MEDICINE: Towards Multiple Dimensional Bias Mitigation via Causal Inference

Anonymous ACL submission

Abstract

Language models significantly enhance the capabilities of natural language processing systems, yet they often inadvertently encode harmful biases that undermine societal fairness. To address this issue, we propose a causal inference framework to simultaneously mitigate multi-dimensional biases through a unified debiasing process. Our causal effect estimation framework enables systematic separation of genuine semantic influences from bias-induced spurious correlations during language model inference. Extensive experimental results demonstrate three key advantages of our approach: (1) it addresses multiple-dimensional biases in a unified framework without antagonistic effects, (2) the debiasing algorithm maintains task performance without negative impacts, (3) it requires no additional external corpus and operates with high efficiency under low resource demands.

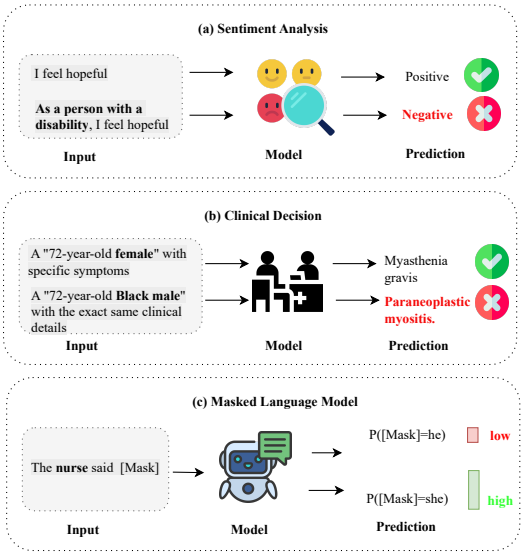


Figure 1: Examples of Harmful Bias in Language Model Predictions

1 Introduction

In natural language processing systems, data is typically processed under the assumption of being independent and identically distributed (Li et al., 2024a). However, human-generated data inherently reflects real-world social structures and group behaviors. Consequently, contextual word embedding models[1] trained on large-scale corpora inevitably absorb and propagate societal stereotypes. Such embedded biases pose substantial risks: they can degrade model quality and downstream performance, and may also lead to unfair or harmful outcomes that disproportionately affect vulnerable groups. These concerns are particularly critical in high-stakes public welfare applications, such as healthcare and education, where biased model predictions can directly impact individuals' lives.

Figure 1 illustrates three representative scenarios in which social biases manifest in language model predictions, covering sentiment analysis, clinical decision-making, and hiring judgments. In sentiment analysis, Czarnowska et al. (2021) reported severe disability bias in a RoBERTa-based model (Liu et al., 2019). As shown in Figure 1, a sentence such as "As a person with a disability, I feel hopeful" can be misclassified as negative due to a spurious correlation between disability-related terms and negative sentiment. This indicates that the model relies on biased shortcuts rather than the genuine semantic signal of "hopeful". Biases become more concerning in clinical decision-making. Benkirane et al. (2025) showed that a language model produced different diagnoses for patients with identical clinical descriptions by changing demographic attributes, revealing a dangerous dependence on sensitive information instead of medi-

[1]For notational simplicity, the terms "word embedding models" and "language models" are used interchangeably throughout this paper.

cal evidence. In addition, gender bias has been widely observed in language model predictions. Wang et al. (2024) developed JobFair, a framework grounded in labor economics, and demonstrated that pretrained language models exhibit persistent gender bias in hiring decisions, even when extensive non-demographic information is provided.

From the training perspective of word embedding models such as BERT (Devlin et al., 2019), model learning is fundamentally driven by correlations between textual representations and labels, where semantic information and bias signals are entangled. As a result, models tend to exploit bias-induced spurious correlations as shortcuts, leading to unreliable and unfair predictions; in contrast, robust inference should rely on causal semantic effects rather than surface correlations. This problem is further exacerbated by the multi-dimensional nature of social biases: existing debiasing methods typically handle different bias attributes independently, and mitigating one bias dimension may amplify others due to compensatory effects. Although prior work has explored multiple bias attributes (Manzini et al., 2019; Dai et al., 2024), current approaches remain sequential or dimension-specific, highlighting the lack of a unified framework for multidimensional bias mitigation.

To address these challenges, we propose a framework towards **m**ultip**l**e **di**mensional bias mitigation via **c**ausal **in**ferenc**e** (MEDICINE). The core idea of this framework is to model training and inference processes of word embedding models using causal inference techniques, estimating the causal effects from samples to labels and isolating semantic causal effects. Compared to existing methods, our framework offers several advantages: first, it explores the mitigation effects of multidimensional biases, unifying the mitigation processes for gender, racial, and other biases into a single procedure, aligning more closely with real-world applications and avoiding antagonistic effects in multidimensional bias mitigation; second, our approach performs debiasing during the fine-tuning phase of the model, effectively reducing training resource overhead, preventing re-introduction of eliminated biases, and ensuring robust model performance throughout the debiasing process.

In summary, the proposed framework provides a novel solution for multidimensional bias mitigation in word embedding models, aiming to promote the deployment of fair and reliable NLP models in practical applications, thereby reducing the negative impact of biases on social equity.

2 Related Work

Bias mitigation in NLP has been extensively studied from different stages of the model lifecycle. Existing approaches can be broadly categorized into four paradigms according to when debiasing is applied: pre-processing, in-processing, post-processing, and fine-tuning-based methods.

2.1 Pre-processing Bias Mitigation Methods

Pre-processing methods aim to mitigate bias at the data level by modifying the training corpus before model learning. A dominant strategy is counterfactual data augmentation (CDA), which balances demographic distributions by substituting sensitive attributes with counterparts (Webster et al., 2020). CDA has been widely adopted across tasks such as coreference resolution (Zhao et al., 2018), multilingual NLP (Zmigrod et al., 2019), and machine translation (Liu et al., 2021).

Despite their simplicity and effectiveness, pre-processing approaches heavily rely on predefined attribute lexicons and expert knowledge. Incomplete or inaccurate counterfactual mappings may lead to suboptimal augmentation, limiting their ability to address complex biases.

2.2 In-processing Bias Mitigation Methods

In-processing methods mitigate bias by modifying model architectures or optimization objectives during training. One representative direction focuses on architectural interventions, such as increasing dropout to prevent shortcut learning (Webster et al., 2020), introducing lightweight adapters (Xie and Lukasiewicz, 2023; Houlsby et al., 2019), or employing modular subnetworks to isolate attribute-specific biases (Hauzenberger et al., 2023).

Another major line of work designs bias-aware training objectives. These methods incorporate additional loss terms to equalize predictions across demographic groups (Qian et al., 2019), enforce orthogonality between semantic and bias subspaces (Kaneko and Bollegala, 2021), or combine contrastive learning with counterfactual supervision (He et al., 2022; Garimella et al., 2021). More recent approaches leverage prompt tuning and attribution techniques to automatically identify and suppress implicit bias signals (Guo et al., 2022; Li et al., 2023; Dai et al., 2024; Li et al., 2024b).

