

ARCA: Attribution-Regularized Causal Alignment for Task-Aware Bias Mitigation

Botao Wang¹, Lin Li^{1*}, Kaize Shi², Jiahao Liu¹,
Chong Lyu¹, Xian Zhong¹

¹Wuhan University of Technology, China, ²University of Southern Queensland, Australia

¹{botaowang, cathylin, jh.liu, 346816, zhongx}@whut.edu.cn

²Kaize.Shi@unisq.edu.au

Abstract

Pretrained masked language models (MLMs) remain important backbones for natural language understanding (NLU) tasks, as they often provide a more favorable balance between task performance and computational efficiency than large decoder-only models. However, existing debiasing methods often decouple bias mitigation from downstream fine-tuning, making it difficult to preserve the mitigation effects after task adaptation. To address this, we propose ARCA (Attribution-Regularized Causal Alignment), an end-to-end framework that unifies debiasing quality and task utility. ARCA uses differential gradient attribution to identify task-specific bias words and regularize the intervention for causal alignment on counterfactual samples. This single-pipeline approach enables joint optimization of task utility and debiasing. Experiments on three mainstream NLU tasks show that ARCA achieves a strong trade-off between bias mitigation and task utility compared with representative debiasing baselines.¹

1 Introduction

Pretrained masked language models (MLMs) remain strong and efficient backbones for natural language understanding (NLU), providing a favorable balance between task performance and computational efficiency compared with large decoder-only language models (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Zhao et al., 2021). By learning generalizable linguistic patterns from large-scale corpora, MLMs provide a robust foundation for various downstream applications. However, these models inevitably encode social biases (e.g., gender and racial stereotypes) embedded in their training data (Bolukbasi et al., 2016; Caliskan et al., 2016; Rudinger et al., 2017; Bender et al., 2021).

*Corresponding author.

¹Code is available at: <https://github.com/WUT-IDEA/ARCA>

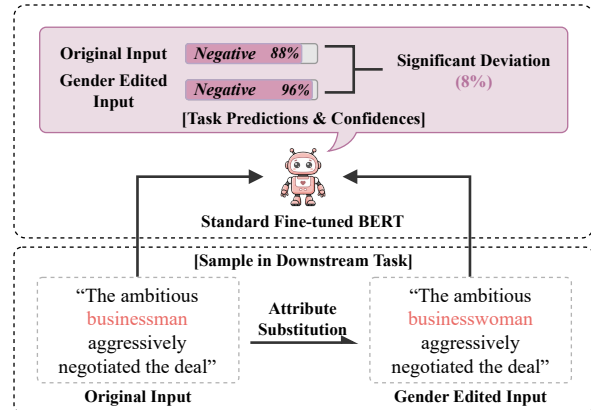


Figure 1: Illustration of gender bias re-introduction during downstream fine-tuning. The standard fine-tuned BERT exhibits a significant confidence gap (8%) between gender-paired inputs.

During downstream fine-tuning, such implicit biases (Kurita et al., 2019; Nadeem et al., 2021) are often amplified into consistent and discriminatory predictions against specific demographic groups, leading to unfair outcomes in high-stakes scenarios (e.g., resume screening and legal judgment assistance) (De-Arteaga et al., 2019; Blodgett et al., 2020). Driven by growing regulatory and ethical concerns, mitigating task-level social bias in MLMs while preserving utility remains a critical challenge (Liang et al., 2021).

Despite extensive research into bias mitigation for pretrained language models (Tan and Celis, 2019; May et al., 2019; Nangia et al., 2020; Kaneko and Bollegala, 2021; Parrish et al., 2022), existing methods still face a fundamental drawback: bias mitigation is often decoupled from downstream fine-tuning, making it difficult to preserve the mitigation effects after task adaptation (Shen et al., 2022; Zhou et al., 2023). As a result, debiasing performed before downstream training may not consistently translate into stable task-level bias mitigation, as illustrated in Figure 1. This sepa-

ration is further compounded by the reliance on high-coverage manual lexicons, since building a task-specific lexicon is labor-intensive, difficult to scale, and hard to verify for complete coverage of context-specific bias-propagating words. Consequently, the disconnect between debiasing and downstream fine-tuning impedes end-to-end gradient flow and makes it difficult to jointly optimize debiasing quality and task utility.

To address these challenges, we propose ARCA (Attribution-Regularized Causal Alignment), an end-to-end debiasing framework for task-aware bias mitigation in pretrained MLMs. ARCA treats social bias as a spurious causal pathway from protected attributes to model predictions, a formulation motivated by the structural causal model perspective (Pearl, 2009). Specifically, ARCA first collects task-specific gradient signals and uses differential gradient attribution to construct a causal intervention lexicon, thereby regularizing causal alignment by constraining counterfactual interventions to task-specific bias-propagating words. During downstream fine-tuning, this lexicon guides the on-the-fly generation of counterfactual pairs, while the model is optimized with both task classification loss and causal alignment loss between original and counterfactual representations. This integrated pipeline enables stable task-aware debiasing while preserving task utility.

Overall, our contributions are the following: (i) We formulate task-level social bias mitigation from a causal perspective, treating protected attributes as spurious causal pathways that may influence downstream predictions. (ii) We introduce ARCA, an end-to-end framework for pretrained MLMs that unifies automatic causal intervention lexicon construction, on-the-fly counterfactual generation, and representation-level causal alignment during downstream fine-tuning. (iii) Experiments on three mainstream NLU tasks demonstrate that ARCA achieves a stronger trade-off between bias mitigation and task utility than representative baselines.

2 Related Work

2.1 Data-level Debiasing

Data-level methods mitigate bias by modifying or augmenting training data to reduce skewed statistical associations. A widely adopted paradigm is Counterfactual Data Augmentation (CDA) (Lu et al., 2020), which constructs paired samples by substituting protected attribute terms. Early re-

search utilized rule-based substitutions to generate gender-balanced corpora (Zhao et al., 2018; Webster et al., 2021), while recent approaches leverage large language models to produce contextually rich counterfactuals with improved diversity (Han et al., 2024; Parihar et al., 2026). However, these methods depend heavily on manual high-coverage lexicons or external generators and often fail to capture implicit, task-specific bias.

2.2 Model-Level Debiasing

Model-level approaches directly modify representations or training objectives to remove bias, often grounded in counterfactual fairness (Kusner et al., 2017) via causal intervention. Existing strategies primarily fall into three categories: **(i) Representation Projection:** Methods like INLP (Ravfogel et al., 2020) and Sentence-Debias (Liang et al., 2020) remove biased information via orthogonal projection or principal component analysis (PCA). However, these methods assume that bias directions are linearly identifiable and may remove task-relevant semantic information entangled with bias signals in deep transformer representations. **(ii) Contrastive Learning:** FairFil (Cheng et al., 2021), MABEL (He et al., 2022), and TriConFair (Lyu et al., 2026) promote bias-invariance by minimizing distances between counterfactual pairs. However, contrastive methods are sensitive to counterfactual construction and objective design, and noisy pair relations may blur the boundary between spurious bias and task-relevant semantics. **(iii) Parameter-efficient and Regularization-based:** Recent work explores parameter-efficient debiasing through adapter-based architectures (Jain et al., 2024) or integrates fairness constraints via adversarial learning (Qian et al., 2019) and divergence-based regularization (Li et al., 2025; Gallegos et al., 2024). Among debiasing approaches integrated with downstream fine-tuning, Causal-Debias (Zhou et al., 2023) employs causal invariant learning over intervention environments constructed from predefined demographic factors. However, these methods often lack precise bias localization and may still struggle to balance bias mitigation with task utility across diverse tasks.

