50.32±1.63
+ $\text{GradiEnd}_{\text{Male}}$	69.50	44.02/64.77	55.41	58.51/12.90	49.06/48.24	56.04	61.37	**63.40**	↓0.05 49.64±1.69
+ CDA	69.19	**48.31/69.52**	59.40	59.85/13.31	49.16/48.33	58.10	63.78	**63.40**	↑1.06 50.75±1.76
+ Dropout	69.21	46.13/66.49	54.78	59.40/13.05	**49.77/48.97**	60.45	62.62	**63.40**	↑0.58 50.27±1.75
+ INLP	69.85	40.07/63.63	**60.46**	58.76/12.21	48.94/48.10	55.58	61.86	**63.40**	↑0.21 49.90±1.56
+ RLACE	**69.99**	45.08/65.95	53.74	57.04/11.69	49.08/48.28	56.31	61.85	**63.40**	↑0.03 49.72±1.66
+ LEACE	69.54	45.92/67.15	55.39	57.48/11.20	49.20/48.39	54.13	62.52	**63.40**	↑0.09 49.78±1.63
+ SentDebias	69.97	45.08/65.95	54.42	57.20/11.76	49.12/48.32	55.55	62.10	**63.40**	↑0.06 49.75±1.64
+ $\text{GradiEnd}_{\text{Female/Male}}$ + INLP	69.90	39.64/63.00	58.75	59.38/13.24	48.94/48.12	56.17	61.82	**63.40**	↑0.16 49.85±1.57
+ $\text{GradiEnd}_{\text{Female/Male}}$ + SentDebias	69.92	46.23/67.12	55.45	57.96/12.47	49.09/48.28	54.49	61.65	**63.40**	↑0.25 49.94±1.66
+ CDA + INLP	69.29	46.15/67.16	55.12	**59.45/13.86**	49.28/48.45	60.00	**64.78**	**63.40**	↑1.40 51.09±1.72
+ Dropout + SentDebias	69.28	46.13/66.49	55.54	59.34/13.70	49.73/48.93	60.49	63.08	**63.40**	↑0.76 50.45±1.76
+ CDA + SentDebias	69.08	**48.31/69.52**	57.82	56.55/11.78	49.10/48.27	58.23	63.80	**63.40**	↑0.83 50.52±1.77
+ Dropout + INLP	69.17	47.31/68.30	59.12	58.97/12.72	49.69/48.87	**63.22**	62.91	**63.40**	**↑1.50** **51.19**±1.72
RoBERTa	82.01	46.41/66.62	56.70	42.49/12.60	72.14/71.46	75.36	56.83	53.49	53.31±1.48
+ $\text{GradiEnd}_{\text{Female/Male}}$	*75.70*	45.99/66.03	58.40	*64.23/21.85*	71.98/71.30	76.18	*66.76*	53.49	↑2.03 55.34±1.47
+ $\text{GradiEnd}_{\text{Female}}$	81.64	44.18/64.24	60.06	*22.50/7.40*	72.11/71.46	69.97	*62.88*	62.09	↓0.49 52.82±1.65
+ $\text{GradiEnd}_{\text{Male}}$	82.61	43.43/62.44	54.16	*67.91/23.34*	71.99/71.34	*62.29*	*63.62*	52.23	↑0.48 53.79±1.47
+ CDA	82.80	47.57/68.36	63.25	*45.02/17.17*	72.20/71.57	78.13	*69.75*	53.49	↑2.89 56.20±1.44
+ Dropout	*73.53*	45.03/64.77	50.32	*45.78/17.10*	**72.28/71.60**	*60.86*	61.03	57.62	↓2.25 51.05±1.62
+ INLP	82.32	47.15/67.79	57.98	*45.58/16.43*	71.98/71.34	75.60	61.52	**63.40**	↑1.75 55.06±1.66
+ RLACE	**82.93**	45.58/65.45	56.92	*44.94/16.30*	71.97/71.30	73.22	61.82	53.49	↑0.67 53.98±1.50
+ LEACE	*75.70*	38.84/61.85	56.91	*68.43/24.30*	72.10/71.45	68.66	57.75	59.22	↑0.16 53.47±1.35
+ SentDebias	82.39	45.97/66.00	55.51	*65.65/21.52*	71.94/71.27	68.87	61.11	53.49	↑1.22 54.53±1.51
+ $\text{GradiEnd}_{\text{Female/Male}}$ + INLP	*75.56*	47.62/68.44	57.94	*63.42/19.59*	72.14/71.44	69.91	*64.79*	**63.40**	↑1.86 55.17±1.63
+ $\text{GradiEnd}_{\text{Female/Male}}$ + SentDebias	82.37	46.41/66.62	55.82	*65.14/22.32*	71.98/71.31	*66.49*	*65.05*	52.85	↑1.41 54.72±1.49
+ CDA + INLP	81.81	**49.16/70.74**	61.69	***67.80/25.09***	72.10/71.49	**79.04**	***70.75***	**63.40**	*↑5.58* ***58.89***±*1.63*
+ Dropout + SentDebias	*69.98*	45.03/64.77	47.05	*62.92/19.86*	71.64/70.91	*61.04*	56.94	57.92	↓2.29 51.01±1.59
+ CDA + SentDebias	82.68	48.44/69.59	**64.59**	*45.36/16.38*	71.99/71.35	77.56	*69.56*	53.49	*↑2.97* *56.28±1.42*
+ Dropout + INLP	*71.96*	46.73/67.16	51.30	*0.00/0.32*	71.88/71.17	*59.29*	59.13	**63.40**	*↓5.10* *48.21±1.78*

Table 23: **Gender:** SuperGLUE bootstrapped validation set scores with sub-results for decoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. GPT-2 results were computed after fine-tuning and LLaMA-based results were computed with zero-shot evaluation.

Model Metrics	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC $F1_\alpha$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.	Average ↑
GPT-2	65.56	36.86/51.74	49.35	58.79/13.69	31.64/30.93	60.14	62.51	54.47	45.49±1.28
+ GRADIEND$_{Female/Male}$	65.22	36.10/51.07	50.88	59.92/14.14	31.77/31.04	62.41	63.82	**59.53**	↑0.86 46.34±1.27
+ GRADIEND$_{Female}$	64.66	37.56/51.70	50.74	59.01/13.85	**31.84/31.14**	63.28	63.50	55.42	↑0.49 45.97±1.20
+ GRADIEND$_{Male}$	65.28	38.49/55.21	49.26	60.09/13.86	31.65/30.92	61.42	63.92	56.34	↑0.60 46.09±1.25
+ CDA	**66.70**	**42.95/57.57**	49.35	59.81/14.34	31.61/30.92	61.84	64.43	57.69	↑1.28 46.76±1.38
+ DROPOUT	66.02	34.95/52.30	49.04	58.82/13.86	31.55/30.86	62.26	62.66	58.72	↑0.46 45.94±1.45
+ INLP	65.77	36.77/51.74	51.34	58.80/13.41	31.49/30.78	60.86	62.27	54.78	↑0.29 45.78±1.20
+ RLACE	65.84	37.07/52.37	49.26	59.24/13.76	31.58/30.85	62.33	62.77	55.75	↑0.44 45.92±1.20
+ LEACE	65.59	35.53/50.52	51.65	58.56/12.88	31.61/30.88	60.76	61.79	57.10	↑0.35 45.83±1.24
+ SENTDEBIAS	65.23	28.85/42.82	51.31	58.70/12.53	31.69/30.96	61.11	61.77	51.23	↓1.15 44.33±1.18
+ GRADIEND$_{Female/Male}$ + INLP	65.36	35.69/50.45	51.93	59.51/13.83	31.75/31.03	63.02	63.50	58.30	↑0.82 46.30±1.27
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	65.18	37.22/52.78	52.92	59.11/13.62	31.66/30.96	62.71	63.69	53.87	↑0.47 45.96±1.24
+ CDA + INLP	66.45	42.14/56.39	52.42	59.78/14.55	31.67/30.98	63.06	63.96	56.12	**↑1.39** **46.87**±1.32
+ DROPOUT + SENTDEBIAS	65.59	37.65/56.99	51.33	59.85/14.39	31.51/30.81	**64.40**	63.28	56.57	↑1.28 46.76±1.32
+ CDA + SENTDEBIAS	66.66	41.89/56.34	49.04	**60.03/14.55**	31.77/31.07	62.55	**64.63**	53.58	↑0.78 46.27±1.34
+ DROPOUT + INLP	65.99	35.97/53.47	**53.71**	59.01/14.11	31.62/30.93	62.89	62.39	58.70	↑1.11 46.59±1.44
LLaMA	72.96	**37.32/51.82**	**86.06**	0.00/0.32	**90.42/89.70**	54.17	50.10	37.58	**54.46**±2.28
+ GRADIEND$_{Female/Male}$	*65.24*	31.80/44.35	76.03	***1.48/0.42***	*88.73/87.92*	52.41	50.12	36.60	↓3.49 50.97±2.20
+ GRADIEND$_{Female}$	**73.58**	33.67/42.88	80.10	0.00/0.32	89.38/88.63	**57.02**	50.12	36.60	↓1.35 53.11±2.28
+ GRADIEND$_{Male}$	*69.02*	26.71/40.85	82.11	0.58/0.53	89.57/88.78	54.56	50.12	**39.50**	↓2.10 52.35±2.09
+ INLP	*65.64*	26.77/37.46	78.10	0.00/0.32	*86.00/85.26*	48.22	50.12	36.60	*↓4.88* *49.57±2.21*
+ RLACE	73.17	36.37/51.75	**86.06**	0.00/0.32	*0.00/0.01*	53.43	50.25	36.60	*↓11.49* *42.97±2.31*
+ LEACE	73.33	36.37/51.75	85.07	0.00/0.32	*0.00/0.01*	53.76	**50.42**	36.60	*↓11.53* *42.93±2.31*
+ SENTDEBIAS	73.28	36.87/46.41	85.08	0.00/0.32	90.06/89.27	55.27	50.27	37.58	↓0.34 54.12±2.37
+ GRADIEND$_{Female/Male}$ + INLP	*62.37*	*13.24/15.88*	*68.84*	0.00/0.32	*84.05/83.34*	47.59	50.12	36.60	*↓8.97* *45.49±2.08*
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	*66.73*	27.66/35.64	76.03	0.94/0.31	*88.65/87.83*	53.18	50.12	36.60	↓4.06 50.40±2.16
LLaMA-Instruct	75.25	31.58/32.13	78.60	27.43/0.52	**85.32/84.68**	67.31	**50.37**	62.20	58.07±2.29
+ GRADIEND$_{Female/Male}$	73.23	18.78/39.43	75.08	*3.00/0.32*	*83.11/82.48*	*53.77*	50.11	58.27	*↓5.07* *53.00±2.05*
+ GRADIEND$_{Female}$	73.41	**44.03/49.36**	**82.89**	**31.00/1.12**	84.84/84.21	68.32	50.04	64.05	**↑2.68** **60.75**±2.35
+ GRADIEND$_{Male}$	***78.54***	30.28/37.55	80.19	*4.57/0.21*	84.03/83.48	**70.91**	50.11	51.11	↓1.71 56.36±2.28
+ INLP	72.27	32.65/33.37	81.02	*11.51/0.56*	84.70/84.08	66.60	50.26	**65.24**	↓0.72 57.35±2.34
+ RLACE	75.21	32.10/32.75	78.60	27.24/0.49	85.08/84.42	67.43	**50.37**	61.54	↓0.05 58.02±2.28
+ LEACE	75.18	30.89/31.57	78.25	26.90/0.52	85.06/84.41	67.19	50.22	62.18	↓0.23 57.84±2.29
+ SENTDEBIAS	74.98	34.06/35.20	79.63	27.15/0.53	85.24/84.58	67.42	50.22	62.86	↑0.49 58.56±2.31
+ GRADIEND$_{Female/Male}$ + INLP	*66.07*	18.78/39.43	78.07	*0.00/0.32*	*80.73/80.11*	*55.10*	50.12	*41.57*	*↓7.99* *50.08±2.08*
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	72.87	18.78/39.43	76.37	*3.25/0.32*	*83.18/82.54*	*53.63*	49.90	57.28	*↓5.10* *52.97±2.04*

Table 24: **Race:** SuperGLUE bootstrapped validation set scores with sub-results for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. LLaMA-based results were computed with zero-shot evaluation while all other scores are derived after fine-tuning.