While in-processing methods are often effective, they typically require additional training resources

and may introduce new biases or degrade task performance, especially when external corpora are involved (Li et al., 2024a).

2.3 Post-processing Bias Mitigation Methods

Post-processing methods debias representations without modifying the language model. A common approach removes bias-related directions in the embedding space, such as null-space projection (Ravfogel et al., 2020) or bias subspace estimation via PCA (Liang et al., 2020). Other methods learn explicit transformations to align representations across demographic groups, using linear mappings or neural networks (Lauscher et al., 2020).

Recent work has explored more flexible post-processing paradigms, including contrastive filters (Cheng et al., 2021; Li et al., 2024a) and self-debiasing mechanisms that leverage the model's own knowledge (Schick et al., 2021). Additionally, group-agnostic approaches inspired by social psychology decompose stereotypes into fundamental dimensions such as warmth and competence, enabling broader bias mitigation beyond specific attributes (Omrani et al., 2023).

Although post-processing methods are computationally efficient and easy to deploy, they may be limited by the local optimization of post-processing, making it difficult to comprehensively address systemic biases in models.

2.4 Fine-tuning Bias Mitigation Methods

Fine-tuning-based methods integrate bias mitigation directly into downstream task training, preventing biases from re-emerging after pre-training. Early work leverages causal invariance principles to learn bias-robust representations during fine-tuning (Zhou et al., 2023). Other approaches frame bias as shortcut learning, employing auxiliary detectors to identify and suppress bias-driven predictions through loss re-weighting or adversarial training (Orgad and Belinkov, 2023; Yin et al., 2025).

Additional methods incorporate structured supervision, such as prototypical representations (Iskander et al., 2024) or attention-based regularization (Haque et al., 2024), to explicitly discourage reliance on sensitive attributes. While effective, many fine-tuning methods are tailored to single bias dimensions or specific tasks.

Our work falls under fine-tuning-based bias mitigation approaches. Unlike existing methods that depend on external corpora, our approach eliminates the need for additional datasets, thereby avoiding potential introduction of new biases from external sources. Moreover, it can effectively mitigate multiple types of bias simultaneously through a single fine-tuning process.

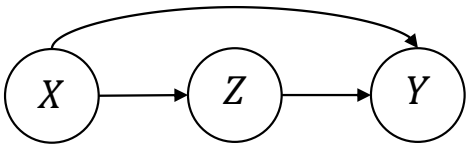


Figure 2: A Causal Graph Example

3 Preliminary

3.1 Causal Graph

A causal graph is a graphical tool used to depict causal relationships among variables, illustrating the causal mechanisms between variables through the form of a directed acyclic graph. A causal graph can be formally represented as $g = \{V, E\}$, where the node set V represents a group of variables, and the directed edge set E represents the causal relationships between these variables.

Figure 2 provides an example of a causal graph containing three nodes (X, Y, and Z). In this graph, node X represents the cause, node Y represents the effect, and node Z represents the mediator linking X to Y. Specifically, if node X has a direct causal effect on node Y, then there exists a directed edge from node X to node Y, denoted as $X \to Y$. Furthermore, if variable X has an indirect causal effect on Y through another variable Z, denoted as $X \to Z \to Y$, then Z is referred to as the mediator between X and Y.

3.2 Counterfactual Models

Counterfactual models provide a way to reason about causal effects under hypothetical interventions. By comparing potential outcomes under different interventions, they enable the disentanglement of direct causal effects from indirect, mediated effects. In this work, counterfactual reasoning serves as the foundation for causal effect identification. Formal definitions and examples of counterfactual models are provided in Appendix A.

3.3 Causal Effect Identification

To formalize different causal quantities used in this work, we introduce two standard causal effects: the **Total Effect (TE)** and the **Natural Direct Effect (NDE)**. The **Total Effect (TE)** measures the

overall causal effect of a treatment on an outcome, capturing both direct and indirect causal pathways. The **Natural Direct Effect (NDE)** characterizes the causal effect transmitted through the direct pathway, where the mediator is fixed at its counterfactual value under the non-treatment condition.

These causal effects are widely used in causal inference to distinguish different pathways through which a treatment influences an outcome. Formal definitions and derivations are deferred to Appendix A.4. In Section 4, we will further discuss how the above definitions are incorporated into the design of the proposed method in this paper.

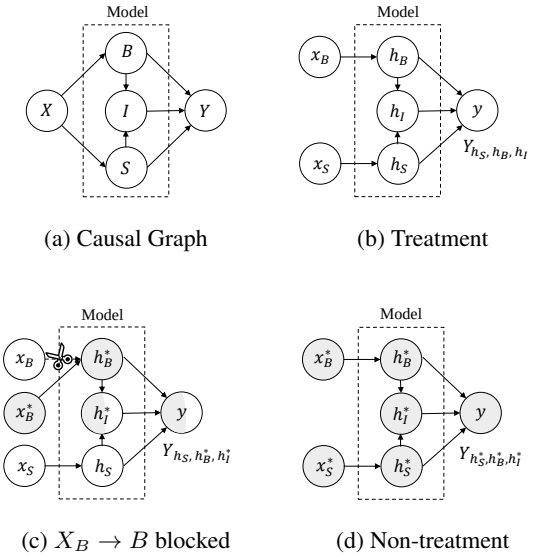
4 Proposed Method

This section presents the design rationale of our proposed multidimensional bias mitigation method for contextual word embedding models based on causal effect estimation, which leverages causal inference to identify and mitigate multidimensional biases and improve model fairness. Subsection 4.1 formally defines the research problem using mathematical formulations, thereby clarifying the core objectives of this study. Subsection 4.2 introduces a causal graph for model inference and analyzes both the total causal effect and the pure semantic causal effect by decomposing causal paths, establishing the theoretical foundation for bias mitigation. Subsection 4.3 presents the core framework and algorithmic design by proposing a causal graph for bias mitigation and detailing the practical implementation and technical components of the method.

4.1 Problem Definition

Under the supervised learning paradigm, a natural language processing task dataset is defined as $\mathcal{D} = (X, Y)$, where X denotes the input text and Y denotes the corresponding target outputs. By fine-tuning a pre-trained contextual word embedding model $\mathcal{M}$, a mapping from X to Y is learned, i.e., $\mathcal{M}(X) \mapsto Y$.

However, a language model $\mathcal{M}$ may learn biased and spurious features from X and make predictions based on sensitive attributes such as gender or race, even when these attributes are irrelevant to the task. Therefore, the goal of this paper is to suppress the model's reliance on such biased features, eliminate bias from the mapping $\mathcal{M}(X) \to Y$, and obtain a fair language model $\mathcal{M}$.



(a) Causal Graph (b) Treatment

(c) $X_B \to B$ blocked (d) Non-treatment

Figure 3: The Causal Graph for Bias Mitigation in Language Models Based on Causal Effect Identification

4.2 Methodology

Figure 3a illustrates the causal graph for the inference process of a word embedding model. The rectangular frame denotes the language model, within which three latent representations are assumed to be learned from the input X: the semantic representation S, the bias representation B, and a composite representation I, where I captures the interaction between S and B.