Despite these efforts, stable task-level bias mitigation during downstream adaptation remains challenging. Data-level methods rely on manual high-coverage lexicons or external generators, while model-level methods often lack precise localization of task-specific bias-propagating features. A com-

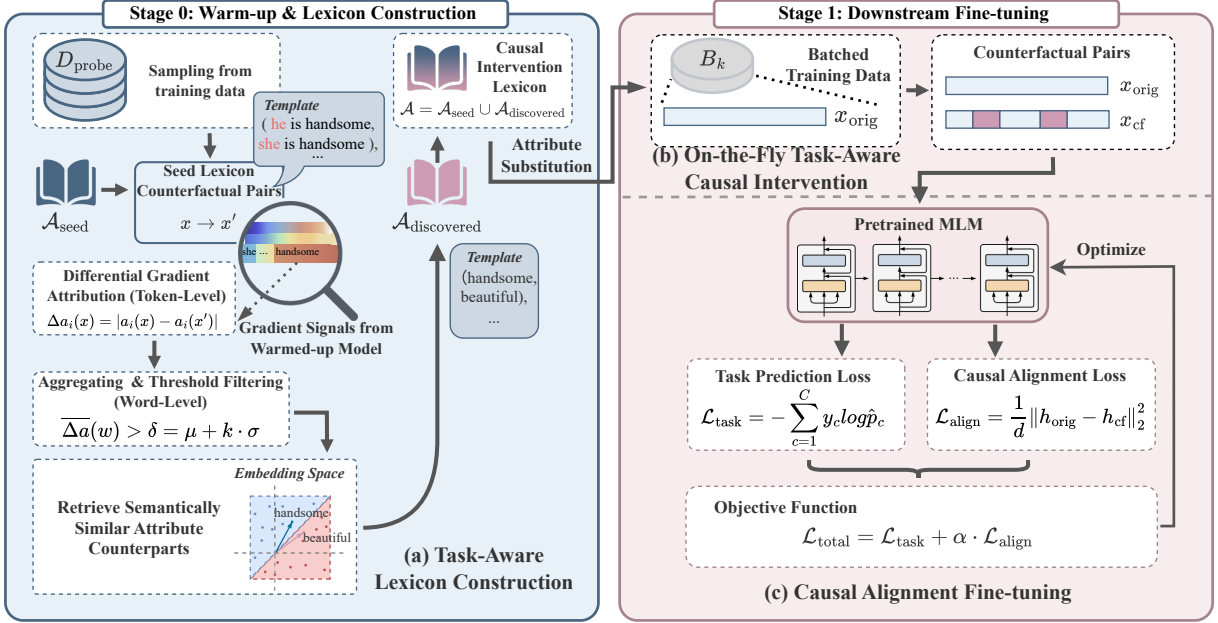


Figure 2: Overview of ARCA. (a) Task-aware lexicon construction discovers task-specific bias-propagating words. (b) On-the-fly causal intervention generates counterfactual pairs via attribute substitution. (c) Causal alignment fine-tuning jointly optimizes task prediction and representation alignment through divergence-based regularization.

mon limitation is that debiasing is often decoupled from downstream fine-tuning, making it difficult to preserve mitigation effects after task adaptation. In contrast, ARCA builds task awareness into each stage of debiasing: it constructs intervention lexicons from task-specific attribution signals, generates counterfactual pairs on the fly during downstream fine-tuning, and applies representation-level causal alignment within the same training pipeline. This unified framework enables task-aware debiasing while preserving task utility.

3 Methodology

3.1 Overview

Figure 2 illustrates the overall architecture of ARCA. ARCA is designed as an end-to-end framework for task-aware bias mitigation in pretrained MLMs, which often provide a more favorable balance between task performance and computational efficiency. The framework consists of two consecutive stages. In Stage 0, ARCA first performs a short downstream warm-up and constructs a task-specific causal intervention lexicon from model-internal attribution signals. In Stage 1, ARCA uses the constructed lexicon during downstream fine-tuning through two coupled operations: on-the-fly counterfactual generation via attribute substitution, followed by causal alignment fine-tuning that jointly optimizes task classification and representation-

level causal alignment. In this sense, the attribution signals regularize causal alignment by constraining the intervention space to task-specific bias-propagating words, rather than applying alignment to arbitrary or manually exhaustive substitutions.

ARCA adopts a causal framing while making no claim of full causal identification. Motivated by the structural causal model perspective (Pearl, 2009), ARCA treats social bias as a spurious pathway from protected attributes to model predictions. Specifically, lexical co-occurrences such as female–nurse reflect statistical associations rather than causal dependencies. Nevertheless, during downstream fine-tuning, MLMs may exploit protected attributes A as shortcut features for predictions $\hat{Y}$, inducing a spurious path $A \rightarrow \hat{Y}$ alongside the task-relevant semantic path $S \rightarrow \hat{Y}$, where S denotes task-relevant semantic features. Inspired by counterfactual fairness, ARCA approximates the intervention $do(A = a')$ through protected-attribute substitution, where a' denotes a counterfactual attribute value. Under this intervention-based formulation, ARCA uses such interventions to regularize representation learning by aligning original and counterfactual representations when the substitution preserves task semantics. This encourages the model to reduce its sensitivity to changes in A , thereby weakening the protected-attribute shortcut while retaining task-relevant semantic information.

3.2 Task-Aware Lexicon Construction

As shown in Figure 2 (a), the first component of ARCA aims to construct a causal intervention lexicon via differential gradient attribution for generating counterfactual inputs during downstream training. Instead of relying on manually constructed high-coverage lexicons, ARCA uses task-specific attribution signals to discover bias-propagating words from the warmed-up model. Formally, the causal intervention lexicon is defined as:

$$\mathcal{A} = \mathcal{A}_{\text{seed}} \cup \mathcal{A}_{\text{discovered}} \quad (1)$$

where $\mathcal{A}_{\text{seed}}$ is a small seed lexicon of linguistically unambiguous core attribute pairs, e.g., *(he, she)* and *(him, her)* in the gender bias setting. It serves as a compact anchor set rather than an exhaustive lexicon. $\mathcal{A}_{\text{discovered}}$ is automatically expanded from attribution signals computed by a warmed-up model, enabling ARCA to capture task-specific bias-propagating words from partially adapted gradient signals.

3.2.1 Token-Level Gradient Attribution

ARCA leverages model-internal gradient signals to discover bias-propagating words from token-level attribution signals. A probing set D_{probe} is first sampled from the training data. For each input x , we construct a counterfactual x' by replacing protected attributes using $\mathcal{A}_{\text{seed}}$. Following gradient $\times$ input attribution methods (Ancona et al., 2018; Wang et al., 2020), token-level attribution scores are computed as:

$$a_i(x) = \left\| e_i \odot \frac{\partial \mathcal{L}_{\text{task}}(x)}{\partial e_i} \right\|_2 \quad (2)$$

where e_i denotes the embedding vector of the i -th token, and $\mathcal{L}_{\text{task}}$ denotes the task prediction loss, instantiated as cross-entropy loss for NLU classification tasks. The differential gradient attribution score at token positions is defined as:

$$\Delta a_i(x) = |a_i(x) - a_i(x')| \quad (3)$$

A large $\Delta a_i(x)$ indicates that the model changes its reliance on the i -th token when the protected attribute is intervened upon. ARCA uses such changes to identify candidate bias-propagating lexical features. Here, “bias-propagating” refers to a model-level attribution pattern rather than an inherent property of the lexical item. A word need not carry explicitly biased or harmful semantics;

instead, a substantial attribution change indicates that its contribution to the prediction varies with the protected attribute, potentially reflecting contextual or statistical associations acquired during pretraining or downstream adaptation.

3.2.2 Word-Level Aggregation and Filtering

We then aggregate differential gradient attribution scores across the probing set. Let $\mathcal{P}(w)$ denote the set of positions (j, i) where word w appears in the probing set, with j and i indicating the sentence index and the token position, respectively. The aggregated score of word w is defined as:

$$\overline{\Delta a}(w) = \frac{1}{|\mathcal{P}(w)|} \sum_{(j,i) \in \mathcal{P}(w)} \Delta a_i^{(j)} \quad (4)$$

Candidate words are selected based on attribution strength and frequency $\text{Freq}(w)$. We define the attribution threshold as:

$$\delta = \mu_{\overline{\Delta a}} + k \cdot \sigma_{\overline{\Delta a}} \quad (5)$$

where $\mu_{\overline{\Delta a}}$ and $\sigma_{\overline{\Delta a}}$ denote the mean and standard deviation of aggregated differential attribution scores over non-seed words, respectively, and k is a scaling factor. A word w is retained if:

$$\overline{\Delta a}(w) > \delta \quad \text{and} \quad \text{Freq}(w) \geq n_{\min} \quad (6)$$

where $n_{\min}$ is the minimum frequency threshold. From this candidate set, we retrieve nearest semantic neighbors in the embedding space of the pre-trained model that satisfy attribute opposition constraints, and validate them via substitution consistency. The validated pairs are then used to construct the expanded lexicon $\mathcal{A}_{\text{discovered}}$. Detailed procedures for counterpart selection and substitution validation are provided in Appendix A.1.