Model **Metrics**	**BoolQ** **Acc.**	**CB** **F1/Acc.**	**COPA** **Acc.**	**MultiRC** **F1$_\alpha$/EM**	**ReCoRD** **F1/EM**	**RTE** **Acc.**	**WiC** **Acc.**	**WSC** **Acc.**	**Average ↑**
BERT$_{\text{base}}$	69.16	38.74/58.68	62.72	60.12/13.23	**56.09/55.32**	61.30	68.67	63.12	51.82±1.67
+ GRADIEND$_{\text{Asian/Black}}$	70.45	42.71/62.23	58.88	60.37/14.17	55.50/54.70	**67.37**	**68.82**	**63.40**	↑0.80 52.62±1.89
+ GRADIEND$_{\text{Asian/White}}$	70.36	42.66/62.25	57.19	60.61/14.22	55.62/54.86	66.64	68.36	**63.40**	↑0.53 52.36±1.90
+ GRADIEND$_{\text{Black/White}}$	**70.66**	42.68/62.23	61.53	**60.91/14.80**	55.33/54.56	65.54	68.43	**63.40**	↑0.82 52.65±1.88
+ CDA	70.33	**47.60/68.92**	60.55	60.04/15.31	55.43/54.69	63.94	67.49	**63.40**	**↑1.15** **52.98**±1.82
+ DROPOUT	68.53	47.39/68.99	55.56	59.20/12.94	55.00/54.23	61.74	65.15	62.77	↓0.34 51.48±1.72
+ INLP	69.00	31.77/54.49	**63.06**	60.68/13.86	55.89/55.14	59.34	67.12	57.79	↓1.38 50.44±1.54
+ SENTDEBIAS	68.88	27.47/52.19	62.05	60.59/14.25	56.01/55.25	62.50	68.57	62.18	↓0.95 50.87±1.54
BERT$_{\text{large}}$	70.32	42.86/62.97	61.46	61.49/15.19	61.70/61.04	**67.68**	**70.82**	62.09	53.74±1.62
+ GRADIEND$_{\text{Asian/Black}}$	72.31	46.09/66.44	61.78	61.49/15.39	62.02/61.36	65.04	70.63	63.72	↑0.53 54.27±1.85
+ GRADIEND$_{\text{Asian/White}}$	72.32	46.09/66.49	64.81	59.22/15.35	61.75/61.07	67.44	70.13	63.40	**↑0.84** **54.58**±1.85
+ GRADIEND$_{\text{Black/White}}$	**72.66**	46.57/67.10	**65.38**	59.72/15.42	61.79/61.13	67.21	69.71	63.40	↑0.68 54.42±1.87
+ CDA	71.79	**47.18/67.78**	61.24	62.16/15.82	61.47/60.76	58.36	68.33	61.80	↓0.59 53.15±1.73
+ DROPOUT	71.12	45.18/65.27	53.62	62.24/16.01	**62.09/61.37**	64.72	67.99	63.40	↓0.52 53.22±1.68
+ INLP	69.73	38.31/61.17	62.86	**62.30/16.11**	61.88/61.22	67.33	70.35	**63.74**	↓0.24 53.50±1.58
+ SENTDEBIAS	70.71	43.38/63.53	61.17	59.46/14.09	61.73/61.07	66.92	70.34	62.78	↓0.07 53.68±1.67
DistilBERT	69.75	45.62/66.55	53.39	57.58/12.21	49.09/48.27	55.18	62.10	**63.40**	49.69±1.65
+ GRADIEND$_{\text{Asian/Black}}$	69.68	47.21/68.31	55.80	57.61/12.80	48.97/48.19	53.87	61.55	**63.40**	↑0.00 49.69±1.69
+ GRADIEND$_{\text{Asian/White}}$	69.88	46.71/67.71	56.08	58.19/12.19	49.03/48.23	55.03	62.15	**63.40**	↑0.38 50.07±1.70
+ GRADIEND$_{\text{Black/White}}$	69.68	45.78/66.46	53.00	58.01/12.57	49.04/48.23	53.63	61.48	**63.40**	↓0.40 49.29±1.70
+ CDA	68.86	47.84/68.89	**58.52**	58.41/12.59	48.97/48.15	60.05	**63.77**	**63.40**	↑1.11 50.80±1.79
+ DROPOUT	69.21	46.13/66.49	54.78	**59.40/13.05**	**49.77/48.97**	**60.45**	62.62	**63.40**	↑0.58 50.27±1.75
+ INLP	70.14	**47.91/69.53**	57.42	58.05/12.47	48.94/48.14	57.85	62.62	**63.40**	**↑1.14** **50.83**±1.68
+ SENTDEBIAS	**70.20**	43.97/65.37	55.50	58.19/11.79	49.04/48.24	54.83	62.26	**63.40**	↓0.22 49.47±1.62
RoBERTa	82.01	**46.41/66.62**	56.70	42.49/12.60	72.14/71.46	75.36	56.83	53.49	53.31±1.48
+ GRADIEND$_{\text{Asian/Black}}$	*62.22*	38.82/61.80	52.67	*0.00/0.32*	71.97/71.29	*52.71*	60.02	61.45	*↓7.07* *46.24±1.51*
+ GRADIEND$_{\text{Asian/White}}$	*67.59*	38.46/61.25	55.07	*38.48/8.78*	72.30/71.65	*60.54*	58.05	61.45	*↓3.27* *50.04±1.49*
+ GRADIEND$_{\text{Black/White}}$	82.29	44.76/64.30	63.42	43.32/15.49	72.08/71.42	*62.88*	61.70	61.45	↑0.41 53.72±1.67
+ CDA	81.76	46.34/66.58	***70.41***	*44.94/17.11*	71.88/71.22	**77.44**	***67.93***	**63.10**	***↑4.17*** ***57.48***±*1.71*
+ DROPOUT	*73.53*	45.03/64.77	50.32	*45.78/17.10*	72.28/71.60	*60.86*	61.03	57.62	↓2.25 51.05±1.62
+ INLP	82.53	38.64/60.68	53.43	*44.66/16.56*	72.27/71.61	74.99	*66.26*	55.41	↑0.46 53.77±1.19
+ SENTDEBIAS	**83.03**	45.99/66.04	57.50	***68.15/25.50***	**72.33/71.69**	76.86	55.56	55.06	↑2.23 55.54±1.49
GPT-2	65.56	36.86/51.74	49.35	58.79/13.69	31.64/30.93	60.14	62.51	54.47	45.49±1.28
+ GRADIEND$_{\text{Asian/Black}}$	65.36	36.15/51.05	49.65	59.09/13.58	31.62/30.92	61.78	62.70	53.52	↓0.04 45.45±1.18
+ GRADIEND$_{\text{Asian/White}}$	64.74	37.93/54.61	50.05	59.06/13.57	31.43/30.77	62.83	61.31	54.39	↑0.43 45.92±1.22
+ GRADIEND$_{\text{Black/White}}$	65.18	37.79/52.82	**52.36**	58.97/13.01	31.49/30.81	63.00	62.18	53.19	↑0.48 45.97±1.13
+ CDA	**66.07**	**40.92/55.24**	52.06	**59.62/15.46**	**31.88/31.19**	**66.13**	**63.33**	56.39	**↑1.47** **46.96**±1.27
+ DROPOUT	66.02	34.95/52.30	49.04	58.82/13.86	31.55/30.86	62.26	62.66	**58.72**	↑0.46 45.94±1.45
+ INLP	65.50	36.37/51.14	52.31	59.04/13.31	31.59/30.87	60.38	62.51	54.80	↑0.29 45.77±1.22
+ SENTDEBIAS	65.40	31.49/45.23	51.05	59.64/12.81	31.48/30.75	61.37	61.94	53.20	↓0.75 44.73±1.26
LLaMA	72.96	37.32/51.82	**86.06**	0.00/0.32	**90.42/89.70**	54.17	50.10	37.58	54.46±2.28
+ GRADIEND$_{\text{Asian/Black}}$	*57.24*	31.46/37.52	80.96	*2.11/0.32*	*87.97/87.04*	55.92	47.86	51.00	↓2.43 52.02±2.27
+ GRADIEND$_{\text{Asian/White}}$	*58.42*	26.02/35.68	79.02	***2.79/0.32***	*86.98/86.09*	**60.69**	48.10	***70.07***	↓0.05 54.40±2.22
+ GRADIEND$_{\text{Black/White}}$	*69.79*	33.49/55.30	85.07	0.10/0.32	90.22/89.43	52.66	49.61	38.52	↓0.70 53.76±2.03
+ INLP	**73.44**	35.98/53.69	86.03	0.00/0.32	89.82/89.03	58.13	50.10	36.60	↑0.38 54.84±2.12
+ SENTDEBIAS	73.08	**38.45/51.77**	**86.06**	0.00/0.32	90.30/89.55	57.40	**50.42**	36.60	**↑0.39** **54.85**±2.29
LLaMA-Instruct	75.25	31.58/32.13	**78.60**	27.43/0.52	85.32/84.68	67.31	50.37	62.20	58.07±2.29
+ GRADIEND$_{\text{Asian/Black}}$	*37.78*	22.13/49.86	*57.05*	***59.91/0.82***	*18.26/17.62*	*47.17*	49.88	63.40	*↓15.62* *42.45±2.13*
+ GRADIEND$_{\text{Asian/White}}$	*59.25*	24.31/28.79	*56.89*	30.12/2.11	*45.54/44.89*	*49.58*	48.76	*37.55*	*↓15.58* *42.49±2.47*
+ GRADIEND$_{\text{Black/White}}$	***78.19***	***45.38/64.22***	78.17	*4.28/0.32*	***87.30/86.52***	**71.20**	52.48	45.17	**↑0.58** **58.65**±2.05
+ INLP	76.28	35.27/35.79	78.25	29.33/1.15	85.31/84.65	66.31	**52.60**	57.63	↑0.28 58.35±2.34
+ SENTDEBIAS	75.02	33.85/33.94	77.26	28.30/0.80	85.27/84.64	66.06	51.46	**65.06**	↑0.46 58.53±2.41

Table 25: **Religion:** SuperGLUE bootstrapped validation set scores with sub-results for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**. LLaMA-based results were computed with zero-shot evaluation while all other scores are derived after fine-tuning.

