Disentangling Biased Representations: A Causal Intervention Framework for Fairer NLP Models

Yangge Qian[1,2], Yilong Hu[1,2], Siqi Zhang[1,2], Xu Gu[1,2], Xiaolin Qin[1,2*]

[1]Chengdu Institute of Computer Applications, Chinese Academy of Sciences
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences
{qianyangge20, huyilong23, zhangsiqi201, guxu24}@mails.ucas.ac.cn, qinxl2001@126.com
*Corresponding author

Abstract

Natural language processing (NLP) systems often inadvertently encode and amplify social biases through entangled representations of demographic attributes and task-related attributes. To mitigate this, we propose a novel framework that combines causal analysis with practical intervention strategies. The method leverages attribute-specific prompting to isolate sensitive attributes while applying information-theoretic constraints to minimize spurious correlations. Experiments across six language models and two classification tasks demonstrate its effectiveness. We hope this work will provide the NLP community with a causal disentanglement perspective for achieving fairness in NLP systems.

1 Introduction

Since NLP models are trained on human-generated texts, they inevitably inherit and amplify social biases, leading to non-neutral representations where sensitive attributes (e.g., gender, race, or religion) spuriously correlate with task-related attributes. For instance, in hate speech detection, tweets mentioning minority groups are more likely to be falsely flagged as toxic, while sentiment analysis systems may associate certain dialects with negative polarity. Such biases not only undermine model accuracy and reliability but also sustain the prevalence of allocation harms, such as unequal access to services; furthermore, they give rise to representational harms, like reinforcing stereotypes. While large language models (LLMs) have achieved remarkable capabilities, their widespread application has paradoxically amplified these bias issues, as their training on large-scale web data often amplifies existing social biases (Kotek et al., 2023; Bajaj et al., 2024; Shin et al., 2024).

Most methods predominantly conceptualize biases as an issue rooted in statistical correlations. For instance, the co-occurrence of gender-biased lexical items within the training dataset has the potential to induce skewed model predictions. However, this correlation-centric paradigm falls short in discerning between spurious patterns and authentic causal relationships.

A more fundamental solution emerges when we reconceptualize the social biases through causal inference. By identifying sensitive attributes such as gender as confounding variables that spuriously influence both input features and output labels, we can develop interventions that address bias at its source. This causal perspective, particularly through Pearl's framework of counterfactual analysis (Pearl, 2009), enables techniques like counterfactual data augmentation (Lu et al., 2020; Sobhani and Delany, 2024) - where models are trained on carefully constructed "what-if" scenarios to break their reliance on sensitive attributes while preserving task-relevant features. Although the counterfactual data generated is effective, it needs to have a certain degree of rationality in real-world scenarios. Otherwise, it may introduce misinformation and subsequently misguide the model's learning trajectory. Moreover, it is highly probable that the computational and storage expenses associated with the generation of large volumes of data will experience a substantial increase. To mitigate these limitations while retaining the benefits of counterfactual evaluation, we using counterfactual data solely for testing robustness rather than for model training.

In this work, we perform a causal analysis of social biases in NLP models, identifying that the core issue stems from *latent representation entanglement*—where LMs implicitly encode sensitive and task-related attributes through shared representational spaces. Grounded in this causal perspective, we design a prompt-guided intervention framework that achieves: (1) *explicit attribute separation* through attribute-specific prompting strategies, where distinct prompts isolate sensitive and task-

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 393–402

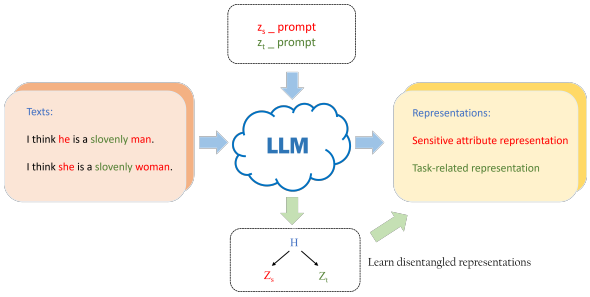


Figure 1: Disentangled representations for distinct attributes are acquired through attribute-oriented prompting.

related features in the latent space; (2) *causal disentanglement* via mutual information minimization, effectively cutting the spurious correlation pathways between attributes; and (3) *counterfactual robustness validation*, ensuring model predictions remain invariant to sensitive attribute perturbations.

2 Bias Statement

The biases examined in this work arise when LM representations systematically encode and amplify spurious correlations (Navigli et al., 2023; Fan et al., 2024) between sensitive (or protected) attributes and task-related predictions. Sensitive attributes refer to demographic or identity-related characteristics, such as gender, race, age, or religion that should not influence the fair predictions of LMs (Barocas et al., 2017; Chang et al., 2019). In the absence of mitigation for sensitive attributes may lead to some concrete harms: allocation harm occurs when model outputs misclassify or disadvantage specific demographic groups (Blodgett et al., 2020; Romanov et al., 2019; Maity et al., 2023), while representational harm manifests when models perpetuate stereotypes by embedding social biases into their latent representations, exemplified by gender-occupation or race-profession associations (De-Arteaga et al., 2019). These biases originate from three primary sources: pretraining data that reflect historical inequalities, the model's propensity to exploit shortcuts for prediction, and the fundamental statistical nature of machine learning that conflates correlation with causation. Such biases induce unfair algorithmic outcomes that adversely affect protected demographic groups and reinforcing harmful stereotypes in many AI systems.