By selecting words according to attribution strength, ARCA constructs a task-specific intervention space composed of validated bias-propagating words. Appendix A.2 provides representative discovered pairs with attribution statistics, together with the complete set of automatically discovered intervention pairs. As illustrated by these examples, the discovered lexicon extends beyond canonical protected-attribute terms such as pronouns and identity nouns, and can include context-dependent lexical carriers whose attribution changes under protected-attribute substitution. The current formulation therefore focuses on lexical-level intervention patterns associated with protected-attribute substitutions.

3.3 On-the-Fly Task-Aware Causal Intervention

The second component of ARCA, illustrated in Figure 2 (b), performs on-the-fly causal intervention by generating counterfactual pairs on the fly during training using the constructed task-specific lexicon $\mathcal{A}$. For each input sentence x_{orig} , ARCA scans the input for lexical matches in $\mathcal{A}$. If no match is found, the sample is only used for standard task training. Otherwise, the matched words are replaced with their attribute counterparts to produce a counterfactual x_{cf} . This procedure yields counterfactual pairs $(x_{\text{orig}}, x_{\text{cf}})$ that differ only in protected attribute terms while preserving the original semantic content and sharing the same task label. Guided by the causal intervention lexicon $\mathcal{A}$, the attribute substitution process serves as a targeted causal intervention during training. Since generation is deterministic and based on lexicon lookup, it is computationally efficient with time complexity $O(n)$; semantic preservation is encouraged by restricting substitutions to validated attribute counterparts. The detailed algorithm is provided in Appendix B.1.

3.4 Causal Alignment Fine-Tuning

The final component of ARCA, illustrated in Figure 2 (c), jointly optimizes downstream task performance and causal alignment between original and counterfactual representations.

Task Prediction Loss. For NLU classification tasks, the task prediction objective is naturally formulated as cross-entropy over task labels. The task loss is computed on original inputs:

$$\mathcal{L}_{\text{task}} = - \sum_{c=1}^C y_c \log \hat{p}_c \quad (7)$$

where C is the number of categories, $\hat{p}$ is the predicted probability distribution produced by the model, and y is the true label.

Causal Alignment Loss. For each causal intervention pair, we extract [CLS] representations $h_{\text{orig}}, h_{\text{cf}} \in \mathbb{R}^d$, then implement causal divergence regularization as a representation-level alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{d} \|h_{\text{orig}} - h_{\text{cf}}\|_2^2 \quad (8)$$

This encourages the model to produce invariant representations under protected-attribute intervention, thereby suppressing bias propagation before the prediction layer.

Objective Function. The overall optimization objective of ARCA is defined as a weighted sum of the task loss and the causal alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \cdot \mathcal{L}_{\text{align}} \quad (9)$$

where $\alpha \geq 0$ is a hyperparameter that controls the strength of the bias mitigation constraint. We provide the training algorithm in Appendix B.2.

4 Experimental Setup

4.1 Datasets

ARCA is evaluated on three representative NLU tasks from the GLUE benchmark (Wang et al., 2019): MNLI (Williams et al., 2018), SST-2 (Socher et al., 2013), and MRPC (Dolan and Brockett, 2005). MNLI is a three-way natural language inference task containing 392,702 training examples. SST-2 contains 67,349 movie-review sentences for binary sentiment classification. MRPC is a binary paraphrase identification benchmark with 3,668 training examples. Results are reported on the standard validation splits for SST-2 and MRPC, and on the matched validation split for MNLI.

StereoSet (Nadeem et al., 2021) is additionally included as an auxiliary reference for conventional intrinsic bias evaluation. It measures stereotypical associations across gender, race, and religion, and is commonly used to assess bias before downstream task adaptation. Since ARCA constructs task-specific intervention spaces from downstream attribution signals, direct evaluation of the task-adapted ARCA model on StereoSet would introduce a distributional mismatch. Therefore, StereoSet results are reported for existing debiasing methods, including additional decoder-only backbones, as reference results, while ARCA is evaluated on downstream NLU tasks targeted by the proposed framework. Details of datasets are presented in Appendix C.1.

4.2 Evaluation Metrics

For downstream NLU tasks, both task utility and task-level debiasing performance are evaluated. Task utility is measured by classification accuracy (Acc). Debiasing performance is measured by Prediction Flip Rate (PFR) and Probability Divergence Score (PDS), which capture label-level and probability-level sensitivity to protected-attribute substitution, respectively. PFR and PDS are intended to measure task-level counterfactual robustness under controlled protected-attribute substitu-

tions, rather than comprehensive fairness across demographic groups. For all methods, PFR and PDS are computed on the held-out validation split using a consistent counterfactual generation procedure and metric implementation. The intervention lexicon used to construct counterfactual samples follows the corresponding experimental configuration.

For conventional intrinsic bias evaluation, StereoSet metrics are additionally reported as auxiliary references for existing debiasing methods. Specifically, stereotype scores (SS) are reported for gender, race, and religion, where values closer to 50 indicate weaker stereotypical preference. LM Score and ICAT are also reported to measure language modeling ability and the overall trade-off between language modeling quality and stereotype bias, respectively. Since ARCA is task-aware and constructs its intervention space from downstream attribution signals, StereoSet serves as a reference for traditional intrinsic bias evaluation rather than the primary evaluation target of ARCA. Detailed definitions of the metrics are provided in Appendix C.2.

4.3 Baselines and Implementation

Representative debiasing baselines are selected to cover major categories of existing debiasing approaches, including Standard Fine-tuning, CDA (Lu et al., 2020), INLP (Ravfogel et al., 2020), Sentence-Debias (Liang et al., 2020), DBR (Li et al., 2025), and TriCon-Fair (Lyu et al., 2026). DBR is included as the closest objective-level fine-tuning-integrated baseline under our protocol, since it directly regularizes divergence between original and intervened inputs during downstream adaptation. This enables a controlled comparison with ARCA at the training-objective level while keeping the downstream task setting consistent. Official implementations are used when available, and all hyperparameters are tuned on the validation set under the same experimental protocol. The main downstream experiments are conducted on two widely adopted MLMs: BERT-base-uncased (Devlin et al., 2019) with 110M parameters and RoBERTa-base (Liu et al., 2019) with 125M parameters. For intrinsic bias reference on StereoSet, additional decoder-only backbones, including GPT-2 (Radford et al., 2019) and Llama2-7B (Touvron et al., 2023), are included only as auxiliary comparisons rather than target architectures of ARCA. More implementation details are provided in Appendix C.3.

5 Results and Analysis

In this section, we analyze ARCA from three perspectives: the relation between conventional intrinsic bias evaluation and downstream task-level debiasing in Section 5.1, the contribution of lexicon construction in Section 5.2, and sensitivity to key hyperparameters in Section 5.3. Instance-level case studies are presented in Appendix D.

5.1 Main Results

Table 1 and Figure 3 present the main results. Table 1 places conventional intrinsic bias evaluation and downstream task-level evaluation side by side. StereoSet results are reported as auxiliary references for existing debiasing methods, while MNLI results evaluate downstream fine-tuning.

Intrinsic Bias Reference vs. Downstream Bias.