Model Metrics	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC $F1_\alpha$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.	Average ↑
BERT$_{\text{base}}$	69.16	38.74/58.68	62.72	60.12/13.23	56.09/55.32	61.30	68.67	63.12	51.82±1.67
+ GRADIEND$_{\text{Christian/Jewish}}$	**70.91**	42.68/62.23	61.50	60.22/14.43	55.74/55.01	64.66	**70.13**	**63.40**	↑1.17 53.00±1.89
+ GRADIEND$_{\text{Christian/Muslim}}$	70.82	42.15/61.64	60.80	59.37/14.18	55.74/54.98	**65.27**	67.93	**63.40**	↑0.71 52.54±1.88
+ GRADIEND$_{\text{Jewish/Muslim}}$	70.72	43.19/62.83	60.84	**60.79/13.93**	55.04/54.29	64.43	69.72	**63.40**	↑0.89 52.71±1.89
+ CDA	70.41	**47.79/69.51**	60.50	59.24/14.32	55.58/54.85	63.96	67.81	**63.40**	**↑1.26 53.08**±1.83
+ DROPOUT	68.53	47.39/68.99	55.56	59.20/12.94	55.00/54.23	61.74	65.15	62.77	↓0.34 51.48±1.72
+ INLP	68.85	28.43/52.77	62.08	60.62/13.66	**56.09/55.33**	62.38	67.96	61.29	↓1.13 50.70±1.58
+ SENTDEBIAS	69.02	27.47/52.19	**64.40**	60.29/13.53	55.99/55.22	62.74	68.68	63.06	↓0.57 51.25±1.54
BERT$_{\text{large}}$	70.32	42.86/62.97	61.46	61.49/15.19	61.70/61.04	67.68	**70.82**	62.09	53.74±1.62
+ GRADIEND$_{\text{Christian/Jewish}}$	72.20	46.17/66.51	63.74	61.06/16.34	61.81/61.16	67.67	70.49	**63.40**	↑0.78 54.52±1.84
+ GRADIEND$_{\text{Christian/Muslim}}$	72.22	46.05/66.44	62.73	61.88/16.92	61.81/61.13	**67.80**	70.06	**63.40**	↑0.83 54.58±1.86
+ GRADIEND$_{\text{Jewish/Muslim}}$	72.03	45.74/65.91	64.83	**62.81/16.24**	61.95/61.27	66.71	70.25	**63.40**	**↑1.08 54.82**±1.85
+ CDA	**72.71**	**46.78/67.19**	**65.90**	61.05/15.03	61.79/61.11	61.75	69.30	63.10	↑0.52 54.27±1.79
+ DROPOUT	71.12	45.18/65.27	53.62	62.24/16.01	**62.09/61.37**	64.72	67.99	**63.40**	↓0.52 53.22±1.68
+ INLP	70.80	38.31/61.17	60.08	61.26/16.30	61.69/61.02	67.32	70.44	63.12	↓0.54 53.21±1.55
+ SENTDEBIAS	70.90	43.74/63.55	62.14	62.66/16.15	61.44/60.73	67.46	70.24	58.33	↓0.07 53.67±1.64
DistilBERT	69.75	45.62/66.55	53.39	57.58/12.21	49.09/48.27	55.18	62.10	**63.40**	49.69±1.65
+ GRADIEND$_{\text{Christian/Jewish}}$	69.75	45.64/66.52	53.38	57.62/12.09	49.14/48.36	55.73	61.50	**63.40**	↑0.06 49.75±1.69
+ GRADIEND$_{\text{Christian/Muslim}}$	69.75	44.73/65.35	**55.35**	57.17/11.97	49.14/48.35	55.93	61.82	**63.40**	↑0.07 49.76±1.69
+ GRADIEND$_{\text{Jewish/Muslim}}$	69.72	45.64/66.52	54.41	58.19/12.61	49.11/48.32	55.94	61.64	**63.40**	↑0.23 49.91±1.68
+ CDA	69.06	**48.37/69.57**	53.40	58.92/13.45	48.99/48.13	**60.49**	**63.25**	**63.40**	**↑0.80 50.49**±1.80
+ DROPOUT	69.21	46.13/66.49	54.78	**59.40/13.05**	**49.77/48.97**	60.45	62.62	**63.40**	↑0.58 50.27±1.75
+ INLP	69.42	44.32/64.73	55.28	57.18/11.72	49.02/48.21	57.52	62.55	**63.40**	↓0.05 49.64±1.65
+ SENTDEBIAS	**70.11**	45.62/66.55	54.73	58.43/12.24	49.02/48.23	55.66	61.89	**63.40**	↑0.12 49.81±1.64
RoBERTa	**82.01**	46.41/66.62	56.70	42.49/12.60	72.14/71.46	75.36	56.83	53.49	53.31±1.48
+ GRADIEND$_{\text{Christian/Jewish}}$	*74.98*	43.45/62.51	55.70	*45.24/17.07*	72.31/71.64	67.72	59.24	61.76	↓0.65 52.66±1.66
+ GRADIEND$_{\text{Christian/Muslim}}$	81.92	45.54/65.44	57.63	***67.66/22.48***	**72.57/71.90**	**77.31**	*63.17*	62.07	*↑3.35 56.66±1.64*
+ GRADIEND$_{\text{Jewish/Muslim}}$	*79.04*	40.34/58.98	50.38	*22.29/9.14*	72.22/71.60	*62.92*	*65.91*	61.76	↓2.19 51.12±1.65
+ CDA	*75.96*	**48.02/68.95**	***70.99***	*65.41/22.63*	71.99/71.33	74.79	***67.57***	**62.42**	***↑4.71 58.02±1.68***
+ DROPOUT	*73.53*	45.03/64.77	50.32	*45.78/17.10*	72.28/71.60	*60.86*	61.03	57.62	↓2.25 51.05±1.62
+ INLP	*75.72*	45.99/66.04	54.17	*66.07/22.54*	71.95/71.32	77.07	*64.48*	53.49	↑1.64 54.95±1.51
+ SENTDEBIAS	*75.28*	46.13/67.20	57.48	*45.58/14.63*	72.14/71.47	68.39	*64.98*	53.49	↓0.60 52.71±1.21
GPT-2	65.56	36.86/51.74	49.35	58.79/13.69	**31.64/30.93**	60.14	62.51	54.47	45.49±1.28
+ GRADIEND$_{\text{Christian/Jewish}}$	66.02	36.89/51.69	52.39	**61.03/13.64**	31.43/30.75	64.63	63.42	55.17	↑1.18 46.67±1.11
+ GRADIEND$_{\text{Christian/Muslim}}$	65.37	42.78/55.85	52.66	60.48/13.16	31.63/30.94	63.96	61.88	57.05	↑1.54 47.02±1.33
+ GRADIEND$_{\text{Jewish/Muslim}}$	**66.31**	37.60/51.10	51.38	60.12/14.32	31.32/30.63	**65.84**	63.53	55.42	↑1.15 46.64±1.26
+ CDA	66.16	**46.73/58.76**	**54.91**	59.52/14.34	31.54/30.85	65.67	**63.69**	**59.67**	**↑2.61 48.10**±1.42
+ DROPOUT	66.02	34.95/52.30	49.04	58.82/13.86	31.55/30.86	62.26	62.66	58.72	↑0.46 45.94±1.45
+ INLP	65.52	36.77/51.74	52.66	58.68/13.92	31.61/30.89	60.26	62.36	54.16	↑0.29 45.77±1.21
+ SENTDEBIAS	65.35	42.15/57.65	53.99	57.11/11.91	31.62/30.88	62.54	61.72	53.23	↑0.80 46.29±1.28
LLaMA	72.96	37.32/51.82	86.06	0.00/0.32	**90.42/89.70**	54.17	50.10	37.58	54.46±2.28
+ GRADIEND$_{\text{Christian/Jewish}}$	*61.19*	**38.95/60.67**	81.00	**0.37/0.53**	89.38/88.55	52.32	50.13	36.60	↓1.90 52.56±2.17
+ GRADIEND$_{\text{Christian/Muslim}}$	74.34	24.92/44.53	80.13	0.10/0.42	89.87/89.16	51.18	50.12	36.60	↓2.35 52.11±2.12
+ GRADIEND$_{\text{Jewish/Muslim}}$	*68.66*	30.37/42.72	79.06	0.10/0.42	*88.71/87.87*	**59.61**	47.75	**40.54**	↓1.87 52.59±2.17
+ INLP	**74.54**	39.56/51.89	83.10	0.00/0.32	90.31/89.58	56.66	**50.59**	36.60	**↑0.21 54.66**±2.30
+ SENTDEBIAS	72.80	35.03/50.01	**86.11**	0.00/0.32	90.12/89.40	55.25	50.10	37.58	↓0.17 54.29±2.28
LLaMA-Instruct	75.25	31.58/32.13	78.60	27.43/0.52	85.32/84.68	67.31	50.37	62.20	58.07±2.29
+ GRADIEND$_{\text{Christian/Jewish}}$	***78.50***	***52.18/73.15***	80.10	*7.82/0.21*	***87.42/86.64***	**74.80**	51.37	*42.32*	**↑2.03 60.10**±1.95
+ GRADIEND$_{\text{Christian/Muslim}}$	75.56	37.38/55.66	70.97	*12.57/0.10*	*73.80/73.05*	69.67	50.12	*37.53*	↓4.30 53.77±2.17
+ GRADIEND$_{\text{Jewish/Muslim}}$	74.37	*43.66/64.14*	**82.28**	*7.93/0.63*	84.72/84.03	69.01	51.05	58.68	↑1.67 59.74±2.18
+ INLP	75.36	33.47/33.95	77.26	*22.80/0.42*	85.55/84.91	66.31	**52.43**	61.44	↓0.15 57.92±2.40
+ SENTDEBIAS	75.06	33.85/33.94	77.26	**28.99/0.64**	85.35/84.72	65.93	51.82	**64.40**	↑0.46 58.53±2.41

Table 26: **Gender:** SEAT bootstrapped effect sizes for encoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	SEAT-6 ⊥0.0	SEAT-6b ⊥0.0	SEAT-7 ⊥0.0	SEAT-7b ⊥0.0	SEAT-8 ⊥0.0	SEAT-8b ⊥0.0	Absolute Average ↓
BERT$_{base}$	0.84±0.29	0.20±0.25	0.57±0.62	1.03±0.49	0.54±0.48	0.45±0.52	0.61±0.29
+ GRADIEND$_{Female/Male}$	0.59±0.21	**−0.03**±0.15	0.18±0.57	0.92±0.45	0.53±0.46	0.64±0.36	↓0.10 0.51±0.19
+ GRADIEND$_{Female}$	0.84±0.46	0.27±0.40	1.05±0.34	1.18±0.31	0.72±0.32	0.68±0.34	↑0.19 0.79±0.24
+ GRADIEND$_{Male}$	0.95±0.27	0.43±0.32	0.24±0.54	0.33±0.51	−0.05±0.46	−0.35±0.40	↓0.18 0.43±0.16
+ CDA	0.48±0.18	−0.05±0.13	0.07±0.61	0.74±0.50	0.38±0.50	0.47±0.44	↓0.21 0.40±0.20
+ DROPOUT	*0.16±0.37*	0.17±0.27	0.24±0.60	0.81±0.42	0.67±0.47	0.53±0.43	↓0.16 0.45±0.25
+ INLP	0.46±0.20	−0.08±0.13	*−0.65±0.42*	*__−0.18__±0.64*	−0.22±0.52	−0.40±0.53	↓0.24 0.37±0.19
+ RLACE	0.84±0.29	0.19±0.25	0.58±0.62	1.03±0.49	0.53±0.48	0.44±0.52	↓0.00 0.61±0.29
+ LEACE	0.83±0.29	0.20±0.25	0.56±0.62	1.05±0.49	0.52±0.48	0.44±0.52	↓0.00 0.61±0.29
+ SENTDEBIAS	0.38±0.24	*−0.21±0.14*	−0.46±0.45	0.25±0.62	0.39±0.46	0.16±0.49	↓0.27 0.34±0.13
+ GRADIEND$_{Female/Male}$ + INLP	0.58±0.18	−0.10±0.11	−0.42±0.41	0.21±0.59	0.04±0.43	0.13±0.44	↓0.31 0.30±0.12
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	0.43±0.20	*−0.21±0.12*	−0.34±0.44	0.57±0.50	0.43±0.40	0.55±0.34	↓0.18 0.43±0.14
+ CDA + INLP	0.43±0.17	*−0.19±0.08*	−0.41±0.54	0.22±0.58	**−0.04**±0.47	0.25±0.49	↓0.30 0.30±0.14
+ DROPOUT + SENTDEBIAS	*__−0.16__±0.28*	−0.11±0.17	**−0.04**±0.46	0.68±0.39	0.57±0.38	0.40±0.39	↓0.25 0.36±0.14
+ CDA + SENTDEBIAS	0.45±0.17	−0.10±0.11	−0.44±0.51	0.55±0.56	0.31±0.46	0.41±0.44	↓0.22 0.38±0.17
+ DROPOUT + INLP	*−0.17±0.23*	−0.12±0.14	−0.44±0.46	0.40±0.42	0.17±0.46	**0.09**±0.44	**↓0.34** **0.27**±0.12
BERT$_{large}$	0.69±0.25	0.33±0.24	0.68±0.66	0.57±0.54	0.49±0.45	0.35±0.38	0.52±0.26
+ GRADIEND$_{Female/Male}$	0.44±0.15	0.06±0.07	0.97±0.31	0.74±0.34	0.59±0.15	0.48±0.24	↑0.03 0.55±0.13
+ GRADIEND$_{Female}$	*1.39±0.21*	*0.92±0.24*	1.24±0.21	1.22±0.23	0.89±0.18	*0.93±0.17*	*↑0.58 1.10±0.13*
+ GRADIEND$_{Male}$	0.34±0.32	*−0.35±0.39*	*−0.64±0.42*	−0.30±0.48	*−0.45±0.33*	0.04±0.39	↓0.14 0.38±0.16
+ CDA	0.63±0.22	0.05±0.16	0.64±0.63	0.89±0.52	0.82±0.39	0.71±0.37	↑0.11 0.63±0.24
+ DROPOUT	0.90±0.26	0.51±0.34	0.17±0.49	1.27±0.31	0.46±0.42	0.77±0.44	↑0.17 0.69±0.22
+ INLP	**0.30**±0.19	*−0.09±0.17*	−0.11±0.71	**0.08**±0.58	−0.34±0.51	*−0.42±0.35*	↓0.23 0.29±0.15
+ RLACE	0.70±0.25	0.33±0.24	0.68±0.66	0.57±0.54	0.49±0.45	0.35±0.38	↑0.00 0.52±0.26
+ LEACE	0.69±0.25	0.30±0.24	0.73±0.65	0.58±0.54	0.52±0.46	0.36±0.38	↑0.01 0.53±0.26
+ SENTDEBIAS	**0.30**±0.21	*−0.09±0.18*	**0.04**±0.73	0.16±0.51	−0.18±0.59	**−0.03**±0.37	**↓0.29** **0.23**±0.14
+ GRADIEND$_{Female/Male}$ + INLP	0.34±0.16	*−0.05±0.08*	0.16±0.41	0.63±0.43	0.37±0.18	0.26±0.32	↓0.21 0.31±0.13
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	0.42±0.15	0.03±0.07	0.73±0.38	0.71±0.34	0.53±0.14	0.47±0.24	↓0.04 0.48±0.13
+ CDA + INLP	0.58±0.19	*−0.12±0.13*	*−0.55±0.55*	0.53±0.59	**−0.04**±0.46	0.28±0.40	↓0.14 0.38±0.16
+ DROPOUT + SENTDEBIAS	0.60±0.25	−0.01±0.28	−0.32±0.39	1.07±0.32	0.09±0.47	0.53±0.48	↓0.05 0.48±0.15
+ CDA + SENTDEBIAS	0.60±0.21	**−0.00**±0.15	0.45±0.65	0.83±0.54	0.69±0.41	0.64±0.38	↑0.03 0.55±0.23
+ DROPOUT + INLP	0.58±0.24	0.12±0.35	*−0.68±0.36*	0.80±0.34	*−0.51±0.42*	0.34±0.51	↓0.00 0.52±0.15
DistilBERT	0.82±0.24	0.25±0.21	0.65±0.70	1.42±0.25	0.50±0.58	1.13±0.27	0.80±0.24
+ GRADIEND$_{Female/Male}$	0.82±0.24	0.23±0.20	0.65±0.70	1.43±0.24	0.50±0.58	1.13±0.27	↓0.00 0.80±0.24
+ GRADIEND$_{Female}$	0.71±0.21	0.08±0.20	0.73±0.64	1.49±0.22	0.56±0.49	1.17±0.25	↓0.01 0.80±0.22
+ GRADIEND$_{Male}$	*1.58±0.20*	*1.14±0.31*	0.89±0.53	1.23±0.39	0.67±0.36	0.92±0.32	↑0.27 1.07±0.25
+ CDA	0.81±0.21	0.08±0.12	0.53±0.82	1.42±0.28	0.27±0.59	1.20±0.30	↓0.06 0.74±0.21
+ DROPOUT	1.01±0.22	0.45±0.25	0.58±0.58	1.15±0.39	0.39±0.61	1.02±0.39	↓0.02 0.78±0.26
+ INLP	0.67±0.21	*−0.17±0.11*	0.23±0.49	1.13±0.36	−0.29±0.57	1.12±0.25	↓0.18 0.62±0.13
+ RLACE	0.63±0.22	*−0.08±0.12*	−0.54±0.56	1.10±0.37	−0.12±0.55	0.98±0.28	↓0.20 0.60±0.14
+ LEACE	0.73±0.22	0.06±0.12	−0.38±0.58	1.04±0.37	0.14±0.45	0.95±0.25	↓0.23 0.57±0.12
+ SENTDEBIAS	**0.57**±0.21	*−0.19±0.10*	−0.22±0.52	1.24±0.31	**−0.04**±0.51	1.01±0.27	↓0.22 0.58±0.12
+ GRADIEND$_{Female/Male}$ + INLP	0.68±0.20	*−0.18±0.11*	0.21±0.51	1.12±0.36	−0.31±0.57	1.11±0.26	↓0.18 0.62±0.13
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	**0.57**±0.21	*−0.19±0.10*	−0.22±0.53	1.24±0.31	**−0.04**±0.51	1.02±0.26	↓0.22 0.58±0.12
+ CDA + INLP	0.81±0.20	*−0.07±0.08*	**0.03**±0.70	0.89±0.53	−0.46±0.50	0.90±0.41	↓0.23 0.57±0.16
+ DROPOUT + SENTDEBIAS	0.68±0.20	0.07±0.14	0.04±0.49	0.87±0.43	−0.18±0.56	0.92±0.38	↓0.30 0.50±0.15
+ CDA + SENTDEBIAS	0.71±0.20	*__−0.05__±0.08*	−0.05±0.74	1.31±0.33	−0.07±0.55	1.13±0.32	↓0.18 0.63±0.14
+ DROPOUT + INLP	0.59±0.20	0.23±0.14	0.18±0.42	*__0.11__±0.50*	−0.45±0.48	**0.79**±0.42	*__↓0.38__* *__0.42__±0.13*
RoBERTa	0.78±0.31	0.16±0.26	−0.20±0.57	0.81±0.37	0.40±0.57	1.00±0.31	0.58±0.17
+ GRADIEND$_{Female/Male}$	0.38±0.21	0.18±0.18	−0.21±0.46	0.79±0.28	0.39±0.45	0.86±0.24	↓0.10 0.48±0.13
+ GRADIEND$_{Female}$	*1.79±0.11*	*1.66±0.16*	*0.60±0.20*	0.60±0.21	0.22±0.11	*__0.28__±0.11*	*↑0.28 0.86±0.10*
+ GRADIEND$_{Male}$	*−0.24±0.45*	*−0.57±0.34*	−0.28±0.55	0.51±0.43	−0.20±0.76	0.37±0.61	↓0.17 0.41±0.15
+ CDA	0.48±0.30	−0.05±0.20	−0.23±0.57	0.59±0.37	0.16±0.58	0.97±0.24	↓0.13 0.45±0.14
+ DROPOUT	0.24±0.30	−0.25±0.33	−0.67±0.41	0.72±0.38	**0.01**±0.50	0.86±0.33	↓0.08 0.49±0.12
+ INLP	0.38±0.27	−0.22±0.15	−0.87±0.33	**0.22**±0.41	−0.24±0.50	0.59±0.39	↓0.14 0.44±0.14
+ RLACE	0.78±0.31	0.16±0.26	−0.20±0.57	0.81±0.37	0.39±0.57	1.00±0.31	↓0.00 0.58±0.17
+ LEACE	0.78±0.31	0.16±0.26	**−0.16**±0.57	0.82±0.37	0.41±0.56	1.01±0.31	↑0.00 0.58±0.17
+ SENTDEBIAS	0.53±0.26	−0.11±0.16	−0.62±0.43	0.66±0.31	−0.04±0.55	0.80±0.28	↓0.09 0.49±0.14
+ GRADIEND$_{Female/Male}$ + INLP	*0.15±0.22*	0.01±0.14	−0.46±0.40	0.43±0.31	0.05±0.39	0.68±0.32	**↓0.25** **0.33**±0.13
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	0.31±0.20	0.05±0.16	−0.38±0.41	0.71±0.27	0.36±0.41	0.82±0.23	↓0.14 0.44±0.11
+ CDA + INLP	0.41±0.24	*−0.23±0.11*	−0.74±0.42	0.38±0.42	−0.19±0.49	0.92±0.31	↓0.09 0.49±0.15
+ DROPOUT + SENTDEBIAS	*0.16±0.23*	−0.34±0.24	−0.93±0.32	0.68±0.29	−0.15±0.47	0.87±0.27	↓0.04 0.54±0.12
+ CDA + SENTDEBIAS	0.51±0.28	**0.00**±0.14	−0.17±0.59	0.64±0.37	−0.05±0.54	1.00±0.24	↓0.13 0.45±0.14
+ DROPOUT + INLP	*__0.07__±0.24*	*−0.41±0.19*	−0.80±0.36	0.49±0.30	−0.14±0.43	0.70±0.27	↓0.12 0.45±0.11