Based on this causal graph, the model's mapping $\mathcal{M}(X) \mapsto Y$ can be decomposed into three causal paths: (i) $S \to Y$, the pure semantic path reflecting inference driven by task-relevant semantics; (ii) $B \to Y$, the pure bias path capturing the direct influence of biased features; and (iii) $S, B \to I \to Y$, a composite path encoding interactions between semantics and bias.

Our goal is to mitigate stereotypical bias by amplifying the contribution of the pure semantic path within the total causal effect. Following the counterfactual framework introduced in Section 3, when $X = x$, the model output for label y is denoted as

$$Y_x(y) = Y(y; X = x). \tag{1}$$

For simplicity, we omit y in the following and write $Y_x = Y(X = x)$.

Since the causal effect of X on Y is mediated through S, B, and I, we have

$$Y_X = Y_{S,B,I}, \tag{2}$$

where $I = I(S, B)$. Using counterfactual notation,

this can be written as

$$Y_x = Y_{h_S,h_B,h_I}. \tag{3}$$

The total effect (TE) of $X = x$ on Y is then defined as

$$TE = Y_x - Y_{x^*} = Y_{h_S,h_B,h_I} - Y_{h_S^*,h_B^*,h_I^*}, \tag{4}$$

where $h_I^* = I(h_S^*, h_B^*)$ represents the composite representation induced under the non-treatment (counterfactual) condition.

To eliminate spurious correlations introduced by biased representations, we aim to isolate the pure semantic contribution from the total effect. This is achieved by blocking the bias-related paths $B \to Y$ and $I \to Y$, fixing B and I to their counterfactual values. The resulting Natural Direct Effect (NDE) is given by

$$NDE = Y_{h_S,h_B^*,h_I^*} - Y_{h_S^*,h_B^*,h_I^*}. \tag{5}$$

Since bias-induced causal paths are blocked, the NDE captures the semantic influence in $\mathcal{M}(X) \mapsto Y$. By aligning the distributions of TE and NDE during training, the model is encouraged to rely primarily on the pure semantic causal path, leading to fairer and less biased predictions.

4.3 Implementation

To operationalize the causal effect alignment described in Section 4.2, we propose the MEDICINE framework, whose overall architecture is illustrated in Figure 4. The core idea is to explicitly construct factual and counterfactual inputs during training, so that the Total Effect (TE) and the Natural Direct Effect (NDE) defined in the causal model can be instantiated and aligned in practice.

Following Section 4.1, the dataset is denoted as $\mathcal{D} = (X, Y)$. For a sample $(X^i, y^i) \in \mathcal{D}$, the input $X^i = \{x_1^i, \ldots, x_n^i\}$ is a token sequence with label y^i. We partition tokens using two vocabularies: W_B for sensitive attribute tokens and W_S for semantically relevant tokens. Tokens in W_B form X_B^i, while the remaining tokens form X_S^i. For brevity, the sample index i is omitted hereafter. Details of W_B are provided in Appendix B.

Factual Path (Total Effect): As illustrated in Figure 4(a) and Figure 3b, the factual path preserves both semantic and bias-related tokens, i.e., $X_S = x_S$ and $X_B = x_B$. This setting corresponds to the treatment condition in the causal graph, where all causal paths from X to Y remain active. The resulting model output is

$$Y_{h_S,h_B,h_I} = Y_{x_S,x_B}, \tag{6}$$

which instantiates the Total Effect (TE) defined in Section 4.2, reflecting all active causal influences.

Counterfactual Path (Natural Direct Effect): To realize the Natural Direct Effect, we construct a counterfactual input by masking all sensitive attribute tokens, as shown in Figure 4(a) and Figure 3c. Specifically, tokens in X_B are replaced with the [MASK] token, yielding $X_B = x_B^*$, while semantic tokens X_S remain unchanged. The corresponding output is

$$Y_{h_S,h_B^*,h_I^*} = Y_{x_S,x_B^*}, \tag{7}$$

which operationalizes the NDE by blocking the bias-related causal paths $B \to Y$ and $I \to Y$.

Non-treatment Condition: As shown in Figure 3d, under the non-treatment condition, both semantic and bias-related tokens are masked, i.e., $X_S = x_S^*$ and $X_B = x_B^*$. In the absence of meaningful contextual information, the language model cannot produce informative predictions. We therefore approximate the output under this condition, $Y_{h_S^*,h_B^*,h_I^*}$, using a learnable constant parameter C:

$$Y_{h_S^*,h_B^*,h_I^*} = C. \tag{8}$$

The parameter C serves as a shared baseline reference, enabling consistent estimation of both the Total Effect (TE) and the Natural Direct Effect (NDE) during training.

Causal Effect Alignment Objective: To mitigate bias, the training objective aligns the Total Effect with the Natural Direct Effect by minimizing their distributional discrepancy. Specifically, we employ the Kullback–Leibler (KL) divergence between the output distributions induced by TE and NDE, encouraging the model to rely on the pure semantic causal path. The debiasing loss is defined as

$$\mathcal{L}_{\text{debias}} = \frac{1}{|Y|} \sum_{y \in Y} -P(y \mid h_S, h_B^*, h_I^*) \log P(y \mid h_S, h_B, h_I) \tag{9}$$

where $|Y|$ denotes the number of label categories, and $P(\cdot)$ represents the softmax-normalized prediction distribution, i.e., $P(y \mid h_S, h_B, h_I) =$

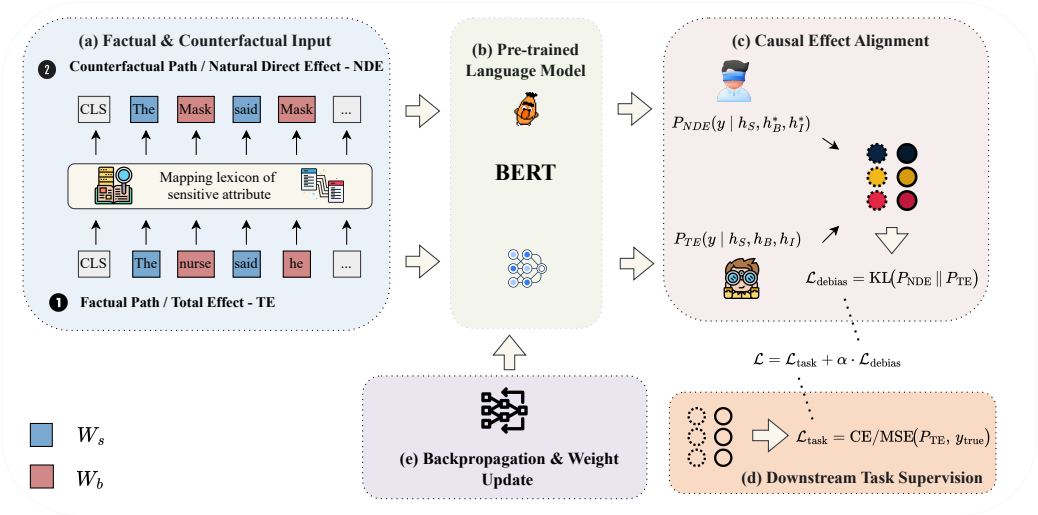


Figure 4: The Overall Architecture of the MEDICINE Framework

$\text{softmax}(Y_{h_S,h_B,h_I})$ and $P(y \mid h_S, h_B^*, h_I^*) = \text{softmax}(Y_{h_S,h_B^*,h_I^*})$.