The comparison between StereoSet and MNLI shows that strong performance on conventional intrinsic bias benchmarks does not necessarily guarantee preserved debiasing effects after downstream fine-tuning. On StereoSet, results show that conventional debiasing methods can reduce intrinsic stereotype scores across different backbones. For example, on BERT, TriCon-Fair reduces Gender SS from 60.28 to 55.68 and improves ICAT from 69.02 to 72.05. Similar trends on decoder-only backbones are reported separately in Table 2 as auxiliary intrinsic-bias references across architectures. For instance, TriCon-Fair reduces GPT-2’s Gender SS from 62.65 to 55.43 and improves ICAT from 69.90 to 77.86; on Llama2-7B, it improves ICAT from 87.97 under CDA to 89.95. These results indicate that stereotype bias is not limited to encoder-only MLMs, and that conventional debiasing methods can mitigate intrinsic bias to some extent under traditional bias benchmarks.

However, improvements on intrinsic bias benchmarks do not necessarily eliminate downstream task-level sensitivity. On MNLI, conventional debiasing methods still exhibit non-negligible PFR and PDS after fine-tuning. For instance, INLP reduces BERT’s PFR to 2.3%, but its PDS remains 0.045; TriCon-Fair obtains a lower PFR of 1.8%, but still has a PDS of 0.035. These results suggest that intrinsic debiasing and downstream task-level debiasing capture different aspects of bias mitigation, motivating the need for end-to-end task-aware evaluation. Since ARCA is designed for MLM-based downstream fine-tuning and relies on task-specific attribution signals, decoder-only results

Backbone	Method	StereoSet					MNLI		
		Gender SS → 50	Race SS → 50	Religion SS → 50	LM score↑	ICAT↑	Acc(%)↑	PFR(%)↓	PDS↓
BERT	Original	60.28	57.03	59.70	84.17	69.02	84.5	4.2	0.085
	CDA	59.61	56.73	58.37	83.08	69.39	83.9	2.8	0.052
	INLP	57.25	57.29	60.31	80.63	67.28	84.3	2.3	0.045
	Sentence-Debias	59.37	57.78	58.73	84.20	69.67	84.4	2.7	0.053
	TriCon-Fair	55.68	56.82	57.13	82.89	72.05	<u>84.3</u>	<u>1.8</u>	<u>0.035</u>
	DBR	-	-	-	-	-	84.2	2.5	0.048
	ARCA (ours)	-	-	-	-	-	84.4	0.9	0.018
BERT Variants	Original	59.93	57.51	60.32	89.77	73.16	87.6	3.8	0.072
	CDA	55.85	53.15	58.70	77.11	68.01	87.1	2.4	0.045
	INLP	58.05	55.00	63.77	86.58	71.10	87.2	2.0	0.038
	Sentence-Debias	58.38	57.95	56.09	88.98	75.69	<u>87.4</u>	2.3	0.044
	TriCon-Fair	56.33	55.42	56.58	86.71	76.11	87.3	<u>1.5</u>	<u>0.029</u>
	DBR	-	-	-	-	-	87.3	2.1	0.041
	ARCA (ours)	-	-	-	-	-	87.5	0.7	0.014

Table 1: Main results comparing conventional intrinsic bias evaluation and downstream task-level debiasing. StereoSet results are reported as auxiliary references for existing debiasing methods on backbones, while MNLI results evaluate downstream fine-tuning on BERT-base and RoBERTa-base. "-" indicates that the corresponding setting is not evaluated. The best results on MNLI are highlighted in bold and the second-best results are underlined.

Backbone	Method	StereoSet				
		Gender SS → 50	Race SS → 50	Religion SS → 50	LM score↑	ICAT↑
GPT-2	Original	62.65	58.90	63.26	91.01	69.90
	CDA	64.02	57.31	63.55	90.36	69.34
	INLP	60.17	58.96	63.95	91.62	71.41
	Sentence-Debias	56.05	56.43	59.62	87.43	74.54
	TriCon-Fair	55.43	57.33	58.31	90.58	77.86
Llama2-7B	Original	56.25	43.36	-	-	-
	CDA	55.71	44.74	56.31	92.12	87.97
	INLP	55.21	45.28	55.81	91.57	87.72
	Sentence-Debias	55.84	43.98	55.71	91.76	88.38
	TriCon-Fair	52.53	45.47	56.12	92.48	89.95

Table 2: StereoSet results as auxiliary references for existing debiasing methods on decoder-only backbones. "-" indicates that the corresponding setting is not evaluated.

are included only as intrinsic-bias references rather than direct targets of our method.

Task Utility. According to the MNLI results in Table 1 and the SST-2 and MRPC results shown in Figure 3, ARCA largely preserves the performance of the models across downstream tasks. On MNLI, ARCA achieves 84.4% accuracy with BERT-base, close to standard fine-tuning at 84.5%, and obtains 87.5% with RoBERTa-base, compared with 87.6% under standard fine-tuning. On SST-2, ARCA reaches 92.7% accuracy, only 0.1 percentage points lower than standard fine-tuning. On MRPC, ARCA achieves 87.26% accuracy, outperforming standard fine-tuning at 84.55%. These results indicate that ARCA does not substantially impair task-relevant semantic discrimination and may provide an implicit regularization effect on small-scale semantic matching tasks.

Debiasing performance. ARCA achieves strong downstream debiasing performance, especially in reducing probability-level attribute sensitivity. On MNLI, ARCA reduces PDS from 0.085 to 0.018 with BERT-base and from 0.072 to 0.014 with RoBERTa-base. It also reduces PFR from 4.2% to 0.9% on BERT-base and from 3.8% to 0.7% on RoBERTa-base. On SST-2, ARCA further reduces PDS from 0.068 to 0.006 and PFR from 3.5% to 0.3%. On MRPC, ARCA obtains the lowest PDS of 0.0178 and reduces PFR from 3.05% to 2.10%, although TriCon-Fair achieves a slightly lower PFR. Notably, the gains are not unique to a method simply being exposed to counterfactual substitutions during training. CDA and TriCon-Fair also employ substitution-based counterfactual training, yet they retain higher PFR and PDS than ARCA under the identical evaluation protocol. This suggests

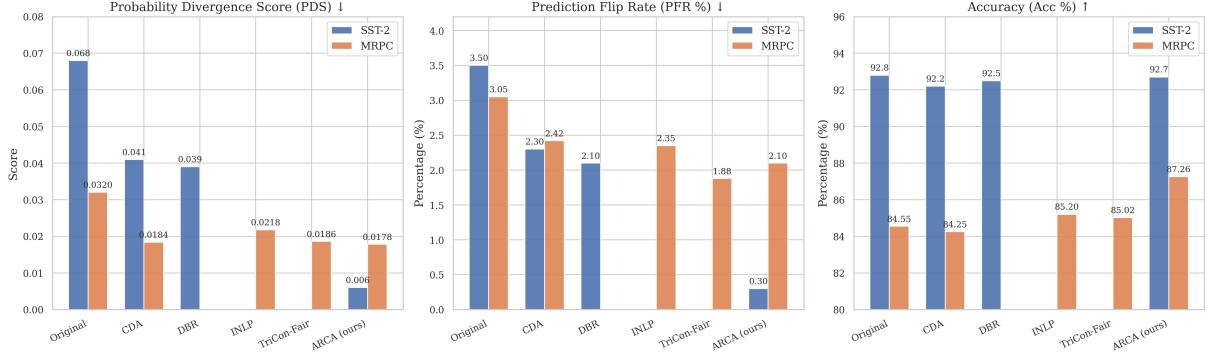


Figure 3: Results on SST-2 and MRPC with BERT as the backbone model.

Lexicon	Size (Pairs)	Acc(%)↑	PFR(%)↓	PDS↓
S.L.	~ 8	92.6	2.1	0.040
M.L.	~ 60	92.5	0.9	0.017
R.L.	~ 60	92.4	3.3	0.063
C.L.(ours)	~ 55	92.7	0.8	0.015

Table 3: Ablation results of different lexicon construction strategies.

that the improvement is associated with the quality of the task-specific intervention space, rather than exposure to the intervention type alone.

Overall, ARCA provides a strong trade-off between task utility and debiasing performance. The consistently low PDS reflects the effect of attribution-regularized causal alignment: attribution signals constrain interventions to task-specific bias-propagating words, while representation-level alignment reduces probability shifts between original and counterfactual inputs. Although ARCA does not always achieve the lowest PFR, its stable reduction of PDS across tasks suggests stronger robustness at the probability level while maintaining competitive task accuracy.