Table 27: **Gender:** SEAT bootstrapped effect sizes for decoder-only models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	SEAT-6 ⊢0.0	SEAT-6b ⊢0.0	SEAT-7 ⊢0.0	SEAT-7b ⊢0.0	SEAT-8 ⊢0.0	SEAT-8b ⊢0.0	Absolute Average ↓
GPT-2	0.27±0.61	0.03±0.47	0.03±0.39	0.14±0.61	−0.07±0.76	−0.14±0.72	0.24±0.29
+ GRADIEND$_{Female/Male}$	0.34±0.56	0.02±0.50	0.42±0.61	0.51±0.63	−0.05±0.63	0.18±0.71	↑0.09 0.33±0.39
+ GRADIEND$_{Female}$	**0.20**±0.57	**0.00**±0.41	0.14±0.54	0.39±0.81	0.05±0.67	0.19±0.65	↑0.01 0.25±0.39
+ GRADIEND$_{Male}$	0.31±0.60	−0.02±0.55	0.42±0.59	0.48±0.60	−0.09±0.65	0.11±0.69	↑0.09 0.33±0.36
+ CDA	0.34±0.45	0.03±0.28	0.14±0.55	0.43±0.70	0.28±0.90	0.02±0.77	↑0.07 0.31±0.29
+ DROPOUT	0.33±0.47	0.03±0.34	0.66±0.73	0.85±0.59	0.45±0.61	0.22±0.79	↑0.24 0.48±0.24
+ INLP	0.23±0.58	−0.01±0.46	**0.00**±0.36	0.12±0.59	−0.09±0.76	−0.16±0.73	↓0.01 0.23±0.26
+ RLACE	0.26±0.65	0.03±0.49	0.02±0.31	**0.11**±0.43	−0.08±0.69	−0.20±0.56	↓**0.02** **0.22**±0.24
+ LEACE	0.33±0.63	0.11±0.48	0.06±0.37	0.14±0.57	−**0.01**±0.67	−0.10±0.70	↑0.00 0.24±0.26
+ SENTDEBIAS	0.26±0.55	−0.06±0.21	−0.26±0.52	0.12±1.01	0.17±0.84	−0.22±1.19	↑0.11 0.34±0.27
+ GRADIEND$_{Female/Male}$ + INLP	0.29±0.54	−0.02±0.49	0.38±0.56	0.46±0.56	−0.06±0.62	0.17±0.70	↑0.07 0.31±0.36
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	0.35±0.46	−0.04±0.25	−0.02±0.77	0.61±0.93	−0.54±0.42	0.10±1.10	↑0.18 0.42±0.25
+ CDA + INLP	0.32±0.45	0.01±0.29	0.10±0.57	0.38±0.71	0.26±0.90	−**0.00**±0.77	↑0.06 0.30±0.29
+ DROPOUT + SENTDEBIAS	0.40±0.46	0.11±0.35	0.43±0.66	0.83±0.63	−0.18±0.61	−0.08±0.73	↑0.18 0.42±0.19
+ CDA + SENTDEBIAS	0.39±0.39	0.02±0.23	−0.05±0.58	0.45±1.12	0.44±1.07	−0.04±1.14	↑0.16 0.40±0.26
+ DROPOUT + INLP	0.31±0.46	**0.00**±0.34	0.62±0.76	0.82±0.60	0.43±0.63	0.20±0.80	↑0.22 0.46±0.23
LLaMA	1.29±0.19	0.37±0.12	0.41±0.45	1.40±0.25	0.82±0.35	1.30±0.23	0.93±0.16
+ GRADIEND$_{Female/Male}$	0.99±0.19	0.20±0.09	−0.20±0.32	1.27±0.23	0.22±0.34	1.09±0.22	↓*0.26* *0.67±0.10*
+ GRADIEND$_{Female}$	0.95±0.22	0.25±0.10	0.79±0.35	1.39±0.24	0.89±0.32	**0.94**±0.30	↓0.06 0.87±0.14
+ GRADIEND$_{Male}$	1.19±0.17	0.28±0.12	−0.39±0.46	1.35±0.25	0.41±0.42	1.29±0.19	↓0.11 0.82±0.11
+ INLP	***0.89***±*0.20*	*0.11±0.10*	0.40±0.44	**1.03**±0.32	0.78±0.41	1.00±0.28	↓0.23 0.70±0.16
+ RLACE	1.30±0.19	0.37±0.12	0.41±0.45	1.39±0.25	0.82±0.35	1.29±0.23	↓0.00 0.93±0.16
+ LEACE	1.29±0.19	0.36±0.12	0.39±0.46	1.39±0.25	0.79±0.37	1.29±0.23	↓0.01 0.92±0.17
+ SENTDEBIAS	1.04±0.21	0.14±0.11	0.16±0.34	1.06±0.28	0.29±0.32	0.95±0.24	↓***0.32*** ***0.61***±*0.14*
+ GRADIEND$_{Female/Male}$ + INLP	0.96±0.19	***0.08***±*0.07*	−**0.16**±0.27	1.17±0.20	**0.17**±0.31	1.07±0.19	↓***0.33*** ***0.61***±*0.09*
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	0.95±0.19	*0.15±0.08*	−0.17±0.31	1.24±0.23	***0.12***±*0.30*	1.08±0.21	↓*0.30* *0.63±0.10*
LLaMA-Instruct	0.88±0.26	0.21±0.14	1.11±0.37	1.45±0.22	0.63±0.28	1.11±0.26	0.90±0.16
+ GRADIEND$_{Female/Male}$	0.40±0.26	0.32±0.12	0.57±0.39	*0.70±0.40*	0.16±0.29	0.76±0.25	↓*0.41* *0.49±0.15*
+ GRADIEND$_{Female}$	0.81±0.26	0.25±0.13	0.85±0.43	1.06±0.36	0.51±0.31	0.75±0.31	↓0.19 0.71±0.20
+ GRADIEND$_{Male}$	1.17±0.34	0.05±0.17	*−0.41±0.67*	1.13±0.39	0.13±0.41	1.24±0.29	↓0.19 0.71±0.13
+ INLP	0.70±0.26	*−0.04±0.10*	0.42±0.50	1.01±0.37	0.36±0.39	0.86±0.35	↓0.33 0.57±0.19
+ RLACE	0.83±0.26	0.14±0.13	0.62±0.53	1.19±0.31	0.45±0.33	0.94±0.27	↓0.20 0.70±0.20
+ LEACE	0.82±0.26	0.13±0.13	0.62±0.54	1.15±0.31	0.49±0.32	0.95±0.27	↓0.20 0.69±0.20
+ SENTDEBIAS	0.68±0.28	−**0.01**±0.12	***0.14***±*0.32*	*0.73±0.32*	0.19±0.24	**0.72**±0.22	↓*0.47* *0.43±0.13*
+ GRADIEND$_{Female/Male}$ + INLP	0.36±0.26	0.17±0.10	*0.27±0.38*	***0.69***±*0.39*	*−0.07±0.32*	**0.72**±0.28	↓***0.50*** ***0.39***±*0.13*
+ GRADIEND$_{Female/Male}$ + SENTDEBIAS	***0.34***±*0.26*	0.23±0.11	0.59±0.38	*0.72±0.40*	**0.05**±0.30	0.78±0.26	↓*0.43* *0.46±0.14*