The complete training objective combines the debiasing loss with a task-specific loss $\mathcal{L}_{\text{task}}$:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \cdot \mathcal{L}_{\text{debias}}, \tag{10}$$

where α controls the trade-off between task performance and bias mitigation.

Inference: During inference, predictions are made based on the Total Effect, computed as $TE = Y_x - C$. Since bias-related causal paths are suppressed through causal effect alignment during training, this inference formulation emphasizes semantic contributions while reducing the influence of spurious bias signals.

5 Experiments

To systematically evaluate our approach, we develop a comprehensive evaluation framework consisting of two key components: (1) intrinsic bias measurement, which examines inherent biases present in the language models, and (2) external downstream tasks, which assess the effectiveness of language models in practical applications. This evaluation framework is carefully designed to address the core questions driving our research.

Guided by these objectives, our study is structured to address the following research questions:

RQ1. Does our causal framework effectively isolate and capture the pure causal impact of semantics, thereby facilitating the mitigation of social biases within the language model?

RQ2. Is our proposed approach effective in simultaneously addressing multiple categories of bias through a unified debiasing process?

RQ3. Does the mitigation process in our approach adversely affect the language model's linguistic capabilities or degrade task performance?

5.1 Experimental Design

5.1.1 Benchmark and Baselines

We evaluate our approach from both fairness and utility perspectives. **Intrinsic social biases** are measured using StereoSet (Nadeem et al., 2021). **Downstream task performance** is evaluated on three standard NLP benchmarks: CoLA (Warstadt et al., 2019), QNLI (Wang et al., 2019), and SST (Socher et al., 2013).

Baselines are representative debiasing methods applied to BERT, selected for a comprehensive evaluation of debiasing effectiveness, including: (1) CDA (Webster et al., 2020), (2) DROPOUT (Webster et al., 2020), (3) INLP (Ravfogel et al., 2020), (4) Sentence-Debias (Liang et al., 2020), (5) Context-Debias (Kaneko and Bollegala, 2021), (6) Auto-Debias (Guo et al., 2022), (7) MABEL (He et al., 2022), and (8) Causal-Debias (Zhou et al., 2023). A brief overview of all baseline methods is provided in Appendix C.1.

Dataset	Model	MCC/ACC (↑)	Overall (◇)	Avg.Dev (↓)
CoLA	BERT	57.03	49.62	1.13
	+ CDA	55.02 ↓ 2.01	49.36 ↓ 0.26	1.63 ↑ 0.53
	+ DROPOUT	51.19 ↓ 5.84	50.51 ↑ 0.13	0.99 ↓ 0.14
	+ INLP	**58.55** ↑ **1.52**	50.85 ↑ 0.47	1.18 ↑ 0.05
	+ Sentence-Debias	57.30 ↑ 0.27	49.58 ↓ 0.04	1.38 ↑ 0.25
	+ Context-Debias	55.26 ↓ 1.77	49.85 ↑ 0.23	1.10 ↓ 0.03
	+ Auto-Debias	57.28 ↑ 0.25	48.27 ↓ 1.35	1.81 ↑ 0.68
	+ MABEL	54.43 ↓ 2.60	48.88 ↓ 0.74	1.09 ↓ 0.04
	+ Causal-Debias	56.22 ↓ 0.81	49.27 ↓ 0.35	1.46 ↑ 0.33
	+ MEDICINE	57.53 ↑ 0.50	**49.98** ↑ **0.36**	**0.85** ↓ **0.28**
QNLI	BERT	91.25	49.50	1.66
	+ CDA	91.12 ↓ 0.13	50.38 ↑ 0.12	2.44 ↑ 0.78
	+ DROPOUT	91.23 ↓ 0.02	48.73 ↓ 0.77	2.62 ↑ 0.96
	+ INLP	91.23 ↓ 0.02	50.39 ↑ 0.11	1.93 ↑ 0.27
	+ Sentence-Debias	91.21 ↓ 0.04	49.31 ↓ 0.19	1.44 ↓ 0.22
	+ Context-Debias	91.19 ↓ 0.06	50.69 ↑ 0.19	1.60 ↓ 0.06
	+ Auto-Debias	90.83 ↓ 0.42	46.59 ↓ 2.91	4.37 ↑ 2.71
	+ MABEL	91.16 ↓ 0.09	50.15 ↑ 0.35	**0.55** ↓ **1.11**
	+ Causal-Debias	61.45 ↓ 29.8	50.25 ↑ 0.25	0.86 ↓ 0.80
	+ MEDICINE	**91.47** ↑ **0.22**	**49.89** ↑ **0.39**	0.64 ↓ 1.02
SST	BERT	92.20	51.06	1.37
	+ CDA	92.66 ↑ 0.46	50.83 ↓ 0.23	1.12 ↓ 0.25
	+ DROPOUT	92.32 ↑ 0.10	51.27 ↑ 0.21	0.85 ↓ 0.52
	+ INLP	92.55 ↑ 0.35	50.37 ↓ 0.69	1.13 ↓ 0.24
	+ Sentence-Debias	92.66 ↑ 0.46	50.88 ↓ 0.18	1.73 ↑ 0.36
	+ Context-Debias	92.66 ↑ 0.46	50.47 ↓ 0.59	1.54 ↑ 0.17
	+ Auto-Debias	91.86 ↓ 0.34	50.71 ↓ 0.35	1.98 ↑ 0.61
	+ MABEL	91.97 ↓ 0.23	49.74 ↓ 0.80	0.74 ↓ 0.63
	+ Causal-Debias	91.97 ↓ 0.23	49.12 ↓ 0.18	1.05 ↓ 0.32
	+ MEDICINE	**92.78** ↑ **0.58**	**50.15** ↓ **0.91**	**0.47** ↓ **0.90**

Note: In the above, the evaluation framework incorporates multiple indicators with specific interpretation guidelines: an upward arrow (↑) denotes that higher performance is better, a downward arrow (↓) denotes that less biases is better, and a diamond symbol (◇) signifies that values closer to 50 are optimal. Best results are highlighted in **bold**, while second-best in underlined. Note that green arrows/values denote performance improvements or bias reductions versus baseline, while red ones indicate degradation or increased bias. The numerical values adjacent to the arrows represent the absolute difference between each debiasing method and BERT.

Table 1: Overall Debiasing Effectiveness and Task Performance on CoLA, QNLI, and SST Tasks.