5.2 Ablation Study

We conduct ablation experiments on the lexicon construction strategy on SST-2 using BERT-base as the backbone model, with $\alpha = 2$. We design four lexicon variants for comparison: a seed-only lexicon (S.L.), a manual high-coverage lexicon (M.L.), a random lexicon (R.L.), and the causal intervention lexicon (C.L.) used by ARCA. All variants use the same training and alignment objective and differ only in the construction of the intervention lexicon, allowing us to isolate the contribution of the intervention space.

The results are shown in Table 3. The seed-only lexicon yields a significantly higher PFR than the

k	Acc(%)↑	PFR(%)↓	PDS↓
1.0	92.20	0.46	0.0790
1.5	93.12	0.00	0.0064
2.0	92.20	0.46	0.0790

Table 4: Sensitivity analysis of ARCA to the threshold scaling factor k , where $k \in \{1.0, 1.5, 2.0\}$.

full lexicon, indicating that covering only core pronouns is insufficient for capturing the broader set of attribute-sensitive lexical patterns relevant to downstream prediction. The random lexicon provides little debiasing improvement and even hurts task accuracy, suggesting that arbitrary substitutions may introduce semantic noise rather than effective debiasing signals. In contrast, the causal intervention lexicon achieves performance comparable to, and slightly better than, the manual high-coverage lexicon. This shows that attribution-driven lexicon construction can identify task-specific and context-dependent lexical carriers of protected-attribute sensitivity that are difficult to exhaustively cover through manual lexicon construction.

5.3 Sensitivity Analysis

5.3.1 Effect of Regularization Weight

As shown in Figure 4, we evaluate the sensitivity of ARCA to α on SST-2 with BERT-base, varying α from 0.0 to 3.0. The weight α controls the relative contribution of the causal alignment loss to the task prediction loss: larger values impose stronger representation consistency between original and counterfactual samples, while smaller values make the training objective closer to standard fine-tuning.

When $\alpha = 0$, ARCA reduces to standard fine-tuning and exhibits the highest PFR and PDS. As α increases, both metrics generally decrease, indicating that causal alignment effectively suppresses

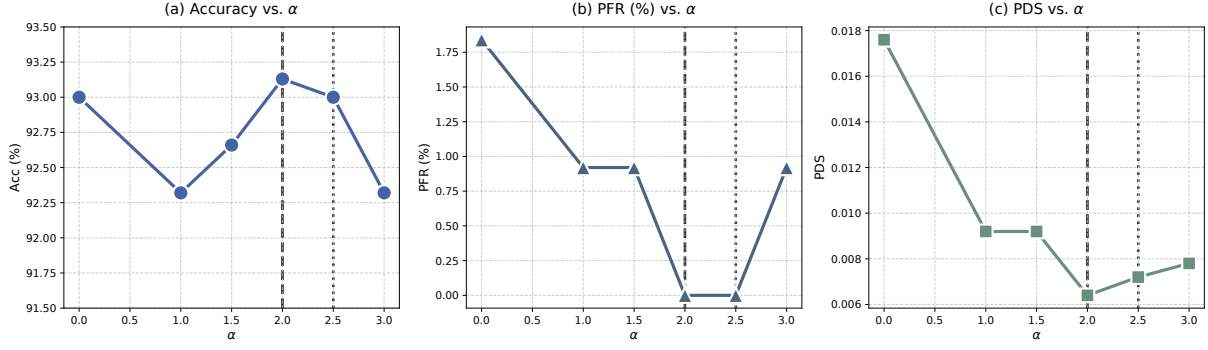


Figure 4: Sensitivity analysis of ARCA to the regularization weight α on SST-2 with BERT-base. Subfigures report the changes in accuracy, PFR, and PDS as α varies from 0.0 to 3.0. The dashed vertical lines indicate the recommended range $\alpha \in [2.0, 2.5]$.

attribute-induced prediction sensitivity. The best trade-off is observed around $\alpha = 2.0$, where PFR approaches zero and PDS reaches its lowest value while accuracy remains high. When α further increases to 3.0, PFR slightly rebounds and accuracy decreases, suggesting that overly strong representation alignment may begin to interfere with task-level semantic discrimination. Overall, $\alpha \in [2.0, 2.5]$ offers the most stable trade-off between debiasing performance and task utility. We therefore set $\alpha = 2.0$ in our main experiments.

5.3.2 Effect of Threshold Scaling Factor

The scaling factor k controls the strictness of candidate-word filtering in Eq. 5. A smaller k retains more candidate words but may introduce noisy substitutions, whereas a larger k selects fewer words and may miss effective bias-propagating terms. Table 4 reports the results with $k \in \{1.0, 1.5, 2.0\}$.

The results show that ARCA is relatively robust to the threshold setting, with accuracy and PFR remaining within a small range across different k values. The default setting $k = 1.5$ achieves the best overall trade-off, reaching the highest accuracy of 93.12%, reducing PFR to 0.00%, and obtaining the lowest PDS of 0.0064. Therefore, we set $k = 1.5$ as the default configuration in our experiments.

6 Conclusion

This paper presents ARCA, an end-to-end framework for task-aware bias mitigation in pretrained masked language models. ARCA uses differential gradient attribution to construct a task-specific causal intervention lexicon, generates counterfactual pairs on the fly, and applies representation-level causal alignment during downstream fine-

tuning. Experiments on three NLU tasks show that ARCA achieves a strong trade-off between debiasing performance and task utility, reducing label- and probability-level sensitivity while maintaining competitive accuracy.

Limitations

Although ARCA achieves effective bias mitigation in our experiments, several limitations remain.

Multilingual and cross-lingual settings. Our experiments are conducted primarily on English datasets. However, social bias may manifest differently across languages and cultural contexts. For example, gender bias in Chinese may emerge through different lexical or syntactic patterns, while bias in Arabic can be closely related to grammatical gender. Future work may extend ARCA to multilingual settings and investigate cross-lingual debiasing strategies.

Extension to generative models. ARCA targets masked language models, which are often practical in resource-constrained scenarios, and relies on encoder-side [CLS] representations for alignment. Extending it to decoder-only generative models is non-trivial, as bias may emerge throughout long-form generation. Future work may combine causal alignment with parameter-efficient tuning methods.

Scope of counterfactual evaluation. Our evaluation focuses on task-level robustness to controlled protected-attribute substitutions. All methods are evaluated on held-out validation data using the same PFR/PDS implementation and counterfactual generation mechanism. Thus, these metrics remain intervention-dependent and do not establish broader group-level fairness or transfer to qualita-

tively different forms of bias. Future work should consider fixed, method-independent intervention sets and complementary evaluations of group-level disparities, calibration, intersectional effects, and real-world fairness.

Acknowledgments

This work is partially supported by the Hubei Provincial Natural Science Foundation of China under Grant 2026AFC0615. The authors also gratefully appreciate the reviewers for their valuable comments and constructive suggestions that greatly helped to improve the quality of the paper.