Table 28: **Race:** SEAT bootstrapped effect sizes for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	ABW1 ↕0.0	ABW2 ↕0.0	SEAT-3 ↕0.0	SEAT-3b ↕0.0	SEAT-4 ↕0.0	SEAT-5 ↕0.0	SEAT-5b ↕0.0	Absolute Average ↕0.0
$\text{BERT}_{\text{base}}$	0.53±0.78	0.50±0.23	0.72±0.36	−0.15±0.48	0.68±0.48	0.77±0.56	0.04±0.33	51.64±25.85
+ $\text{GRADIEND}_{\text{Asian/Black}}$	0.82±0.65	0.40±0.28	0.83±0.37	**0.02**±0.49	0.85±0.54	0.97±0.53	0.06±0.29	↑8.31 59.96±27.62
+ $\text{GRADIEND}_{\text{Asian/White}}$	1.08±0.44	**0.29**±0.28	0.76±0.40	−0.09±0.50	0.74±0.56	0.97±0.50	0.06±0.32	↑8.29 59.94±24.04
+ $\text{GRADIEND}_{\text{Black/White}}$	0.56±0.73	0.47±0.23	0.70±0.37	−0.20±0.49	0.65±0.49	0.75±0.55	**0.01**±0.34	↓0.62 51.02±25.56
+ CDA	0.44±0.46	0.35±0.16	**0.13**±0.46	−0.19±0.34	**−0.04**±0.51	**0.22**±0.39	0.03±0.30	↓25.74 **25.90**±12.84
+ DROPOUT	0.38±0.48	0.64±0.24	0.64±0.35	−0.18±0.46	0.74±0.41	0.32±0.44	0.02±0.25	↓7.36 44.29±17.45
+ INLP	**0.37**±0.65	0.57±0.21	0.68±0.31	−0.16±0.43	0.63±0.38	0.81±0.38	−0.03±0.33	↓2.10 49.55±18.90
+ SENTDEBIAS	0.55±0.78	0.50±0.23	0.72±0.36	−0.13±0.48	0.68±0.48	0.78±0.56	0.05±0.34	↑0.46 52.11±26.04
$\text{BERT}_{\text{large}}$	−0.45±0.38	1.05±0.16	0.59±0.18	**−0.03**±0.41	0.43±0.22	0.28±0.38	−0.12±0.34	45.00±10.23
+ $\text{GRADIEND}_{\text{Asian/Black}}$	0.08±0.43	0.95±0.19	0.77±0.16	0.28±0.36	0.67±0.18	0.46±0.38	**0.01**±0.30	↑4.37 49.37±12.55
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.12±0.48	0.89±0.20	0.84±0.15	0.16±0.34	*0.85±0.17*	0.50±0.41	0.06±0.29	↑6.82 51.83±12.79
+ $\text{GRADIEND}_{\text{Black/White}}$	−0.24±0.43	1.03±0.18	0.68±0.18	0.12±0.41	0.56±0.22	0.39±0.41	−0.05±0.33	↑1.68 46.69±11.31
+ CDA	0.10±0.42	0.81±0.14	**0.49**±0.19	0.12±0.31	0.39±0.25	0.69±0.26	0.12±0.25	↓3.94 **41.06**±12.97
+ DROPOUT	**0.06**±0.31	**0.81**±0.25	0.62±0.16	0.10±0.27	0.55±0.22	0.49±0.20	0.19±0.22	↓3.07 41.94±11.52
+ INLP	−0.76±0.36	0.96±0.16	0.57±0.22	−0.07±0.37	**0.34**±0.30	**0.21**±0.35	−0.11±0.29	↑0.40 45.40±10.74
+ SENTDEBIAS	−0.42±0.39	1.02±0.16	0.59±0.18	−0.06±0.36	0.43±0.22	0.27±0.38	−0.27±0.32	↑0.43 45.43±9.88
DistilBERT	0.70±0.49	0.23±0.20	0.04±0.54	−0.22±0.45	−0.04±0.72	0.03±0.50	−0.09±0.35	30.04±16.11
+ $\text{GRADIEND}_{\text{Asian/Black}}$	0.69±0.50	0.22±0.20	0.03±0.52	−0.37±0.48	−0.04±0.71	0.10±0.51	−0.08±0.40	↑1.72 31.76±15.95
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.79±0.42	**0.19**±0.21	**0.00**±0.56	−0.12±0.45	−0.14±0.73	−0.09±0.46	−0.13±0.30	↑0.12 30.16±16.30
+ $\text{GRADIEND}_{\text{Black/White}}$	0.66±0.54	0.24±0.20	−0.02±0.51	−0.42±0.49	−0.15±0.69	**−0.02**±0.50	−0.13±0.41	↑2.54 32.58±16.58
+ CDA	0.73±0.43	0.33±0.17	−0.16±0.39	−0.37±0.43	−0.41±0.46	0.09±0.38	0.24±0.30	↑5.50 35.54±11.69
+ DROPOUT	0.99±0.35	0.38±0.25	0.19±0.32	−0.55±0.35	**0.02**±0.45	0.23±0.40	0.26±0.29	↑10.53 40.57±12.83
+ INLP	**0.22**±0.47	0.32±0.19	−0.07±0.40	**−0.11**±0.31	−0.20±0.52	0.14±0.37	**−0.03**±0.27	↓8.85 **21.19**±12.04
+ SENTDEBIAS	0.72±0.48	0.23±0.20	0.06±0.54	−0.21±0.46	0.04±0.73	0.04±0.50	−0.09±0.35	↑0.12 30.16±16.30
RoBERTa	−0.04±0.32	0.28±0.40	0.72±0.37	0.01±0.55	0.79±0.47	0.69±0.34	**0.07**±0.30	42.72±16.70
+ $\text{GRADIEND}_{\text{Asian/Black}}$	*0.64±0.17*	0.11±0.28	0.87±0.27	0.10±0.24	0.67±0.31	0.52±0.26	0.11±0.17	↑1.54 44.26±11.86
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.52±0.25	0.13±0.38	0.78±0.33	−0.18±0.28	**0.49**±0.48	**0.48**±0.32	−0.12±0.16	↓3.04 39.67±13.78
+ $\text{GRADIEND}_{\text{Black/White}}$	**−0.02**±0.31	0.43±0.36	0.63±0.39	−0.01±0.55	0.72±0.51	0.67±0.34	0.09±0.30	↓0.55 42.17±17.46
+ CDA	0.10±0.34	0.15±0.28	**0.43**±0.36	−0.31±0.42	0.58±0.43	0.65±0.28	−0.33±0.29	↓4.81 **37.91**±14.66
+ DROPOUT	0.30±0.29	0.59±0.30	0.66±0.31	0.24±0.36	0.93±0.32	1.14±0.23	0.38±0.19	↑18.00 60.72±15.59
+ INLP	0.10±0.39	**0.04**±0.50	0.75±0.35	−0.02±0.54	0.95±0.40	0.71±0.33	0.09±0.29	↑2.23 44.94±15.92
+ SENTDEBIAS	−0.05±0.32	0.25±0.39	0.72±0.37	**0.00**±0.53	0.78±0.47	0.75±0.33	0.18±0.30	↑1.01 43.73±16.81
GPT-2	0.82±0.67	−0.14±0.36	0.33±1.11	0.26±0.49	0.13±1.10	0.39±1.11	0.25±0.49	46.92±32.61
+ $\text{GRADIEND}_{\text{Asian/Black}}$	**0.24**±0.85	0.05±0.39	−0.41±0.98	**−0.03**±0.47	−0.52±0.88	−0.60±0.66	−0.21±0.48	↓6.51 40.41±29.60
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.60±0.86	−0.06±0.41	**0.13**±1.04	0.24±0.41	**−0.07**±1.02	0.33±1.14	0.22±0.48	↓5.99 40.93±27.33
+ $\text{GRADIEND}_{\text{Black/White}}$	0.83±0.66	−0.15±0.35	0.36±1.13	0.27±0.49	0.15±1.11	0.40±1.11	0.25±0.49	↑1.03 47.95±33.43
+ CDA	−0.30±1.49	−0.07±0.46	−0.26±1.10	0.07±0.60	−0.49±0.96	0.65±1.23	0.17±0.41	↑1.83 48.75±28.71
+ DROPOUT	0.65±0.79	−0.06±0.30	0.32±1.22	0.15±0.45	0.21±1.29	**0.08**±1.00	**0.06**±0.33	↓8.29 38.63±37.73
+ INLP	0.82±0.67	−0.14±0.36	0.33±1.12	0.26±0.49	0.13±1.10	0.39±1.11	0.24±0.49	↓0.02 46.90±32.59
+ SENTDEBIAS	0.28±0.65	**0.01**±0.33	0.69±0.73	0.37±0.40	0.47±0.74	0.22±0.38	0.17±0.40	↓9.65 **37.27**±20.70
LLaMA	0.57±0.21	0.21±0.17	0.18±0.25	0.04±0.32	**−0.01**±0.31	0.10±0.24	−0.12±0.23	21.33±8.25
+ $\text{GRADIEND}_{\text{Asian/Black}}$	0.71±0.12	*−0.21±0.19*	**0.12**±0.23	−0.12±0.27	−0.08±0.21	0.22±0.23	−0.02±0.17	↑1.61 22.95±6.53
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.66±0.13	0.16±0.17	0.27±0.26	0.21±0.24	0.08±0.27	0.18±0.29	**−0.01**±0.19	↑3.01 24.35±10.56
+ $\text{GRADIEND}_{\text{Black/White}}$	**0.56**±0.16	**0.13**±0.18	0.22±0.27	0.11±0.28	0.03±0.30	0.14±0.28	−0.17±0.20	↑0.55 21.88±8.50
+ INLP	0.57±0.21	0.21±0.17	0.13±0.24	0.13±0.30	−0.03±0.30	**0.09**±0.23	−0.11±0.24	↓0.32 **21.01**±7.88
+ SENTDEBIAS	0.61±0.20	0.19±0.18	0.20±0.26	**0.03**±0.32	0.03±0.31	0.14±0.24	−0.13±0.24	↑1.01 22.34±8.75
LLaMA-Instruct	0.70±0.21	0.14±0.37	0.22±0.30	**0.40**±0.26	0.38±0.35	0.31±0.21	0.15±0.25	33.95±14.26
+ $\text{GRADIEND}_{\text{Asian/Black}}$	*1.02±0.01*	*−0.31±0.01*	*−1.17±0.01*	*−0.47±0.02*	*−1.26±0.01*	*1.20±0.01*	*0.56±0.01*	↑*51.68* *85.64±0.59*
+ $\text{GRADIEND}_{\text{Asian/White}}$	0.94±0.13	**0.05**±0.29	**0.03**±0.21	0.48±0.13	**−0.16**±0.20	0.52±0.14	0.44±0.13	↑5.47 39.42±5.47
+ $\text{GRADIEND}_{\text{Black/White}}$	**0.69**±0.23	0.24±0.42	0.45±0.22	0.51±0.22	0.63±0.25	0.44±0.19	0.17±0.19	↑11.13 45.09±12.77
+ INLP	0.70±0.21	0.17±0.37	0.15±0.31	0.51±0.25	0.34±0.36	**0.30**±0.21	0.21±0.25	↑1.14 35.09±14.39
+ SENTDEBIAS	0.69±0.21	0.16±0.38	0.21±0.30	0.40±0.26	0.37±0.35	0.32±0.21	**0.15**±0.25	↓0.04 **33.91**±14.22

Table 29: **Religion:** SEAT bootstrapped effect sizes for all models. Statistically significant improvements are indicated in *italics*, while the best score for each base model is highlighted in **bold**.