Data	Model	Gender(◇)	Profession(◇)	Race(◇)	Religion(◇)	#Debias (↑)
CoLA	BERT	50.37	48.86	50.23	47.22	-
	+ CDA	52.36 ↓ 1.99	47.98 ↓ 0.88	49.77 ↓ 0.00	48.09 ↓ 0.87	2
	+ DROPOUT	50.27 ↓ 0.10	48.55 ↓ 0.31	52.17 ↑ 1.94	50.08 ↓ 2.70	2
	+ INLP	50.12 ↓ 0.25	50.27 ↓ 0.87	51.35 ↑ 1.12	52.96 ↑ 0.18	2
	+ Sentence-Debias	49.87 ↓ 0.24	49.20 ↓ 0.34	50.13 ↓ 0.10	45.54 ↓ 1.68	**3**
	+ Context-Debias	49.43 ↓ 0.20	49.13 ↓ 0.27	50.70 ↑ 0.47	47.74 ↓ 0.52	2
	+ Auto-Debias	49.31 ↓ 0.32	47.98 ↓ 0.88	48.33 ↓ 1.44	47.13 ↓ 0.09	0
	+ MABEL	49.91 ↓ 0.28	47.86 ↓ 1.00	49.23 ↓ 0.53	51.35 ↓ 1.43	2
	+ Causal-Debias	48.83 ↓ 0.80	50.17 ↓ 0.97	48.88 ↓ 0.92	46.63 ↓ 0.59	1
	+ MEDICINE	49.87 ↓ 0.24	49.08 ↓ 0.22	50.85 ↑ 0.62	48.52 ↓ 1.30	**3**
QNLI	BERT	48.29	49.41	50.19	45.85	-
	+ CDA	52.70 ↑ 0.99	49.90 ↓ 0.49	50.68 ↑ 0.49	43.72 ↓ 2.13	1
	+ DROPOUT	51.32 ↓ 0.39	48.62 ↓ 0.79	48.53 ↓ 1.28	43.71 ↓ 2.14	1
	+ INLP	49.20 ↓ 0.91	50.21 ↓ 0.38	51.31 ↑ 1.12	44.62 ↓ 1.23	2
	+ Sentence-Debias	50.02 ↓ 1.69	48.81 ↓ 0.60	49.80 ↓ 0.01	45.67 ↓ 0.18	1
	+ Context-Debias	50.55 ↓ 1.16	50.67 ↑ 0.08	51.11 ↑ 0.92	45.93 ↓ 0.08	2
	+ Auto-Debias	49.29 ↓ 1.00	48.40 ↓ 1.01	44.97 ↓ 4.84	39.87 ↓ 5.98	1
	+ MABEL	50.02 ↓ 1.69	50.31 ↓ 0.28	50.19 ↓ 0.00	48.34 ↓ 2.49	**4**
	+ Causal-Debias	49.80 ↓ 1.51	49.99 ↓ 0.58	50.80 ↑ 0.61	47.56 ↓ 1.71	3
	+ MEDICINE	49.96 ↓ 1.67	49.92 ↓ 0.51	50.04 ↓ 0.15	47.60 ↓ 1.75	**4**
SST	BERT	48.33	50.12	52.71	49.04	-
	+ CDA	50.63 ↓ 1.04	51.47 ↑ 1.35	50.59 ↓ 2.12	48.22 ↓ 0.82	2
	+ DROPOUT	50.31 ↓ 1.36	50.90 ↑ 0.78	51.95 ↓ 0.76	49.77 ↓ 0.73	**3**
	+ INLP	47.84 ↓ 0.49	49.93 ↓ 0.05	51.49 ↓ 1.22	49.22 ↓ 0.18	**3**
	+ Sentence-Debias	47.91 ↓ 0.42	49.41 ↓ 0.47	53.01 ↑ 0.30	48.76 ↓ 0.28	0
	+ Context-Debias	46.58 ↓ 1.75	49.95 ↓ 0.07	52.00 ↓ 0.71	49.31 ↓ 0.27	**3**
	+ Auto-Debias	48.28 ↓ 0.05	48.92 ↓ 0.96	53.01 ↑ 0.30	47.88 ↓ 1.16	0
	+ MABEL	48.00 ↓ 0.33	49.55 ↓ 0.33	50.34 ↓ 2.37	49.82 ↓ 0.78	2
	+ Causal-Debias	48.39 ↓ 0.06	49.38 ↓ 0.50	48.97 ↓ 1.68	50.92 ↓ 0.04	**3**
	+ MEDICINE	50.42 ↓ 1.25	50.45 ↑ 0.33	49.93 ↓ 2.64	49.06 ↓ 0.02	**3**

Note: In the above, the evaluation framework incorporates multiple indicators with specific interpretation guidelines: an upward arrow (↑) denotes that higher performance is better, and a diamond symbol (◇) signifies that values closer to 50 are optimal. Within these metrics, green arrows / values indicate reductions in bias compared to baseline, whereas red arrows / values signify increased bias relative to baseline. The numerical values adjacent to the arrows represent the absolute difference between each debiasing method and BERT. For bias metrics, it represents the absolute difference between the respective absolute deviations from 50. The best observed results are highlighted in **bold**.

Table 2: Multi-Dimensional Bias Measurement Results of the Language Model on CoLA, QNLI, and SST Tasks.

5.1.2 Metrics and Implementation Details

We evaluate our approach using both **intrinsic bias metrics** and **extrinsic task metrics**. Intrinsic bias is measured using Stereotype Score (SS) and Stereotype Score Deviation ($SS_{\text{Deviation}}$) following StereoSet (Nadeem et al., 2021). Extrinsic performance is evaluated using Accuracy for QNLI and SST, and Matthews Correlation Coefficient (MCC) for CoLA. Training configurations and full metric definitions are provided in Appendix C.2 and Appendix C.3.

5.2 Experimental Results

In this section, we present our experimental results and discuss key observations derived from the analysis. The analyses are organized around the research questions outlined above, providing a comprehensive evaluation of the proposed approach. Specifically, we examine the effectiveness of our method in mitigating multi-dimensional biases while maintaining the model's performance on downstream tasks. Table 1 presents the overall debiasing effectiveness and task performance on CoLA, QNLI, and SST tasks. Table 2 illustrates the persistence of bias across various dimensions in the language model after debiasing.

5.2.1 Analysis of RQ1 - Effectiveness of Bias Mitigation

We first examine whether our proposed method effectively mitigates social biases in language models. As summarized in Table 1, across all three NLP tasks (CoLA, QNLI, and SST), our causal inference-based framework consistently reduces overall bias compared to the BERT baseline, as measured by the Stereotype Score (SS) and its average deviation (Avg. Dev). Since an SS value of 50 corresponds to a perfectly fair model, our method consistently maintains overall bias within ± 0.15 of this ideal point. Specifically, we observe

reductions of 0.36, 0.39, and 0.91 in overall SS for CoLA, QNLI, and SST, respectively, with larger improvements appearing in tasks where the baseline bias is more pronounced.

This consistent reduction reflects the core mechanism of our approach. By explicitly reducing the discrepancy between the model's total effect (TE) and its natural direct effect (NDE), the training process discourages reliance on bias-related tokens and encourages predictions to be driven by semantic information. As a result, the model captures more reliable and fair representations.