References

- Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. [Towards better understanding of gradient-based attribution methods for deep neural networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018*. OpenReview.net.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. [Language \(technology\) is power: A critical survey of "bias" in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356:183 – 186.
- Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. [Fairfil: Contrastive neural debiasing method for pretrained text encoders](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnamurthy Ken- thapadi, and Adam Tauman Kalai. 2019. [Bias in bios: A case study of semantic representation bias in a high-stakes setting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.
- 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- William B. Dolan and Chris Brockett. 2005. [Automatically constructing a corpus of sentential paraphrases](#). In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. [Bias and fairness in large language models: A survey](#). *Computational Linguistics*, 50(3):1097–1179.
- Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. [ChatGPT based data augmentation for improved parameter-efficient debiasing of LLMs](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 73–105, St. Julian's, Malta. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9681–9702. Association for Computational Linguistics.
- Prachi Jain, Ashutosh Sathe, Varun Gumma, Kabir Ahuja, and Sunayana Sitaram. 2024. [MAFIA: multi-adapter fused inclusive language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 627–645. Association for Computational Linguistics.
- Masahiro Kaneko and Danushka Bollegala. 2021. [De-biasing pre-trained contextualised embeddings](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. [Measuring bias in contextualized word representations](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.
- Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. [Counterfactual fairness](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 4069–4079, Red Hook, NY, USA. Curran Associates Inc.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised learning of language representations](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zihao Li, Ruixiang Tang, Lu Cheng, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2025. [DBR: Divergence-based regularization for debiasing natural language understanding models](#). *SIGKDD Explor. Newsl.*, 27(1):1–9.
- 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, ACL 2020, Online, July 5-10, 2020*, pages 5502–5515. Association for Computational Linguistics.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. [Towards understanding and mitigating social biases in language models](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.
- 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.
- Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. [Gender bias in neural natural language processing](#). In *Logic, Language, and Security - Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, Lecture Notes in Computer Science, pages 189–202. Springer.
- Chong Lyu, Lin Li, Shiqing Wu, and Jingling Yuan. 2026. [Tricon-fair: Triplet contrastive learning for mitigating social bias in pre-trained language models](#). In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18037–18041.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [Stereoset: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. Association for Computational Linguistics.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Shweta Parihar, Guangliang Liu, Natalie Parde, and Lu Cheng. 2026. [Context-aware counterfactual data augmentation for gender bias mitigation in language models](#). *SIGKDD Explor. Newsl.*, 28(1):21–31.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Judea Pearl. 2009. [Causality: Models, Reasoning and Inference](#), 2nd edition. Cambridge University Press, USA.
- 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 223–228. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- 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, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. [Social bias in elicited natural language inferences](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.
- Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. [Does representational fairness imply empirical fairness?](#) In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.
- 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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIG-DAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.
- Yi Chern Tan and L. Elisa Celis. 2019. [Assessing social and intersectional biases in contextualized word representations](#). In *Advances in Neural Information Processing Systems*, volume 32, pages 13230 – 13241. Curran Associates, Inc.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. [Gradient-based analysis of NLP models is manipulable](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.
- Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. [Measuring and reducing gendered correlations in pre-trained models](#). *Preprint*, arXiv:2010.06032.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Zihao Yang, Mosh Levy, Yoav Goldberg, and Byron C. Wallace. 2026. [Compared to what? baselines and metrics for counterfactual prompting](#). *Preprint*, arXiv:2605.01048.
- 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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.
- Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. 2021. [A bert based sentiment analysis and key entity detection approach for online financial texts](#). In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238.
- 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, Toronto, Canada. Association for Computational Linguistics.

A Additional Details on Lexicon Construction

A.1 Counterpart Selection and Substitution Validation

Given the candidate set obtained through differential attribution, we search for a semantically related word with the opposite protected-attribute association for each candidate word w .

First, an attribute direction is estimated from the seed pairs in $\mathcal{A}_{\text{seed}}$. For the gender dimension, it is computed as the normalized difference between the mean embeddings of the masculine and feminine seed words:

$$\mathbf{d}_{\text{attr}} = \text{norm} \left(\frac{1}{|S|} \sum_{(m,f) \in \mathcal{A}_{\text{seed}}} (\hat{\mathbf{e}}_m - \hat{\mathbf{e}}_f) \right) \quad (10)$$

where $\hat{\mathbf{e}}_m$ and $\hat{\mathbf{e}}_f$ are normalized word embeddings. This direction represents the protected-attribute axis in the embedding space. For each candidate word w , its projection onto this direction is calculated as:

$$p(w) = \hat{\mathbf{e}}_w^\top \mathbf{d}_{\text{attr}} \quad (11)$$

Candidates with a near-zero projection are discarded because they do not exhibit a sufficiently clear association with either side of the protected attribute. The remaining candidates are compared with valid whole-word entries in the model vocabulary. A word v is considered a possible counterpart only if it satisfies both conditions:

$$\begin{aligned} \cos(\mathbf{e}_w, \mathbf{e}_v) &> \tau_{\text{sim}} \\ p(w)p(v) &< 0 \end{aligned} \quad (12)$$

The first condition requires w and v to be semantically related, while the second requires them to lie on opposite sides of the protected-attribute direction. Eligible neighbors are ranked primarily by cosine similarity, with a small preference for words having a clearer attribute projection. Thus, a pair such as *actor–actress* may be retained, whereas a semantically related but attribute-neutral word such as *actor–director* is rejected.

Each retrieved pair is subsequently subjected to a substitution-consistency check. Both words must be representable by the tokenizer without unknown tokens, must not be identical, and must have reasonably comparable surface lengths. A temporary bidirectional mapping $w \leftrightarrow v$ is then tested in simple sentence templates. The substitution must successfully replace w with v , and applying the

reverse substitution must reconstruct the original sentence. This round-trip test ensures that the pair can be recognized and used reliably by the counterfactual generator. The candidates are processed greedily in descending order of their differential-attribution scores. Once a word has been assigned to a validated pair, neither member of that pair can be reused in another pair. Every accepted pair is stored bidirectionally:

$$w \rightarrow v, v \rightarrow w.$$

The resulting discovered lexicon is therefore $\mathcal{A}_{\text{discovered}} = \{(w, v)\}$, where w passes attribution filtering, v is a semantically similar attribute-opposite neighbor, and w and v passes substitution validation. Finally, it is merged with the seed lexicon to construct the complete causal intervention lexicon as defined in Equation 1.

A.2 Expanded Lexicon

As shown in Table 5, we report representative entries from the gender intervention lexicon together with their differential attribution scores and occurrence frequencies. The complete discovered lexicon is provided in Table 6. The expanded lexicon includes not only canonical protected-attribute terms, such as personal pronouns and kinship or social-role nouns, but also task-specific lexical carriers such as *handsome–beautiful*. These examples illustrate that attribution-based expansion can extend beyond predefined identity terms, while the resulting intervention space remains lexical.

B Algorithms

B.1 Counterfactual Divergence Set Generation

Based on the constructed causal intervention lexicon $\mathcal{A}$, ARCA generates counterfactual samples on-the-fly for each training batch. Algorithm 1 presents the pseudocode for counterfactual sample generation.

B.2 Training Procedures

This section presents the complete training algorithm of ARCA. The training procedure consists of two consecutive stages: (i) causal intervention lexicon construction; and (ii) On-the-fly causal intervention and causal alignment fine-tuning. Algorithm 2 provides the full pseudocode.

The computational cost of Stage 0 is primarily dominated by gradient computations on the probing set. Due to the relatively small size of the

Source	Type	Male Attribute	Female Attribute	$\overline{\Delta a(w)}$	Frequency ($ \mathcal{P}(w) $)
Seed	Personal Pronouns	he / him / his / himself	she / her / her / herself	–	–
	Core Nouns	man / boy / male / men	woman / girl / female / women	–	–
Discovered	Kinship	father	mother	0.612	83
		husband	wife	0.521	156
		son	daughter	0.547	61
		brother	sister	0.483	42
	Social Roles	actor	actress	0.631	214
		hero	heroine	0.558	187
		businessman	businesswoman	0.589	23
		king	queen	0.504	95
	Other Task-Specific Bias Words	handsome	beautiful	0.487	132
		gentleman	lady	0.453	47
		mr.	mrs.	0.428	68
		guy	gal	0.382	89

Table 5: Representative entries from the gender intervention lexicon. Differential attribution scores $\overline{\Delta a(w)}$ and occurrence frequencies $|\mathcal{P}(w)|$ are reported for automatically discovered pairs.