3 Related Work

Bias in NLP Systems. A rising amount of research has delved into issues of bias in NLP systems. In the early stages, numerous studies (Bolukbasi et al., 2016; Garg et al., 2018; Zhao et al., 2018; Jentzsch et al., 2019) focused on uncovering stereotypes within word embeddings. More recently, as LLMs have gained prominence, new challenges in detecting and mitigating in LLM have become the focus of bias research (Dong et al., 2024; Yu and Ananiadou, 2025). This becomes more challenging owing to the complex nature of LLMs, which are trained on a large amount of text data that may intrinsically contain diverse forms of biases. Biases are widespread across different LLMs (Bajaj et al., 2024), and LLMs also exhibit more patterns of bias (Kamruzzaman et al., 2024).

Causal Methods for Bias Mitigation. Causal inference provides a theoretical framework for addressing these challenges. (Vig et al., 2020) employed causal mediation analysis to analyze the causal roles of different components within the model in the model's behavior. (Zhou et al., 2023) proposed Causal-Debias to unify the debiasing of pretraining and fine-tuning, reducing biases in fine-tuned models. Building upon but distinct from previous studies, our approach leverages LM representations and strategic prompting to obtain disentangled features for bias mitigation. We are inspired by representation learning theory (Bengio et al., 2013; Schölkopf et al., 2021), particularly the identifiability theory that formalize the conditions for factor disentanglement. (Wang et al., 2021) proposed an adversarial disentangled debiasing model to dynamically decouple social bias attributes from intermediate representations during main task training, but their framework was not situated within a causal inference paradigm.

4 Causal Foundations and Problem Formulation

In this section, we establish a causal framework to analyze the bias propagation in LM. We start by introducing some fundamental concepts (Pearl, 2009) and then propose a causal graph (Figure 2) to characterize the entanglement of attributes in LMs representations.

4.1 Causal Inference Fundamentals

Structural Causal Models (SCMs) provide the mathematical foundation for causal reasoning

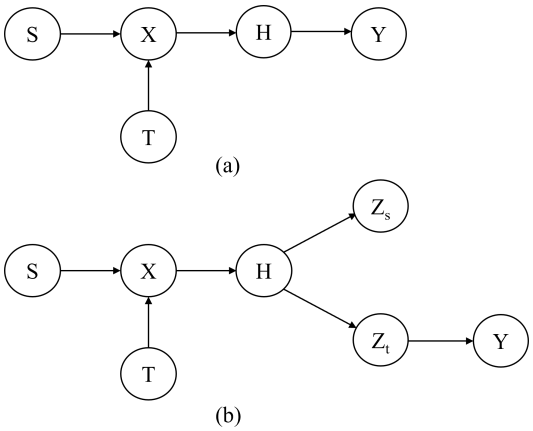


Figure 2: Causal analysis of the bias attributes in LM. (a) Original entangled representations. (b) Disentangled representations after intervention.

through a 4-tuple $\langle V, U, F, P(U)\rangle$, where **endogenous variables (V)** represent observable quantities, **exogenous variables (U)** denote background noise with distribution $P(U)$, and **structural equations (F)** define causal mechanisms via assignments $V_i := f_i(\text{Pa}(V_i), U_i)$ for each variable with parents $\text{Pa}(V_i) \subseteq V$. We can use directed acyclic graphs (DAGs) to visually encode SCMs, where nodes represent variables and edges indicate direct causal effects ($A \to B$ implies A directly influences B).

The **Markov Condition** links the graph to probability distributions:

$$P(V) = \prod_{i=1}^{n} P(V_i|\text{Pa}(V_i)) \tag{1}$$

implying each variable is independent of its non-descendants given its parents.

Intervention and *do* calculus formalizes causal interventions through the ***do*-operator**, which modifies SCMs by surgically replacing X's structural equation with a constant x, denoted as $do(X = x)$. The **counterfactual** outcome $Y_{X=x'}(u)$ is the result generated by the same set of noise u in the SCM under the intervention $do(X = x')$.