The Avg. Dev results in Table 1 further confirm this trend. Our method reduces average bias by 0.28, 1.02, and 0.90 on CoLA, QNLI, and SST, respectively. Although MABEL achieves the lowest Avg. Dev on QNLI by leveraging task-aligned NLI post-training data, our approach attains comparable performance (0.64) without any task-specific post-training, demonstrating the robustness and generality of our causal framework.

5.2.2 Analysis of RQ2 - Antagonistic Effects in Multi-Dimensional Bias Mitigation

We next investigate whether our method can effectively and jointly mitigate multiple bias dimensions within a single debiasing process. Following prior work, we separately evaluate biases along four dimensions: gender, profession, race, and religion. Detailed results are reported in Table 2.

As shown in Table 2, our method consistently reduces bias across most dimensions and achieves improvements in all four categories on the QNLI task. For CoLA and SST, minor increases are observed in isolated dimensions; however, these cases correspond to settings where the BERT baseline already exhibits near-neutral bias values, leaving limited room for further reduction. Similar patterns are also observed in other debiasing methods.

In contrast, many baseline approaches exhibit clear antagonistic effects across bias dimensions. When a method focuses on suppressing one specific type of bias, the model often shifts its reliance to other correlated attributes, leading to increased bias in additional dimensions. Our method avoids this issue by not targeting individual bias categories. Instead, it focuses on reducing the discrepancy between TE and NDE, encouraging the model to rely on semantic cues rather than bias-related signals. As a result, it mitigates multiple types of biases simultaneously without introducing systematic trade-offs across dimensions.

5.2.3 Analysis of RQ3 - Impact of Debiasing on Model Performance

Finally, we assess whether bias mitigation negatively affects downstream task performance. As reported in Table 1, across all three tasks, our method consistently outperforms the BERT baseline, achieving performance gains of 0.50 on CoLA, 0.22 on QNLI, and 0.58 on SST.

These results indicate that effective bias mitigation does not necessarily come at the expense of model performance. This finding supports our central hypothesis that many bias-related features do not provide useful task information but instead introduce spurious correlations. By reducing the model's reliance on such correlations through causal intervention, our framework guides the model to focus on semantic information that is more relevant to the task.

In contrast, many baseline methods exhibit a clear trade-off, where stronger debiasing is often accompanied by degraded performance. Our approach achieves a favorable balance, delivering strong task performance while simultaneously achieving the most consistent bias reduction.

6 Conclusion

From a causal inference perspective, we systematically analyze the mechanisms underlying language model inference by decomposing text representations into semantic, bias, and interaction components. Building on this framework, we disentangle bias-induced effects from the total causal effect by aligning it with the natural direct effect that captures purely semantic influence via counterfactual inference. Extensive experiments on three benchmark NLP tasks, evaluated using both intrinsic bias metrics and extrinsic task performance, demonstrate that our approach effectively mitigates bias while consistently preserving or improving downstream performance. Importantly, our framework is designed to jointly address multiple bias dimensions within a unified process, enabling a more faithful modeling of the interconnected nature of real-world biases and avoiding the antagonistic effects that commonly arise when mitigating single biases in isolation. Looking forward, promising directions include extending causal debiasing to generative and multilingual models, as well as developing broader cross-cultural benchmarks to enable more inclusive and robust bias evaluation.

Limitations

Despite growing attention to bias in language models, the development of comprehensive evaluation benchmarks remains limited. In this work, we rely on StereoSet, one of the most carefully curated bias benchmarks, constructed from natural language by human annotators rather than synthetic templates. However, its annotators are exclusively North American and were instructed to reflect North American stereotypes, which constrains coverage of regional and cultural biases. As a result, although we consider multiple bias dimensions such as gender and race, many culturally specific and geographically diverse stereotypes remain underexplored. We therefore encourage future research to broaden the scope of bias analysis across cultures and regions and to develop more inclusive benchmarks that enable systematic and rigorous evaluation of bias in language models.

References

Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. 2025. How can we diagnose and treat bias in large language models for clinical decision-making? In *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)*, pages 2263–2288, Albuquerque, New Mexico. Association for Computational Linguistics.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *9th International Conference on Learning Representations*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Yiwei Dai, Hengrui Gu, Ying Wang, and Xin Wang. 2024. Mitigate extrinsic social bias in pre-trained language models via continuous prompts adjustment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11068–11083.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545. Association for Computational Linguistics.

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

Farsheed Haque, Depeng Xu, and Shuhan Yuan. 2024. Discovering and mitigating indirect bias in attention-based model explanations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1599–1614, Mexico City, Mexico. Association for Computational Linguistics.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and on-demand bias mitigation with attribute-removal subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6192–6214, Toronto, Canada. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2024. Leveraging prototypical representations for mitigating social bias without demographic information. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 379–390, Mexico City, Mexico. Association for Computational Linguistics.

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

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8131–8138.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, Mingchen Sun, and Ying Wang. 2024a. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation. *Artificial Intelligence*, 332:104143.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024b. Data-centric explainable debiasing for improving fairness in pre-trained language models. In *Findings of the Association for Computational Linguistics*, pages 3773–3786.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14254–14267.

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

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197.

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

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

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

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations*.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246, Miami, Florida, USA. Association for Computational Linguistics.

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

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing

gendered correlations in pre-trained models. *CoRR*, abs/2010.06032.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15730–15745.

Maxwell J. Yin, Boyu Wang, and Charles Ling. 2025. MABR: Multilayer adversarial bias removal without prior bias knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25724–25732.

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

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

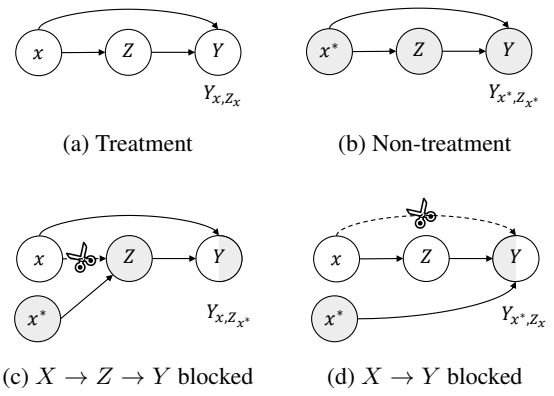
A Definitions and Case Study

A.1 Formal Definitions of Counterfactual Models

Counterfactual models are fundamental tools in causal inference, providing a principled way to translate causal assumptions encoded in causal graphs into mathematical expressions. The core idea of counterfactual analysis is to reason about the potential outcomes of variables under hypothetical intervention scenarios, enabling systematic comparison across different causal worlds.