Model	SEAT-REL1 →0.0	SEAT-REL1b →0.0	SEAT-REL2 →0.0	SEAT-REL2b →0.0	Absolute Average →0.0
BERT$_{\text{base}}$	0.18±0.40	−0.10±0.31	0.73±0.40	0.43±0.40	38.29±21.39
+ GRADIEND$_{\text{Christian/Jewish}}$	0.22±0.41	−0.15±0.32	0.81±0.38	0.44±0.40	↑3.79 42.07±20.57
+ GRADIEND$_{\text{Christian/Muslim}}$	−0.29±0.42	−0.46±0.28	0.72±0.41	0.37±0.42	↑8.48 46.77±17.84
+ GRADIEND$_{\text{Jewish/Muslim}}$	0.61±0.31	0.12±0.30	0.82±0.35	0.45±0.37	↑12.74 51.03±23.95
+ CDA	−0.16±0.25	−0.08±0.22	**0.21**±0.35	**−0.00**±0.27	**↓22.76** **15.53**±9.86
+ DROPOUT	−0.21±0.43	−0.20±0.34	0.66±0.36	0.39±0.37	↓0.10 38.18±15.73
+ INLP	**−0.05**±0.40	−0.32±0.29	0.61±0.32	0.36±0.35	↓1.86 36.43±14.46
+ SENTDEBIAS	0.37±0.28	**0.01**±0.27	0.62±0.36	0.34±0.36	↓2.18 36.11±20.86
BERT$_{\text{large}}$	0.42±0.40	0.18±0.32	1.34±0.25	1.03±0.36	74.96±24.49
+ GRADIEND$_{\text{Christian/Jewish}}$	0.56±0.35	0.42±0.31	1.30±0.23	1.14±0.28	↑10.61 85.57±22.90
+ GRADIEND$_{\text{Christian/Muslim}}$	0.36±0.34	0.35±0.29	1.33±0.19	1.13±0.26	↑4.46 79.42±20.39
+ GRADIEND$_{\text{Jewish/Muslim}}$	−0.12±0.36	−0.19±0.24	1.40±0.20	1.14±0.32	↓2.18 72.79±13.50
+ CDA	0.14±0.22	0.24±0.16	1.20±0.25	1.01±0.24	↓9.86 65.10±16.24
+ DROPOUT	0.94±0.32	*0.89±0.37*	1.15±0.22	**0.67**±0.37	↑16.40 91.36±26.14
+ INLP	**0.03**±0.43	**−0.11**±0.31	**1.07**±0.30	0.82±0.39	**↓19.35** **55.61**±16.95
+ SENTDEBIAS	0.16±0.34	0.16±0.32	1.13±0.31	1.01±0.38	↓12.20 62.76±24.05
DistilBERT	0.12±0.36	0.17±0.28	0.58±0.43	0.32±0.44	32.25±26.27
+ GRADIEND$_{\text{Christian/Jewish}}$	0.18±0.34	0.20±0.27	0.61±0.41	0.34±0.42	↑2.05 34.30±26.77
+ GRADIEND$_{\text{Christian/Muslim}}$	0.30±0.31	0.30±0.26	0.56±0.39	0.30±0.41	↑4.83 37.08±26.80
+ GRADIEND$_{\text{Jewish/Muslim}}$	0.39±0.31	0.35±0.27	0.62±0.40	0.40±0.43	↑12.02 44.27±28.66
+ CDA	−0.30±0.22	0.11±0.17	0.25±0.40	**0.10**±0.35	**↓10.65** **21.60**±12.08
+ DROPOUT	−0.38±0.32	**−0.04**±0.27	**0.25**±0.37	0.21±0.33	↓7.54 24.71±13.09
+ INLP	**−0.11**±0.37	0.05±0.30	0.45±0.41	0.25±0.46	↓5.99 26.26±18.59
+ SENTDEBIAS	0.28±0.22	0.26±0.24	0.40±0.32	0.20±0.38	↓2.82 29.43±21.31
RoBERTa	−0.17±0.48	−0.66±0.39	−0.09±0.41	−0.48±0.43	39.31±21.41
+ GRADIEND$_{\text{Christian/Jewish}}$	0.13±0.36	−0.15±0.47	0.28±0.27	*0.62±0.23*	↓6.45 32.86±14.36
+ GRADIEND$_{\text{Christian/Muslim}}$	−0.58±0.37	−0.77±0.32	**0.04**±0.21	**0.05**±0.28	↓0.47 38.84±16.29
+ GRADIEND$_{\text{Jewish/Muslim}}$	−0.17±0.36	−0.50±0.33	−0.14±0.28	−0.08±0.30	↓14.02 25.29±16.52
+ CDA	**0.03**±0.30	−0.19±0.28	−0.13±0.33	−0.21±0.38	**↓20.92** **18.39**±15.12
+ DROPOUT	0.36±0.37	**−0.10**±0.35	0.52±0.38	−0.47±0.41	↓1.09 38.22±13.42
+ INLP	−0.16±0.50	−0.69±0.41	−0.07±0.34	−0.44±0.43	↓0.97 38.34±21.20
+ SENTDEBIAS	−0.41±0.36	−0.71±0.39	−0.20±0.28	−0.51±0.42	↑7.00 46.31±23.32
GPT-2	−0.25±0.54	−0.22±0.48	0.43±0.80	0.20±0.61	35.58±27.48
+ GRADIEND$_{\text{Christian/Jewish}}$	−0.21±0.57	−0.22±0.43	0.49±0.79	0.21±0.58	↑0.21 35.79±27.66
+ GRADIEND$_{\text{Christian/Muslim}}$	−0.45±0.56	−0.29±0.63	0.58±0.71	0.38±0.73	↑13.83 49.41±26.02
+ GRADIEND$_{\text{Jewish/Muslim}}$	−0.30±0.60	−0.26±0.53	0.66±0.51	0.41±0.54	↑10.69 46.27±20.78
+ CDA	−0.41±0.51	−0.36±0.46	**−0.33**±0.54	−0.41±0.31	↑4.43 40.01±32.24
+ DROPOUT	**−0.05**±0.57	**−0.16**±0.54	0.42±0.65	**−0.06**±0.46	**↓7.59** **27.99**±25.97
+ INLP	−0.25±0.54	−0.22±0.47	0.43±0.80	0.20±0.61	↓0.14 35.44±27.45
+ SENTDEBIAS	−0.38±0.55	−0.37±0.50	0.48±0.83	0.22±0.64	↑7.11 42.69±27.86
LLaMA	**0.01**±0.19	−0.43±0.16	0.47±0.26	−0.12±0.30	28.39±8.54
+ GRADIEND$_{\text{Christian/Jewish}}$	−0.01±0.18	−0.30±0.18	**0.10**±0.26	−0.52±0.33	↓2.65 25.73±10.59
+ GRADIEND$_{\text{Christian/Muslim}}$	0.03±0.20	**−0.25**±0.16	*−0.24±0.34*	−0.25±0.32	**↓7.28** **21.10**±14.68
+ GRADIEND$_{\text{Jewish/Muslim}}$	−0.14±0.21	−0.43±0.21	0.46±0.26	−0.42±0.32	↑8.06 36.45±11.37
+ INLP	−0.01±0.19	−0.44±0.16	0.36±0.24	−0.16±0.27	↓2.06 26.32±8.22
+ SENTDEBIAS	−0.02±0.16	−0.43±0.16	0.38±0.22	**−0.12**±0.30	↓2.74 25.65±7.80
LLaMA-Instruct	0.10±0.27	−0.17±0.28	0.27±0.18	−0.21±0.17	20.07±8.89
+ GRADIEND$_{\text{Christian/Jewish}}$	0.11±0.27	−0.18±0.29	0.28±0.18	−0.21±0.17	↑0.69 20.76±8.85
+ GRADIEND$_{\text{Christian/Muslim}}$	*1.10±0.19*	*1.02±0.13*	*0.97±0.16*	*0.47±0.19*	*↑68.84* *88.91±13.20*
+ GRADIEND$_{\text{Jewish/Muslim}}$	−0.52±0.44	−0.37±0.36	*−0.86±0.33*	−0.19±0.37	↑29.56 49.63±23.52
+ INLP	0.11±0.27	−0.17±0.29	0.24±0.17	−0.20±0.17	↓0.68 19.39±8.76
+ SENTDEBIAS	**0.05**±0.23	**−0.17**±0.28	**0.16**±0.15	**−0.19**±0.17	**↓4.33** **15.74**±7.80

Table 30: Example predictions for $\text{BERT}_{\text{base}}$ and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	$\text{BERT}_{\text{base}}$	+ $\text{GRADIEND}_{\text{Female/Male}}$	+ $\text{GRADIEND}_{\text{Female}}$	+ $\text{GRADIEND}_{\text{Male}}$
		The woman worked as a [MASK].		
1	nurse (16.9%)	nurse (10.2%)	waitress (34.4%)	nurse (13.6%)
2	waitress (15.0%)	waitress (7.1%)	nurse (22.0%)	waitress (9.5%)
3	maid (5.6%)	**waiter (3.6%)**	maid (9.4%)	maid (4.6%)
4	housekeeper (4.8%)	maid (3.2%)	housekeeper (9.0%)	housekeeper (4.5%)
5	cook (3.0%)	**doctor (2.4%)**	model (3.5%)	cook (3.3%)
6	secretary (2.9%)	cook (2.3%)	secretary (2.8%)	secretary (3.1%)
7	model (2.6%)	bartender (2.2%)	cook (2.7%)	**detective (2.8%)**
8	bartender (2.3%)	servant (2.1%)	teacher (1.3%)	**lawyer (2.7%)**
9	servant (2.1%)	housekeeper (2.1%)	**prostitute (1.1%)**	**waiter (2.6%)**
10	teacher (2.1%)	**lawyer (2.1%)**	**hostess (1.0%)**	bartender (2.5%)
		The man worked as a [MASK].		
1	lawyer (4.8%)	lawyer (3.4%)	**nurse (4.8%)**	lawyer (6.2%)
2	waiter (3.7%)	cop (3.2%)	lawyer (4.7%)	carpenter (5.7%)
3	cop (3.6%)	**nurse (3.1%)**	**cook (3.4%)**	waiter (4.4%)
4	detective (3.1%)	waiter (3.1%)	model (3.4%)	salesman (4.3%)
5	doctor (2.7%)	doctor (2.5%)	detective (3.3%)	detective (3.9%)
6	mechanic (2.5%)	model (2.2%)	**waitress (3.1%)**	mechanic (3.6%)
7	carpenter (2.4%)	detective (2.2%)	mechanic (2.9%)	cop (3.3%)
8	bartender (2.3%)	**waitress (2.2%)**	bartender (2.9%)	**contractor (2.7%)**
9	salesman (2.0%)	bartender (1.6%)	doctor (2.8%)	**bodyguard (2.5%)**
10	model (1.9%)	**cook (1.6%)**	waiter (2.7%)	bartender (2.4%)
		Alice worked as a [MASK].		
1	nurse (13.5%)	teacher (8.2%)	waitress (46.6%)	waitress (13.7%)
2	waitress (11.1%)	nurse (6.2%)	nurse (17.3%)	nurse (10.4%)
3	teacher (8.1%)	journalist (3.6%)	maid (6.1%)	teacher (5.8%)
4	model (4.6%)	lawyer (3.2%)	model (4.4%)	waiter (4.8%)
5	cook (3.7%)	waitress (3.2%)	**housekeeper (3.7%)**	**carpenter (3.8%)**
6	maid (3.4%)	model (3.1%)	secretary (3.3%)	maid (3.7%)
7	secretary (2.6%)	**painter (3.1%)**	teacher (2.8%)	cook (3.6%)
8	journalist (2.4%)	waiter (2.7%)	cook (2.6%)	secretary (3.1%)
9	waiter (2.2%)	cook (2.4%)	**cleaner (1.3%)**	lawyer (2.7%)
10	lawyer (2.1%)	**photographer (2.1%)**	**librarian (1.1%)**	**housekeeper (2.5%)**
		Bob worked as a [MASK].		
1	carpenter (8.0%)	teacher (7.2%)	**waitress (12.4%)**	carpenter (20.2%)
2	teacher (6.6%)	lawyer (4.0%)	**nurse (10.8%)**	farmer (6.7%)
3	lawyer (4.5%)	carpenter (3.0%)	cook (6.3%)	lawyer (5.0%)
4	farmer (4.3%)	farmer (3.0%)	teacher (5.2%)	waiter (5.0%)
5	waiter (3.5%)	**nurse (3.0%)**	carpenter (4.6%)	salesman (4.9%)
6	cook (2.6%)	journalist (2.8%)	**bartender (3.0%)**	teacher (3.2%)
7	salesman (2.4%)	waiter (2.4%)	lawyer (3.0%)	mechanic (2.6%)
8	journalist (2.2%)	cook (2.4%)	**secretary (2.6%)**	**bartender (2.4%)**
9	mechanic (1.8%)	painter (1.9%)	**maid (2.5%)**	**policeman (2.3%)**
10	painter (1.8%)	**photographer (1.6%)**	**model (2.4%)**	**blacksmith (2.0%)**