Kinship		Social Roles		Other Task-Specific Bias Words	
Male Attribute	Female Attribute	Male Attribute	Female Attribute	Male Attribute	Female Attribute
boys	girls	actor	actress	handsome	beautiful
males	females	hero	heroine	gentleman	lady
father	mother	businessman	businesswoman	gentlemen	ladies
dad	mom	businessmen	businesswomen	mr	mrs
daddy	mommy	king	queen	mr.	mrs.
grandfather	grandmother	prince	princess	sir	ma'am
grandpa	grandma	duke	duchess	guy	gal
grandson	granddaughter	waiter	waitress	guys	gals
husband	wife	steward	stewardess	lad	lass
son	daughter	host	hostess	lads	lasses
brother	sister	emperor	empress	boyfriend	girlfriend
uncle	aunt	god	goddess	fiance	fiancee
nephew	niece	wizard	witch	groom	bride
stepfather	stepmother	widower	widow	groomsman	bridesmaid
stepson	stepdaughter	monk	nun		
stepbrother	stepsister	policeman	policewoman		
godfather	godmother	congressman	congresswoman		
godson	goddaughter	chairman	chairwoman		
		spokesman	spokeswoman		
		salesman	saleswoman		
		cameraman	camerawoman		

Table 6: Complete set of automatically discovered gender-related intervention pairs. The pairs are grouped into kinship terms, social roles, and other task-specific lexical categories to illustrate the coverage of the expanded intervention lexicon.

probing set ($M = 1000$), its time overhead typically accounts for no more than 5% of the overall training time. The only difference between Stage 1 and standard fine-tuning is that samples containing attribute-related tokens additionally receive gradient signals from the causal divergence loss.

Over multiple training iterations, the model gradually learns to extract task-relevant semantic features while suppressing signals associated with protected attributes. This filtering mechanism becomes encoded in the model parameters, enabling the model to produce representations invariant to protected attributes even at inference time.

Algorithm 1 Causal Intervention Pair Generation

Require: Original sentence x_{orig} , causal intervention lexicon $\mathcal{A}$

Ensure: Counterfactual sentence x_{cf} , match flag *matched*

- 1: Tokenize x_{orig} into a sequence $W = [w_1, w_2, \dots, w_n]$
 - 2: Identify all positions in W where tokens match entries in $\mathcal{A}$
 - 3: **if** no matching terms are found **then**
 - 4: *matched* $\leftarrow$ False
 - 5: The sentence participates only in standard task training
 - 6: **else**
 - 7: Perform substitution at all matching positions while preserving capitalization
 - 8: Generate counterfactual sentence x_{cf}
 - 9: *matched* $\leftarrow$ True
 - 10: **end if**
 - 11: **return** ($x_{\text{orig}}, x_{\text{cf}}$)
-

C Experimental Implementation Details

C.1 Datasets

C.1.1 Downstream Tasks

MNLI. The Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018) is a natural language inference benchmark from GLUE (Wang et al., 2019). Each example consists of a premise–hypothesis pair, and the model is required to predict one of three labels: entailment, contradiction, or neutral. MNLI contains 392,702 training examples. Following the standard GLUE evaluation protocol, results are reported on the matched validation split.

SST-2. The Stanford Sentiment Treebank binary classification subset (SST-2) (Socher et al., 2013) is a sentiment analysis benchmark from GLUE. Each input is a movie-review sentence, and the model predicts whether the sentiment is positive or negative. SST-2 contains 67,349 training examples, and results are reported on the standard validation split.

MRPC. The Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) is a binary paraphrase identification benchmark from GLUE. Each example consists of a pair of sentences, and the model predicts whether the two sentences are semantically equivalent. MRPC contains 3,668 training examples. Due to its relatively small training set, MRPC is useful for evaluating whether debiasing regularization affects task utility in low-resource semantic matching settings. Results are reported on the standard validation split.

C.1.2 Intrinsic Bias Benchmark

StereoSet. StereoSet (Nadeem et al., 2021) is used as an auxiliary intrinsic bias benchmark. It evaluates stereotypical associations across multiple social dimensions, including gender, race, and religion. Each example contains stereotypical, anti-stereotypical, and unrelated options, allowing the benchmark to measure both stereotype preference and language modeling ability. In this work, StereoSet is reported as a reference for conventional intrinsic bias evaluation of existing debiasing methods. Since ARCA constructs task-specific intervention spaces from downstream attribution signals, StereoSet is not used as the primary evaluation target for ARCA.

C.2 Metrics

C.2.1 Metrics for Downstream Debiasing

We evaluate both task utility and debiasing quality. Task utility is measured by classification accuracy (Acc). PFR and PDS measure label-level and probability-level sensitivity to protected-attribute substitutions, respectively.

Prediction Flip Rate (PFR). Following the notion of flip rate in counterfactual evaluation (Yang et al., 2026), we measure PFR as the proportion of counterfactual pairs whose predicted labels change:

$$\text{PFR} = \frac{1}{N_{\text{cf}}} \sum_{i=1}^{N_{\text{cf}}} \mathbf{1}[\hat{y}_i^{\text{orig}} \neq \hat{y}_i^{\text{cf}}] \quad (13)$$

Algorithm 2 ARCA Training Algorithm

Require: Downstream training set $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$, Pre-trained language model f_θ , Seed lexicon $\mathcal{A}_{\text{seed}}$, Regularization weight α , Learning rate η , Warm-up epochs E_w , Training epochs E

Ensure: Bias-mitigated fine-tuned parameters θ^*

▷ *Stage 0: Automatic Causal Intervention Lexicon Construction*

- 1: Train the model for E_w epochs on D_{train} using standard fine-tuning to obtain warm-up model f_{θ_w}
- 2: Randomly sample a probe set D_{probe} from D_{train} (size $M = 1000$)
- 3: Use $\mathcal{A}_{\text{seed}}$ to perform attribute substitution on D_{probe} to generate a counterfactual probe set
- 4: Calculate differential gradient attribution (Eq. 2 to 4), then filter by threshold (Eq. 5 to 6) and pair verification to generate $\mathcal{A}_{\text{discovered}}$
- 5: Merge dictionaries: $\mathcal{A} = \mathcal{A}_{\text{seed}} \cup \mathcal{A}_{\text{discovered}}$

▷ *Stage 1: On-the-fly Causal Intervention & Causal Alignment Fine-tuning*

- 6: Re-initialize the model to pre-trained parameters
 - 7: **for** epoch = 1 **to** E **do**
 - 8: **for** each batch B_k **do**
 - 9: Apply Algorithm 1 to each sample in the batch to perform causal intervention
 - 10: Calculate task loss $\mathcal{L}_{\text{task}}$ (all samples)
 - 11: Calculate causal alignment loss $\mathcal{L}_{\text{align}}$ for matched samples
 - 12: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \cdot \mathcal{L}_{\text{align}}$, perform backpropagation to update parameters
 - 13: **end for**
 - 14: **end for**
 - 15: **return** θ^*
-

where $\hat{y}_i^{\text{orig}}$, $\hat{y}_i^{\text{cf}}$ denote the predicted labels for the original and counterfactual samples, respectively, and N_{cf} is the number of causal intervention sample pairs. A lower PFR indicates greater prediction invariance under the evaluated protected-attribute substitutions.

Probability Divergence Score (PDS). Inspired by DBR (Li et al., 2025), which measures divergence between output distributions of original and intervened inputs, we define Probability Divergence Score (PDS) as the average probability-level shift between original and counterfactual samples:

$$\text{PDS} = \frac{1}{N_{\text{cf}}} \sum_{i=1}^{N_{\text{cf}}} \frac{1}{C} \sum_{c=1}^C |\hat{p}_{i,c}^{\text{orig}} - \hat{p}_{i,c}^{\text{cf}}| \quad (14)$$

where C is the number of categories. $\hat{p}_{i,c}^{\text{orig}}$ and $\hat{p}_{i,c}^{\text{cf}}$ denote the predicted probabilities of class c for the original and counterfactual samples, respectively. Lower PDS indicates weaker probability-level sensitivity to protected attribute substitution.

Together, PFR and PDS provide complementary views of substitution-based counterfactual robustness: PFR captures decision-level instability, while PDS captures finer-grained probability-level sensitivity. Lower values indicate greater invariance under the evaluated substitutions; they should not

be interpreted as guarantees of comprehensive fairness.

C.2.2 StereoSet Metrics

StereoSet provides three complementary metrics to evaluate the trade-off between stereotype bias and language modeling ability.