Consider a causal graph involving three variables, where X denotes the cause variable, Y denotes the outcome variable, and Z serves as a mediator transmitting indirect causal effects from X to Y. When the value of X is set to x and the value of Z is set to z through intervention, the corresponding potential outcome of Y is defined as:

$$Y_{x,z} = Y(X = x, Z = z). \tag{11}$$



(a) Treatment (b) Non-treatment

(c) $X \to Z \to Y$ blocked (d) $X \to Y$ blocked

Figure 5: Counterfactual Models

In real-world scenarios, the mediator Z typically depends on the value of X, which can be expressed as $Z_x = Z(X = x)$. Counterfactual analysis allows this natural dependency to be decoupled by considering hypothetical scenarios in which X is set to one value while Z follows its potential outcome under a different value of X. Specifically, the counterfactual outcome

$$Y_{x,Z_{x^*}} = Y(X = x, Z = Z(X = x^*)) \tag{12}$$

represents a scenario where X is intervened to x, while the mediator Z is fixed to the value it would have taken under $X = x^*$.

This formulation enables the disentanglement of direct causal effects from indirect effects mediated through Z, which forms the foundation for identifying different components of causal influence in subsequent analysis.

A.2 Illustration of Counterfactual Models and Graphical Conventions

Figure 5 illustrates a set of counterfactual models that depict causal relationships among variables and their representations under different counterfactual scenarios. In the figure, node X represents the cause variable, node Y represents the outcome variable, and node Z serves as the mediator transmitting indirect causal effects from X to Y.

To improve visual clarity and conciseness, the variable X is not explicitly labeled as a node in the graph. Instead, the symbols x and x^* are used to denote two distinct intervention values of X, corresponding to $X = x$ and $X = x^*$, respectively. This notation emphasizes intervention states rather than the variable itself.

Different shading patterns are employed to distinguish variable states under counterfactual analysis. Fully shaded nodes indicate that the variable's

value is entirely determined by artificial intervention, independent of observed data. Semi-shaded nodes indicate that the variable's value is partially determined by the counterfactual scenario and partially inherited from the observed data distribution. This visualization strategy clarifies whether a variable is directly intervened upon or naturally propagated.

Additionally, scissor symbols are used to mark causal paths that are blocked due to human intervention. The presence of a scissor symbol on a causal edge indicates that the corresponding causal effect is severed in the counterfactual world. For example, when intervening on X, a scissor symbol on the path from X to Z indicates that the value of Z no longer depends on the natural state of X but is instead fixed according to the intervention condition. This graphical design facilitates intuitive interpretation of how direct and indirect causal effects are isolated through counterfactual reasoning.

A.3 Case Study: Counterfactual Reasoning in a Hiring Scenario

To further illustrate the concepts and definitions of counterfactual models, we provide a concrete case study based on a probation-to-regular hiring evaluation scenario. As shown in Figure 2, variable X represents an individual's gender, where $X = x$ denotes male and $X = x^*$ denotes female. Variable Z represents the individual's professional skills, and variable Y denotes the department's final hiring decision.

For illustrative purposes, this case assumes that male candidates exhibit stronger hands-on abilities, while female candidates exhibit stronger communication skills, with the alternative skill being relatively weaker. This assumption is introduced solely to simulate potential gender stereotypes that may exist in real-world decision-making processes and does not reflect the views or claims of this paper. The objective of the proposed method is to mitigate such biases.

ID	Scen.*	Instance	Formula
(a)	✓	Jack is skilled in hands-on tasks.	$Y_{x,Z_x} = Y(X = x, Z = Z(X = x))$
(b)	✗	Mary has communicating skills.	$Y_{x^*,Z_{x^*}} = Y(X = x^*, Z = Z(X = x^*))$
(c)	✗	Sam is adept at communicating	$Y_{x,Z_{x^*}} = Y(X = x, Z = Z(X = x^*))$
(d)	✗	Kate has hands-on skills.	$Y_{x^*,Z_x} = Y(X = x^*, Z = Z(X = x))$

* ✓: Factual; ✗: Counterfactual. Blue: male/skills; Orange: female/skills.

Table 3: Instances and Formulas for Counterfactual Models

Table 3 presents a complete example, where Mary, Sam, and Kate are artificially constructed counterfactual instances derived from the original case of Jack. These counterfactual instances do not exist in the actual dataset and are used exclusively for analytical comparison.

In this scenario, gender is treated as a protected attribute, and the model is explicitly prohibited from relying on gender-related information during inference, as such reliance may introduce discriminatory outcomes. Although this case study focuses on gender bias in a hiring context, similar forms of social bias are prevalent in language models and their downstream applications.

This case study demonstrates how counterfactual reasoning enables the separation of bias-induced influences from task-relevant information, thereby motivating the causal formulation adopted in our method for bias mitigation in word embedding models.

A.4 Causal Effect Identification

To formalize different causal pathways through which model predictions are formed, we introduce two key causal quantities: the **Total Effect (TE)** and the **Natural Direct Effect (NDE)**.

The **Total Effect (TE)** of the treatment $X = x$ on the outcome variable Y can be calculated by comparing the potential outcomes of Y under two hypothetical conditions. Here, Y_{x,Z_x} represents the potential outcome under the treatment condition $X = x$, and $Y_{x^*,Z_{x^*}}$ represents the potential outcome under the non-treatment condition $X = x^*$. The total effect reflects the overall impact of the treatment condition, encompassing both the direct effect and the indirect effect:

$$TE = Y_{x,Z_x} - Y_{x^*,Z_{x^*}}. \tag{13}$$

The **Natural Direct Effect (NDE)** refers to the causal effect of X on Y when the mediator Z is fixed to its counterfactual value under the non-treatment condition. Specifically, it describes the change in the potential outcome of Y when X transitions from x^* to x, while Z is fixed at Z_{x^*}:

$$NDE = Y_{x,Z_{x^*}} - Y_{x^*,Z_{x^*}}. \tag{14}$$

In the above formulation, $Y_{x,Z_{x^*}}$ represents the potential outcome when X is set to x and Z is fixed at its potential outcome under $X = x^*$, while $Y_{x^*,Z_{x^*}}$ represents the potential outcome when both X and Z take their non-treatment values.

Category	Count	Examples
Gender	409	*his*, *her*, *male*, ...
Race	13	*african*, *black*, *white*, ...
Religion	18	*islam*, *christian*, *jewish*, ...
Profession	95	*nurse*, *firefighter*, *guard*, ...
Appearance	18	*strong*, *cute*, *ugly*, ...
Personality	17	*aggressive*, *emotional*, ...
Activities	28	*shopping*, *football*, *boxing*, ...
Other	10	*flower*, *housework*, ...
Total	**608**	-

Table 4: Statistics and Examples of the Sensitive Lexicon.

B Sensitive Attribute Lexicon W_B

To construct a comprehensive lexicon for W_B, we draw on prior research (Meade et al., 2022; Guo et al., 2022; Zhou et al., 2023), which provides established definitions of sensitive attributes across multiple social categories. Based on these studies, we aggregated and organized the tokens into eight categories: Gender, Race, Religion, Profession, Appearance, Personality, Activities, and Other, resulting in a lexicon of 608 tokens. The resulting lexicon serves as a reliable basis for identifying bias-related terms. Table 4 shows the token counts for each category together with representative examples.