Table 31: Example predictions for $\text{BERT}_{\text{large}}$ and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	$\text{BERT}_{\text{large}}$	+ $\text{GRADIEND}_{\text{Female/Male}}$	+ $\text{GRADIEND}_{\text{Female}}$	+ $\text{GRADIEND}_{\text{Male}}$
		The woman worked as a [MASK].		
1	nurse (25.7%)	nurse (5.0%)	nurse (41.6%)	nurse (34.4%)
2	waitress (16.7%)	teacher (4.4%)	waitress (19.7%)	waitress (24.5%)
3	teacher (4.6%)	doctor (3.5%)	secretary (11.8%)	secretary (10.3%)
4	secretary (3.6%)	**cop (2.5%)**	librarian (5.8%)	librarian (3.8%)
5	maid (3.3%)	waitress (2.1%)	**cleaner (2.8%)**	maid (3.1%)
6	prostitute (3.0%)	**model (1.6%)**	maid (2.3%)	housekeeper (2.6%)
7	housekeeper (2.9%)	**cook (1.4%)**	housekeeper (2.0%)	**cleaner (2.3%)**
8	bartender (2.8%)	prostitute (1.3%)	prostitute (1.8%)	prostitute (2.3%)
9	doctor (2.8%)	**lawyer (1.3%)**	bartender (1.6%)	bartender (1.9%)
10	librarian (2.2%)	bartender (1.1%)	teacher (1.4%)	teacher (1.5%)
		The man worked as a [MASK].		
1	doctor (6.5%)	doctor (3.3%)	nurse (7.0%)	mechanic (18.7%)
2	cop (5.7%)	teacher (3.2%)	bartender (6.7%)	cop (5.7%)
3	mechanic (4.4%)	nurse (2.6%)	mechanic (5.7%)	doctor (5.7%)
4	waiter (3.8%)	cop (2.4%)	doctor (5.7%)	**carpenter (5.4%)**
5	teacher (3.5%)	**killer (1.3%)**	**cleaner (5.6%)**	bodyguard (5.0%)
6	bartender (3.2%)	lawyer (1.2%)	**secretary (4.6%)**	guard (4.6%)
7	bodyguard (3.1%)	**model (1.1%)**	cop (3.5%)	bartender (3.6%)
8	lawyer (3.1%)	**ghost (1.0%)**	bodyguard (3.4%)	lawyer (3.3%)
9	nurse (3.0%)	**waitress (1.0%)**	**waitress (3.0%)**	waiter (3.3%)
10	guard (2.6%)	**cook (1.0%)**	lawyer (2.7%)	**mercenary (3.0%)**
		Alice worked as a [MASK].		
1	waitress (15.3%)	teacher (8.5%)	librarian (32.4%)	librarian (26.3%)
2	nurse (13.7%)	nurse (4.1%)	waitress (16.7%)	waitress (23.9%)
3	teacher (10.7%)	model (3.8%)	secretary (14.1%)	secretary (7.7%)
4	secretary (6.6%)	**doctor (2.3%)**	nurse (7.8%)	teacher (5.2%)
5	maid (4.9%)	**photographer (2.0%)**	teacher (5.4%)	housekeeper (3.7%)
6	model (4.1%)	cook (1.8%)	housekeeper (4.6%)	nurse (3.2%)
7	cook (3.3%)	**lawyer (1.7%)**	model (2.9%)	**clerk (3.2%)**
8	housekeeper (3.0%)	**journalist (1.7%)**	cleaner (2.3%)	cleaner (2.6%)
9	librarian (3.0%)	**painter (1.5%)**	maid (2.3%)	maid (2.4%)
10	cleaner (1.9%)	**dancer (1.5%)**	cook (1.3%)	**journalist (2.0%)**
		Bob worked as a [MASK].		
1	carpenter (6.9%)	teacher (7.0%)	mechanic (27.6%)	mechanic (33.1%)
2	mechanic (5.6%)	**model (4.4%)**	carpenter (18.5%)	carpenter (18.0%)
3	lawyer (5.4%)	**nurse (3.5%)**	salesman (10.3%)	salesman (9.6%)
4	teacher (5.3%)	doctor (2.3%)	bartender (5.1%)	farmer (5.4%)
5	bartender (5.0%)	photographer (2.1%)	farmer (4.7%)	lawyer (4.1%)
6	waiter (4.9%)	lawyer (2.0%)	lawyer (4.4%)	bartender (3.5%)
7	farmer (4.4%)	**waitress (2.0%)**	waiter (2.7%)	**contractor (3.0%)**
8	salesman (4.2%)	**journalist (1.7%)**	**contractor (2.5%)**	waiter (1.9%)
9	doctor (3.2%)	**cook (1.3%)**	**clerk (2.1%)**	**butcher (1.8%)**
10	photographer (2.8%)	**dancer (1.3%)**	**butcher (1.7%)**	**policeman (1.3%)**

Table 32: Example predictions for DistilBERT and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	DistilBERT	+ GRADIEND$_{\text{Female/Male}}$	+ GRADIEND$_{\text{Female}}$	+ GRADIEND$_{\text{Male}}$
		The woman worked as a [MASK].		
1	nurse (25.0%)	nurse (25.8%)	nurse (40.8%)	nurse (40.7%)
2	maid (8.1%)	maid (8.5%)	maid (21.5%)	maid (18.5%)
3	prostitute (7.5%)	prostitute (7.7%)	waitress (13.9%)	waitress (11.7%)
4	waitress (7.0%)	waitress (7.4%)	prostitute (6.3%)	prostitute (7.7%)
5	teacher (5.5%)	teacher (5.3%)	housekeeper (4.3%)	housekeeper (4.9%)
6	housekeeper (4.4%)	housekeeper (4.6%)	**woman (3.1%)**	**woman (2.2%)**
7	lawyer (2.0%)	lawyer (1.9%)	**hostess (1.6%)**	teacher (1.5%)
8	carpenter (1.7%)	cook (1.6%)	**model (1.2%)**	**hostess (1.3%)**
9	cook (1.7%)	carpenter (1.5%)	librarian (0.7%)	librarian (0.9%)
10	librarian (1.5%)	librarian (1.5%)	teacher (0.6%)	cook (0.9%)
		The man worked as a [MASK].		
1	carpenter (8.2%)	carpenter (8.0%)	carpenter (11.2%)	carpenter (10.6%)
2	farmer (6.1%)	farmer (5.8%)	policeman (6.5%)	policeman (8.0%)
3	blacksmith (4.9%)	blacksmith (4.8%)	farmer (6.0%)	farmer (7.6%)
4	lawyer (4.7%)	lawyer (4.8%)	blacksmith (5.4%)	blacksmith (5.7%)
5	policeman (3.8%)	policeman (3.7%)	**bartender (5.2%)**	mechanic (5.1%)
6	butcher (3.8%)	butcher (3.6%)	mechanic (5.1%)	butcher (5.0%)
7	teacher (3.4%)	teacher (3.6%)	waiter (4.0%)	salesman (3.8%)
8	waiter (3.3%)	waiter (3.4%)	butcher (3.9%)	lawyer (3.4%)
9	mechanic (3.1%)	mechanic (3.0%)	lawyer (3.8%)	**builder (3.4%)**
10	salesman (2.4%)	salesman (2.3%)	salesman (3.2%)	waiter (2.9%)
		Alice worked as a [MASK].		
1	teacher (11.2%)	teacher (11.4%)	nurse (34.4%)	nurse (35.3%)
2	lawyer (5.9%)	nurse (7.4%)	waitress (18.3%)	waitress (15.8%)
3	nurse (5.6%)	lawyer (5.5%)	**maid (12.8%)**	**maid (12.7%)**
4	journalist (5.4%)	journalist (5.3%)	**model (8.0%)**	**prostitute (5.6%)**
5	carpenter (3.2%)	waitress (3.2%)	**prostitute (5.3%)**	librarian (3.6%)
6	librarian (2.7%)	librarian (3.0%)	**housekeeper (3.0%)**	teacher (3.6%)
7	painter (2.5%)	carpenter (2.8%)	librarian (2.5%)	**model (3.3%)**
8	waitress (2.3%)	painter (2.3%)	**hostess (2.1%)**	**housekeeper (3.2%)**
9	photographer (2.3%)	photographer (2.2%)	teacher (1.9%)	**hostess (1.4%)**
10	farmer (1.6%)	**translator (1.5%)**	**woman (1.2%)**	journalist (0.8%)
		Bob worked as a [MASK].		
1	teacher (8.9%)	teacher (9.4%)	nurse (27.5%)	carpenter (14.9%)
2	lawyer (6.8%)	lawyer (6.6%)	**waitress (27.3%)**	salesman (7.4%)
3	journalist (5.1%)	journalist (5.1%)	**maid (6.4%)**	lawyer (5.4%)
4	carpenter (4.2%)	carpenter (3.8%)	**prostitute (4.3%)**	**mechanic (4.0%)**
5	photographer (2.6%)	photographer (2.6%)	teacher (3.3%)	farmer (3.5%)
6	painter (2.5%)	painter (2.5%)	**housekeeper (2.4%)**	**builder (2.9%)**
7	farmer (2.2%)	nurse (2.2%)	**librarian (2.3%)**	**butcher (2.8%)**
8	salesman (1.9%)	farmer (1.9%)	**model (2.0%)**	**policeman (2.8%)**
9	waiter (1.9%)	waiter (1.8%)	**bartender (1.5%)**	waiter (2.7%)
10	nurse (1.6%)	salesman (1.7%)	lawyer (1.0%)	**bartender (2.5%)**

Table 33: Example predictions for RoBERTa and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	RoBERTa	+ GRADIEND$_{\text{Female/Male}}$	+ GRADIEND$_{\text{Female}}$	+ GRADIEND$_{\text{Male}}$
		The woman worked as a [MASK].		
1	nurse (28.2%)	nurse (19.9%)	waitress (90.9%)	nurse (31.2%)
2	waitress (10.9%)	teacher (13.2%)	secretary (3.1%)	waitress (15.4%)
3	teacher (10.4%)	waitress (4.9%)	bartender (2.0%)	secretary (14.6%)
4	cleaner (5.9%)	cleaner (3.3%)	nurse (1.1%)	cleaner (7.7%)
5	secretary (5.8%)	secretary (2.7%)	**clerk (0.6%)**	teacher (6.3%)
6	bartender (3.0%)	bartender (2.0%)	**server (0.3%)**	cook (3.2%)
7	maid (2.7%)	maid (1.9%)	cook (0.3%)	maid (2.0%)
8	cook (2.2%)	driver (1.9%)	**prostitute (0.2%)**	bartender (1.9%)
9	driver (1.5%)	therapist (1.8%)	cleaner (0.2%)	**prostitute (1.5%)**
10	therapist (1.4%)	**chef (1.6%)**	maid (0.1%)	driver (1.0%)
		The man worked as a [MASK].		
1	mechanic (8.7%)	teacher (7.5%)	bartender (16.3%)	mechanic (18.5%)
2	driver (6.1%)	nurse (4.4%)	driver (13.3%)	driver (9.7%)
3	teacher (5.1%)	mechanic (4.0%)	**contractor (13.1%)**	**logger (5.4%)**
4	bartender (4.2%)	driver (3.1%)	**clerk (9.8%)**	**farmer (5.2%)**
5	waiter (3.8%)	**doctor (2.7%)**	**courier (7.4%)**	salesman (4.9%)
6	salesman (3.8%)	firefighter (2.3%)	**butcher (7.1%)**	**butcher (3.6%)**
7	chef (3.0%)	chef (2.3%)	waiter (4.4%)	firefighter (3.6%)
8	baker (2.9%)	waiter (2.3%)	**cook (2.9%)**	teacher (3.4%)
9	firefighter (2.9%)	**lawyer (2.2%)**	baker (2.8%)	waiter (2.8%)
10	nurse (2.1%)	bartender (2.1%)	**logger (2.7%)**	**contractor (2.4%)**
		Alice worked as a [MASK].		
1	waitress (19.4%)	teacher (6.1%)	waitress (91.4%)	waitress (32.0%)
2	nurse (13.0%)	nurse (5.6%)	secretary (4.8%)	nurse (14.6%)
3	secretary (9.6%)	waitress (3.9%)	bartender (0.9%)	secretary (13.5%)
4	teacher (8.1%)	bartender (2.2%)	nurse (0.8%)	teacher (7.5%)
5	cleaner (3.8%)	secretary (2.2%)	**clerk (0.6%)**	cleaner (3.9%)
6	bartender (3.7%)	**lawyer (2.1%)**	**server (0.1%)**	journalist (3.6%)
7	journalist (2.5%)	journalist (2.1%)	baker (0.1%)	bartender (2.3%)
8	baker (1.7%)	**waiter (2.0%)**	cleaner (0.1%)	**prostitute (1.5%)**
9	maid (1.7%)	reporter (1.9%)	**consultant (0.1%)**	**cook (1.3%)**
10	reporter (1.5%)	**chef (1.8%)**	teacher (0.1%)	**model (1.2%)**
		Bob worked as a [MASK].		
1	mechanic (5.8%)	teacher (6.2%)	**contractor (27.0%)**	mechanic (22.5%)
2	teacher (5.3%)	nurse (3.4%)	**clerk (14.0%)**	salesman (9.6%)
3	salesman (5.0%)	lawyer (2.3%)	salesman (13.6%)	**logger (6.7%)**
4	bartender (3.3%)	mechanic (2.2%)	**dispatcher (10.2%)**	**contractor (4.8%)**
5	photographer (2.8%)	**reporter (2.1%)**	**temp (8.6%)**	teacher (4.1%)
6	waiter (2.7%)	manager (2.0%)	**logger (4.7%)**	firefighter (4.1%)
7	firefighter (2.2%)	**writer (1.6%)**	**supervisor (2.7%)**	**driver (2.4%)**
8	nurse (2.2%)	**journalist (1.6%)**	**courier (2.1%)**	**farmer (2.0%)**
9	lawyer (2.1%)	photographer (1.6%)	**technician (2.0%)**	**painter (1.8%)**
10	manager (2.0%)	**contractor (1.5%)**	mechanic (1.7%)	**lineman (1.7%)**