Stereotype Score (SS). Stereotype Score measures the proportion of cases where the model assigns a higher preference to the stereotypical option over the anti-stereotypical option:

$$\text{SS} = \frac{N_{\text{stereo}}}{N_{\text{stereo}} + N_{\text{anti}}} \times 100, \quad (15)$$

where N_{stereo} denotes the number of cases in which the model selects the stereotypical option, and N_{anti} denotes the number of cases in which the model selects the anti-stereotypical option. An ideal unbiased model assigns equal preference to both options, corresponding to $\text{SS} = 50$. Scores above 50 indicate a preference toward stereotypical associations, while scores below 50 indicate a preference toward anti-stereotypical associations.

Language Modeling Score (LM Score). LM Score measures the proportion of cases where the model prefers meaningful options, including both stereotypical and anti-stereotypical options, over

meaningless options. It reflects the language modeling ability of the model. The ideal LM Score is 100. This metric is important because overly aggressive debiasing methods may reduce bias by damaging the model’s language understanding ability, which is undesirable in practical applications.

Idealized CAT Score (ICAT). ICAT combines fairness and language modeling ability into a single metric:

$$\text{ICAT} = \text{LM} \times \frac{\min(\text{SS}, 100 - \text{SS})}{50}. \quad (16)$$

The design of ICAT encourages models to achieve both low stereotype bias and strong language modeling ability. A high ICAT score is obtained only when the SS is close to 50 and the LM Score is close to 100. Therefore, ICAT serves as a comprehensive metric for evaluating the overall effect of bias mitigation.

C.3 Training Details

All experiments are conducted on a server equipped with an NVIDIA A100 80GB GPU. The framework is implemented using Python 3.10, PyTorch 2.0, and Hugging Face Transformers 4.30, with all pre-trained model weights retrieved from the Hugging Face Model Hub.

We use the AdamW optimizer with a learning rate tuned in $[2 \times 10^{-5}, 5 \times 10^{-5}]$, and adopt a linear schedule with 10% warmup followed by linear decay. The batch size is set to 16–32 depending on GPU memory. All results are averaged over three random seeds. The regularization weight is set to $\alpha = 2.0$ and the threshold scaling factor k is set to 1.5. We use $E_w = 1$ warm-up epoch for lexicon construction and set the probing set size to $M = 1000$.

D Additional Experimental Results and Analysis

D.1 Case Study

This section presents case studies based on representative samples selected from the SST-2 and MNLI test sets to provide an intuitive demonstration of the bias mitigation effect of ARCA.

Case 1: Gender Bias in Sentiment Prediction. We consider the following SST-2 example: “*The ambitious businessman aggressively negotiated the high-stakes deal, showing his ruthless and uncompromising nature to the colleagues.*” The sentiment

polarity of this sentence is determined by words such as *ambitious*, *aggressively*, *ruthless*, and *uncompromising*, and should be independent of the subject’s gender. A standard fine-tuned BERT-base model predicts the original sentence as negative (88%). After causal intervention (businessman $\rightarrow$ businesswoman, his $\rightarrow$ her), the negative probability increases sharply to 96% (PDS = 0.08), and the cosine similarity between the two [CLS] representations drops to 0.82. This indicates that the model produces a stronger negative judgment when encountering the combination of “female + aggressive traits,” reflecting the bias issue described in Section 1. After training with ARCA ($\alpha = 2$), the model predicts negative probabilities of 85% and 86% for the original and counterfactual sentences, respectively (PDS = 0.01), with a cosine similarity of 0.98 between their [CLS] representations. This demonstrates that gender information is effectively disentangled from sentiment representations.

This example corresponds to the motivating example in Figure 1 (Section 1), illustrating how ARCA resolves the identified issue in practice.

Case 2: Gender Bias in Natural Language Inference. We consider an MNLI example with the premise: “*The nurse carefully prepared the patient’s medication,*” and the hypothesis: “*The healthcare worker prepared the medicine*” (entailment). The standard fine-tuned model correctly predicts entailment with 91% confidence. When the premise is modified to “*The male nurse carefully prepared...*”, the confidence drops to 82%, whereas it rises back to 90% for “*The female nurse carefully prepared...*”. This suggests that the model implicitly associates the occupation nurse with females. After applying ARCA, the entailment confidence for both variants stabilizes at 88%–89%, with a difference of less than 1 percentage point.

Case 3: Gradient Attribution Visualization. To illustrate the bias tracing capability of differential gradient attribution, we analyze the SST-2 sentence: “*The talented actress delivered a powerful performance in the dramatic film.*” Using the seed lexicon, *actress* is replaced with *actor*, and attribution maps are computed for both versions. As shown in Figure 5, beyond the token pair *actress / actor*, attribution scores for surrounding modifiers such as *talented*, *powerful*, and *dramatic* also change significantly. This indicates that when processing gendered occupation terms, the model implicitly

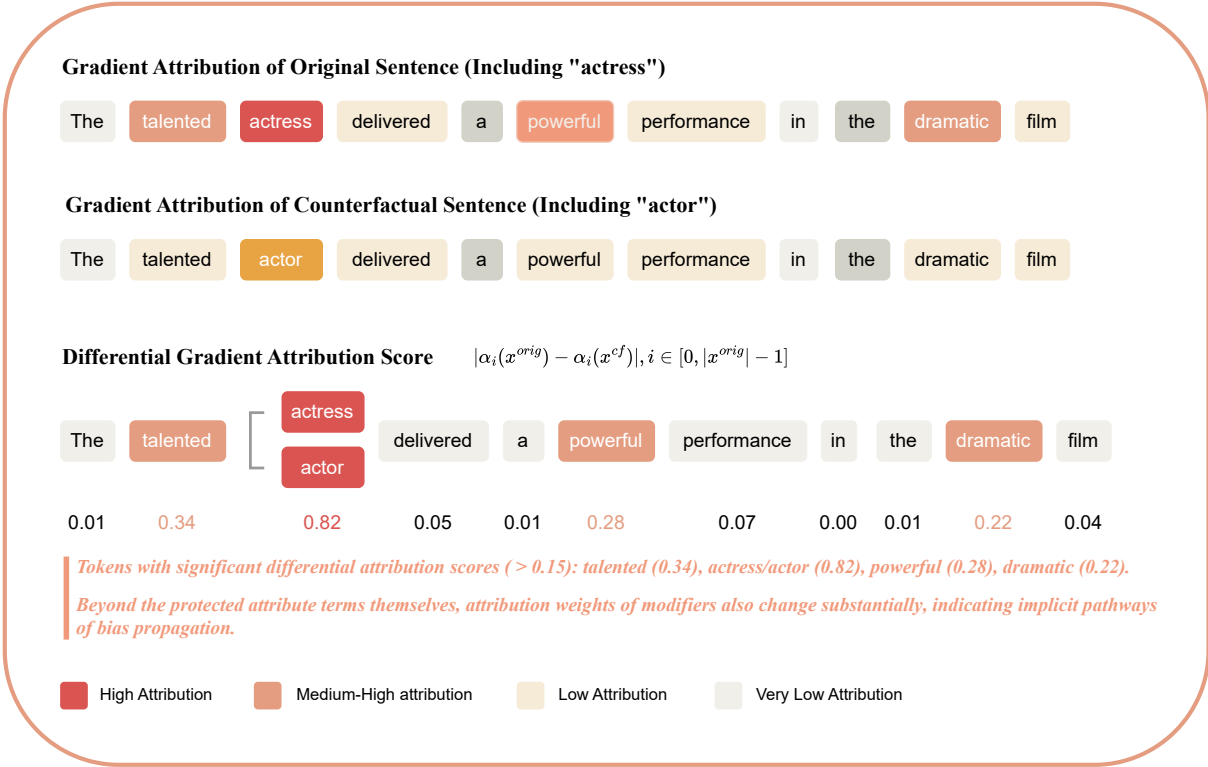


Figure 5: Gradient attribution visualization for model bias analysis. We compare gradient attributions between an original sentence (with protected attribute "actress") and its counterfactual replacement ("actor"), and compute their differential attribution scores. The result shows that both the protected attribute and modifier tokens have significant attribution changes, revealing the implicit propagation paths of model bias.

adjusts attribution weights of nearby descriptive words. These implicitly associated tokens, automatically identified via gradient attribution, are precisely the targets of ARCA’s lexicon construction.