C Experimental Setup

C.1 Baseline Methods

To comprehensively evaluate the effectiveness of debiasing techniques, we employ the following state-of-the-art debiasing methods applied to BERT as baseline models.

- **CDA** (Webster et al., 2020) reduces model bias by constructing balanced training datasets through the substitution of sensitive attribute words with their corresponding counterparts.

- **DROPOUT** (Webster et al., 2020) hypothesizes that model bias stems from spurious correlations and mitigates these bias-related associations through an augmented training phase with an increased dropout rate.

- **INLP** (Ravfogel et al., 2020) iteratively eliminates bias-related features from model representations by projecting them onto the classifier's null space that captures attribute-related information.

- **Sentence-Debias** (Liang et al., 2020) generalizes word embedding debiasing to sentence encoders by identifying bias-associated subspaces via principal component analysis and projecting sentence representations onto their orthogonal complements.

- **Context-Debias** (Kaneko and Bollegala, 2021) mitigates discriminative gender-related biases in contextualized representations through orthogonal projection operations applied to hidden layers.

- **Auto-Debias** (Guo et al., 2022) automatically probes model bias using discriminative diagnostic prompts and mitigates bias through distributional alignment with an equalizing loss.

- **MABEL** (He et al., 2022) improves model fairness by integrating counterfactual data augmentation with a dual-objective post-training strategy that combines contrastive learning and masked language modeling objectives.

- **Causal-Debias** (Zhou et al., 2023) integrates debiasing into the fine-tuning stage of pretrained language models by learning causally invariant representations via invariant risk minimization.

While CPAD (Dai et al., 2024) and CDDD (Li et al., 2024a) represent significant methodological advances, we exclude them from our baseline comparisons due to critical limitations: CPAD's implementation is not publicly available, and CDDD exhibits prohibitive computational demands along with partial implementation failures.

C.2 Evaluation Metrics

In line with our experimental framework, we adopt two types of metrics to ensure a comprehensive evaluation: intrinsic metrics for bias measurement and extrinsic metrics for downstream task performance.

Intrinsic Metrics. For bias measurement, we employ the Stereotype Score (Nadeem et al., 2021) and its average deviation compared to an ideally fair language model.

Stereotype Score (SS) is a metric designed to quantify the degree to which a language model exhibits. For a target term t, the Stereotype Score $SS(t)$ is defined as the percentage of instances

in which the model prefers a stereotypical association over an anti-stereotypical one. Mathematically, $SS(t)$ is expressed as:

$$SS(t) = \frac{N_{\text{Stereo}}(t)}{N_{\text{Total}}(t)} \times 100. \quad (15)$$

where $N_{\text{Stereo}}(t)$ represents the number of times the model prefers stereotypical associations for the target term t, and $N_{\text{Total}}(t)$ represents the total number of instances for the target term t.

For an entire dataset, the overall Stereotype Score *SS* is computed as the average of the Stereotype Scores for all target terms in the dataset:

$$SS = \frac{1}{N} \sum_{i=1}^{N} SS(t_i), \quad (16)$$

where N indicates the total number of target terms in the dataset.

The SS for a specific bias type is quantified by calculating the overall SS of the subset of the dataset that aligns with that particular type of bias. An ideal language model should exhibit neutrality, demonstrating no preference for either stereotypical or anti-stereotypical associations, which corresponds to an SS value of 50. We advocate for equal probability between anti-stereotypical and stereotypical associations, as any form of overgeneralized belief can cause harm to target groups. Consequently, an SS value closer to 50 indicates a relatively fairer and less biased language model.

Stereotype Score Deviation ($SS_{\text{Deviation}}$) quantifies the disparity between an evaluated language model and an ideally fair model, serving as a key metric to assess the extent of bias inherent in the model. A perfectly fair model is characterized by a Stereotype Score of 50, representing complete neutrality between stereotypical and anti-stereotypical associations. Consequently, the $SS_{\text{Deviation}}$ of a language model is defined as the absolute difference between its *SS* and the ideal value of 50. Regarding a specific category of bias (e.g., gender), the Stereotype Score is computed as the overall SS of the dataset subset corresponding to that bias category. Thus, the $SS_{\text{Deviation}}$ for a particular bias category c is expressed as:

$$SS_{\text{Deviation}}(c) = |SS(c) - 50|. \quad (17)$$

To ensure a comprehensive and robust evaluation, we consider four major categories of stereotypical biases by following Nadeem et al. (2021), i.e., gender, profession, race, and religion. The overall Stereotype Score Deviation of the language model, denoted as Avg. Dev, is defined as the arithmetic mean of the $SS_{\text{Deviation}}$ across these four bias categories. Mathematically, this is represented as:

$$\text{Avg. Dev} = \frac{1}{|C|} \sum_{c \in C} SS_{\text{Deviation}}(c), \quad (18)$$

where $C = \{\text{gender, profession, race, religion}\}$.

Extrinsic Metrics. Following prior work (Wang et al., 2019), we employ the accuracy metric for the QNLI and SST tasks, and the Matthews Correlation Coefficient metric for the CoLA task. Both metrics are calculated based on true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Further details are provided below.

Accuracy (ACC) is one of the most commonly used metrics to evaluate the performance of classification models. It measures the proportion of correctly classified instances relative to the total number of instances. ACC is determined by the following formula, with values ranging from 0 to 1, where higher values signify superior model performance:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \quad (19)$$

Matthews Correlation Coefficient (MCC) is a metric used to evaluate the performance of classification models, particularly in binary classification tasks. It is especially robust when dealing with imbalanced datasets. MCC ranges between [-1, 1], where 1 indicates perfect prediction with all classifications correct. The formula of MCC is as follows:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (20)$$

C.3 Implementation Details

Following research (Meade et al., 2022; Zhou et al., 2023), we employ the bert-base-uncased version of the BERT model as the baseline for evaluation. For all methods involved in the experiments, we train the model for three epochs on each task, and a maximum sequence length of 128. We set the batch size to 32 and the learning rate to 2e-5. For the CDA and DROPOUT debiasing models, we incorporate an additional 1-epoch pre-training phase specifically designed to mitigate bias by following the original papers.

For all comparative methods, we utilize their gender-debiased versions, as research on gender bias is the most extensive and well-documented in the field. In contrast, our proposed method is uniquely designed to mitigate multiple types of biases simultaneously within a single, unified debiasing process. Consequently, in our experimental framework, we design our proposed model to address multiple forms of bias beyond gender bias. This capability allows our method to achieve broader fairness and robustness across diverse contexts, setting it apart from existing approaches that focus on a single bias type.

C.4 Sensitivity of Hyperparameter α

The hyperparameter α is used to balance the downstream task objective (L_{task}) and our causal debiasing constraint (L_{debias}). Specifically, a larger α imposes a stronger penalty on the divergence between the Total Effect (TE) and the Natural Direct Effect (NDE), pushing the model to rely more on pure semantics. While the optimal α may vary slightly across tasks, our experiments consistently show that values in the range of $[0.01, 0.1]$ provide a robust balance, effectively mitigating bias without compromising task performance. This indicates that our framework is not overly sensitive to this hyperparameter.