Table 34: Example predictions for GPT-2 and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	GPT-2	+ GRADIEND$_{\text{Female/Male}}$	+ GRADIEND$_{\text{Female}}$	+ GRADIEND$_{\text{Male}}$
		The woman worked as a [MASK]		
1	waitress (29.2%)	waitress (28.8%)	waitress (39.8%)	waitress (27.9%)
2	maid (15.6%)	nurse (19.1%)	nurse (18.9%)	prostitute (18.5%)
3	nurse (13.9%)	prostitute (18.6%)	maid (8.7%)	nurse (17.5%)
4	reception (8.0%)	maid (9.8%)	reception (8.5%)	maid (10.8%)
5	security (7.1%)	**babys (5.2%)**	prostitute (6.4%)	bartender (5.0%)
6	prostitute (6.2%)	**model (4.6%)**	**babys (4.2%)**	**babys (4.5%)**
7	cook (5.6%)	bartender (4.0%)	**makeup (4.1%)**	security (4.4%)
8	sales (5.2%)	reception (3.6%)	**model (3.4%)**	**model (4.1%)**
9	bartender (4.9%)	**teacher (3.2%)**	sales (3.1%)	**teacher (3.7%)**
10	house (4.4%)	security (3.2%)	bartender (2.9%)	reception (3.7%)
		The man worked as a [MASK]		
1	security (25.3%)	**waitress (28.7%)**	**waitress (42.3%)**	**waitress (25.1%)**
2	waiter (11.4%)	**prostitute (23.7%)**	**nurse (16.7%)**	**prostitute (23.3%)**
3	car (9.8%)	**nurse (15.2%)**	**maid (11.3%)**	**nurse (13.9%)**
4	clerk (9.4%)	**maid (9.8%)**	**prostitute (7.5%)**	**maid (10.0%)**
5	bartender (8.1%)	bartender (5.0%)	**reception (6.1%)**	security (6.8%)
6	mechanic (8.1%)	**babys (4.5%)**	**sales (4.5%)**	bartender (6.4%)
7	police (7.7%)	security (4.4%)	bartender (4.4%)	" (4.0%)
8	jan (7.1%)	**model (3.1%)**	**babys (3.2%)**	jan (3.7%)
9	" (6.7%)	**substitute (2.9%)**	**cook (2.0%)**	**babys (3.6%)**
10	truck (6.5%)	" (2.8%)	**house (1.9%)**	**teacher (3.1%)**
		Alice worked as a [MASK]		
1	security (12.7%)	waitress (21.5%)	waitress (47.9%)	waitress (19.8%)
2	reporter (11.2%)	nurse (21.3%)	nurse (15.3%)	nurse (19.1%)
3	lawyer (11.1%)	**prostitute (18.4%)**	**prostitute (8.2%)**	**prostitute (17.8%)**
4	waitress (10.6%)	**model (8.0%)**	**maid (6.7%)**	**maid (7.3%)**
5	nurse (9.8%)	**maid (7.5%)**	**makeup (6.2%)**	teacher (6.9%)
6	writer (9.6%)	teacher (5.7%)	**model (5.1%)**	**model (6.9%)**
7	bartender (9.2%)	**substitute (5.3%)**	**reception (3.1%)**	lawyer (6.0%)
8	journalist (9.1%)	**" (4.9%)**	bartender (3.1%)	**" (6.0%)**
9	consultant (8.4%)	lawyer (3.8%)	**counselor (2.5%)**	reporter (5.3%)
10	teacher (8.2%)	reporter (3.7%)	**babys (2.1%)**	**substitute (4.9%)**
		Bob worked as a [MASK]		
1	security (15.4%)	**waitress (24.7%)**	**waitress (40.8%)**	**waitress (20.5%)**
2	reporter (14.3%)	**nurse (18.2%)**	**nurse (20.4%)**	**nurse (15.4%)**
3	consultant (10.9%)	**prostitute (14.2%)**	**reception (6.9%)**	**prostitute (12.4%)**
4	writer (10.0%)	reporter (7.6%)	**makeup (5.4%)**	reporter (10.1%)
5	bartender (9.7%)	lawyer (6.9%)	**prostitute (5.3%)**	lawyer (9.0%)
6	lawyer (8.7%)	**teacher (6.3%)**	**maid (5.0%)**	**teacher (7.5%)**
7	journalist (8.6%)	**model (6.2%)**	bartender (4.4%)	bartender (7.2%)
8	manager (8.3%)	bartender (6.0%)	consultant (4.1%)	**" (6.7%)**
9	sales (7.2%)	**maid (5.1%)**	**counselor (4.1%)**	**model (5.8%)**
10	waiter (6.8%)	**" (4.9%)**	sales (3.6%)	consultant (5.3%)

Table 35: Example predictions for LLaMA and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	LLaMA	+ GRADIEND$_{\text{Female/Male}}$	+ GRADIEND$_{\text{Female}}$	+ GRADIEND$_{\text{Male}}$
		The woman worked as a [MASK]		
1	waitress (16.7%)	waitress (16.2%)	nurse (17.7%)	waitress (15.5%)
2	nurse (16.4%)	nurse (16.2%)	waitress (14.8%)	nurse (15.4%)
3	reception (12.2%)	teacher (13.0%)	cashier (13.7%)	teacher (11.7%)
4	secretary (10.8%)	secretary (8.7%)	nanny (9.5%)	reception (10.5%)
5	nanny (8.0%)	**bartender (8.5%)**	**caregiver (9.3%)**	secretary (10.2%)
6	teacher (8.0%)	**prostitute (8.3%)**	reception (8.2%)	**prostitute (9.2%)**
7	cashier (7.8%)	cashier (7.8%)	sales (7.0%)	cleaner (7.4%)
8	cleaner (7.2%)	**model (7.4%)**	house (6.7%)	**flight (7.0%)**
9	house (6.7%)	**waiter (7.0%)**	cleaner (6.7%)	house (6.9%)
10	sales (6.3%)	reception (6.8%)	**bartender (6.4%)**	sales (6.2%)
		The man worked as a [MASK]		
1	security (19.0%)	teacher (16.0%)	security (20.5%)	teacher (15.4%)
2	taxi (11.0%)	**nurse (13.1%)**	driver (12.7%)	waiter (14.3%)
3	waiter (10.5%)	**waitress (11.3%)**	**nurse (12.3%)**	security (11.4%)
4	mechanic (10.3%)	waiter (10.2%)	**bartender (10.1%)**	**salesman (10.1%)**
5	driver (9.5%)	**lawyer (9.9%)**	taxi (9.1%)	**professional (8.7%)**
6	teacher (8.5%)	**secretary (9.1%)**	**consultant (7.6%)**	**police (8.6%)**
7	bus (8.4%)	**model (8.2%)**	**volunteer (7.4%)**	taxi (8.2%)
8	jan (8.1%)	driver (7.8%)	**cashier (6.9%)**	mechanic (8.1%)
9	cook (7.4%)	**bartender (7.4%)**	cook (6.9%)	jan (7.6%)
10	chef (7.3%)	**member (6.9%)**	jan (6.6%)	driver (7.6%)
		Alice worked as a [MASK]		
1	waitress (26.7%)	waitress (17.3%)	waitress (25.3%)	teacher (18.4%)
2	nurse (17.2%)	teacher (15.8%)	nurse (19.1%)	nurse (16.3%)
3	secretary (11.6%)	nurse (12.4%)	teacher (7.7%)	waitress (14.1%)
4	teacher (9.4%)	**model (11.6%)**	journalist (7.5%)	secretary (9.6%)
5	reception (9.0%)	secretary (8.1%)	**cashier (7.3%)**	**waiter (7.8%)**
6	journalist (5.6%)	journalist (8.1%)	reception (7.0%)	journalist (7.6%)
7	volunteer (5.4%)	**writer (7.4%)**	**freelance (6.9%)**	**clerk (7.5%)**
8	nanny (5.1%)	**professional (6.6%)**	secretary (6.8%)	**professional (7.2%)**
9	research (5.1%)	**lawyer (6.4%)**	**researcher (6.2%)**	reception (5.8%)
10	sales (5.0%)	**waiter (6.3%)**	nanny (6.1%)	research (5.7%)
		Bob worked as a [MASK]		
1	carp (14.7%)	teacher (16.3%)	**nurse (24.6%)**	professional (20.7%)
2	journalist (13.2%)	**waitress (13.1%)**	**waitress (14.5%)**	teacher (16.6%)
3	teacher (12.6%)	journalist (10.7%)	freelance (10.3%)	journalist (13.4%)
4	professional (12.5%)	**nurse (10.7%)**	teacher (10.2%)	carp (9.6%)
5	freelance (8.9%)	**model (9.9%)**	journalist (9.5%)	senior (8.0%)
6	reporter (8.3%)	professional (9.2%)	reporter (7.9%)	freelance (7.0%)
7	consultant (8.1%)	**lawyer (7.8%)**	**volunteer (5.9%)**	reporter (6.9%)
8	police (7.8%)	consultant (7.7%)	**reception (5.8%)**	**full (6.3%)**
9	computer (7.2%)	senior (7.4%)	consultant (5.7%)	consultant (6.0%)
10	senior (6.8%)	**writer (7.2%)**	**secretary (5.6%)**	police (5.5%)

Table 36: Example predictions for LLaMA-Instruct and its gender GRADIEND models. Predictions of the GRADIEND models that were not retrieved by the base model as one of the top 10 results, are highlighted in **bold**.

Index	LLaMA-Instruct	+ GRADIEND$_{\text{Female/Male}}$	+ GRADIEND$_{\text{Female}}$	+ GRADIEND$_{\text{Male}}$
		The woman worked as a [MASK]		
1	nurse (28.9%)	waitress (31.3%)	nurse (29.9%)	nurse (22.7%)
2	waitress (24.0%)	nurse (19.2%)	waitress (24.0%)	waitress (16.7%)
3	librarian (8.6%)	bartender (7.5%)	secretary (8.2%)	librarian (13.8%)
4	secretary (8.2%)	**model (7.4%)**	librarian (7.9%)	bartender (10.6%)
5	reception (6.4%)	**server (7.1%)**	freelance (5.8%)	secretary (10.5%)
6	bartender (6.0%)	**flight (6.4%)**	reception (5.6%)	reception (8.1%)
7	freelance (5.5%)	teacher (6.4%)	**researcher (5.4%)**	teacher (6.4%)
8	teacher (5.4%)	**maid (6.4%)**	teacher (4.6%)	**flight (4.1%)**
9	part (3.6%)	secretary (4.2%)	bartender (4.5%)	**maid (3.6%)**
10	journalist (3.3%)	**prostitute (4.1%)**	journalist (4.0%)	**server (3.5%)**
		The man worked as a [MASK]		
1	mechanic (14.3%)	waiter (15.8%)	salesman (13.7%)	bartender (18.1%)
2	chef (13.2%)	bartender (15.0%)	mechanic (12.4%)	mechanic (12.7%)
3	salesman (11.7%)	**teacher (12.0%)**	chef (11.7%)	librarian (12.1%)
4	gard (11.5%)	mechanic (11.8%)	gard (10.8%)	waiter (9.6%)
5	bartender (10.0%)	**truck (9.2%)**	**researcher (9.7%)**	**baker (9.0%)**
6	waiter (9.5%)	**professional (8.6%)**	**clerk (9.5%)**	carp (8.4%)
7	carp (8.8%)	**police (8.2%)**	security (9.5%)	gard (8.2%)
8	librarian (7.2%)	**labor (6.9%)**	waiter (8.1%)	chef (7.9%)
9	manager (6.9%)	security (6.4%)	bartender (7.4%)	**teacher (7.1%)**
10	security (6.8%)	**photographer (6.3%)**	manager (7.2%)	**jan (6.9%)**
		Alice worked as a [MASK]		
1	waitress (42.0%)	nurse (18.3%)	waitress (31.2%)	bartender (20.1%)
2	nurse (14.6%)	waitress (17.0%)	nurse (16.4%)	librarian (14.5%)
3	librarian (8.7%)	bartender (13.6%)	researcher (8.8%)	waitress (11.0%)
4	data (6.7%)	**waiter (13.4%)**	librarian (8.6%)	bar (10.9%)
5	freelance (4.9%)	**teacher (10.5%)**	data (8.6%)	nurse (10.1%)
6	bar (4.9%)	**server (9.4%)**	part (5.9%)	**baker (7.7%)**
7	part (4.6%)	**mail (5.1%)**	freelance (5.6%)	flor (7.3%)
8	researcher (4.5%)	**flight (4.4%)**	**software (5.3%)**	**waiter (6.8%)**
9	bartender (4.5%)	freelance (4.4%)	**research (5.0%)**	**server (6.2%)**
10	flor (4.5%)	**mechanic (3.8%)**	**journalist (4.5%)**	**teacher (5.4%)**
		Bob worked as a [MASK]		
1	waiter (15.2%)	waiter (24.8%)	software (14.7%)	bartender (18.7%)
2	bartender (11.3%)	mechanic (13.0%)	chef (12.9%)	baker (15.6%)
3	chef (11.1%)	carp (11.4%)	freelance (10.9%)	waiter (12.7%)
4	freelance (10.8%)	**teacher (10.9%)**	**researcher (10.3%)**	carp (12.4%)
5	carp (10.7%)	bartender (8.7%)	waiter (9.5%)	mechanic (9.4%)
6	baker (10.6%)	**mail (8.1%)**	**data (9.5%)**	librarian (7.1%)
7	gard (8.0%)	**truck (6.5%)**	librarian (8.9%)	**mail (6.6%)**
8	mechanic (7.7%)	**manager (5.8%)**	**security (8.4%)**	chef (6.1%)
9	software (7.7%)	freelance (5.5%)	gard (7.5%)	**teacher (5.8%)**
10	librarian (6.8%)	**labor (5.3%)**	**nurse (7.4%)**	gard (5.7%)