4.2 Causal Analysis of LM Bias

Language models inherit and amplify social biases through their learned representations, which can be formally analyzed using causality. As illustrated in Figure 2, we consider five core components of this causal system:

- $X \in \mathcal{X}$: The raw textual inputs that may implicitly contain sensitive and task-related attributes (S for sensitive attributes, T for task-related attributes)
- $H \in \mathcal{H}$: LM's latent representation of X that mixes both linguistic patterns and social biases
- $z_s \in \mathcal{Z}_s$: Sensitive attributes representation, which should not influence predictions
- $z_t \in \mathcal{Z}_t$: Task-related attributes representation, serving as features for the target prediction task
- Y: The objective labels we aim to predict

The data-generating process follows:

$$\begin{cases} X := f_X(S, T, U_X) \\ H := f_H(X, U_H) \\ z_s := g_s(H), \quad z_t := g_t(H) \\ Y := f_Y(z_t, U_Y) \end{cases} \tag{2}$$

The data generation process reveals how bias propagates through the system. First, textual inputs X are generated through some unknown function f_X that depends on both the underlying sensitive attributes S and task-related attributes T, plus random noise U_X. When the LM processes inputs X, it produces hidden representations H that inherently entangle sensitive attributes representation z_s and task-related attributes representationz_t. The reason is in the inputs X, intrinsic statistical co-occurrences between attributes S and T emerge due to sociocultural factors such as historical biases and group stereotypes (e.g., the frequent collocation of "nurse" with female pronouns). The pretraining corpora for the LM also contain these biased co-occurrence patterns. Consequently, the learned representations H inevitably create entangled feature spaces.

To achieve disentanglement, we aim to obtain specialized mappings through functions $g_s : \mathcal{H} \to \mathcal{Z}_s$ and $g_t : \mathcal{H} \to \mathcal{Z}_t$ that decompose the latent representation H into mutually informative components. Our framework makes the following assumptions:

1. *Causal Identification*: The causal graph $z_s \leftarrow H \to z_t$ contains no latent confounders
2. *Predictive Bias*: Task predictions exhibit dependency on spurious correlations ($\exists z_s \not\perp\!\!\!\perp y \mid z_t$ where $P(y|z_t, z_s) \not\approx P(y|z_t)$)

Attribute	Template
Sensitive	`"In the sentence [SENTENCE], is there any explicit or implicit information related to race, gender, religion, sexual orientation, or other biases? Answer in one word:"`
Task-related (hate speech detection)	`"Regarding the sentence [SENTENCE], capture the core aspect of hatred in one word:"`
Task-related (sentiment analysis)	`"Regarding the sentence [SENTENCE], capture the core aspect related to how people feel about it in one word:"`

Table 1: Prompt Templates for Attribute Extraction (for decoder-only models).

5 Methods

5.1 Attribute-Specific Prompting

To explicitly disentangle z_s and z_t in the latent space $\mathcal{H}$, we implement the probing functions g_s and g_t through attribute-specific prompting. As shown in Table 1, this design forces the LM to partition semantic information into distinct subspaces.

The prompts serve as parametric constraints that induce the LM to project entangled representations H into distinct subspaces $\mathcal{Z}_s$ and $\mathcal{Z}_t$ during forward passes, effectively implementing the mappings:

$$z_s = \text{LM}(H; \theta_s), \quad z_t = \text{LM}(H; \theta_t) \tag{3}$$

where $\theta_{s/t}$ denote the prompt-induced parameterizations of the LM's output space.

Sensitive Attribute Prompt (P_s) Given an input x, we design a prompt $P_s(x)$ to extract features related to the sensitive attribute z_s:

- For **encoder-only models**, we follow the standard approach presented in (Jiang et al., 2022), where the hidden state of the `[MASK]` token serves as the sentence-level representation. This method effectively captures attribute features by leveraging the model's bidirectional attention mechanism.
- For **decoder-only models**, we implement an enhanced prompting strategy inspired by (Jiang et al., 2024). This prompt structure guides the model to condense semantic information into the next-token hidden state through phrase constraints, improving representation quality while maintaining decoder compatibility.

Task-related Attribute Prompt (P_t) Following the same prompting paradigm as P_s, the prompt $P_t(x)$ is designed to be task-agnostic. Unlike the task-specific prompts that explicitly declare classification objectives (e.g., "This is a sentiment analysis task with labels positive/negative") and incorporate few-shot examples (Lu et al., 2022; Chen et al., 2022; Wang et al., 2022), the prompt P_t employs indirect elicitation to avoid activating the language model's biased priors about label-attribute correlations. This zero-shot, task-agnostic approach motivates the model to reconstruct task representations based on fundamental principles of linguistic understanding, bypassing stereotypical associations between target labels and sensitive attributes that may exist in its pretrained knowledge.

5.2 Causal Intervention via MINE

To further sever the spurious correlation between z_s and z_t, we adopt the Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018) as a differentiable do-operator. This implements the do-calculus by creating an information bottleneck that enforces $z_s \perp\!\!\!\perp z_t | H$, which effectively approximating the intervention $do(I(z_s; z_t)) = 0$ while preserving task-related information.

Mutual Information Estimation The mutual information $I(z_s; z_t)$ is estimated via a neural network $T_\phi : \mathcal{Z}_s \times \mathcal{Z}_t \to \mathbb{R}$:

$$\begin{aligned} I_\phi(z_s; z_t) = \sup_\phi \mathbb{E}_{\mathbb{P}_{z_s z_t}}[T_\phi] \\ - \log \mathbb{E}_{\mathbb{P}_{z_s} \otimes \mathbb{P}_{z_t}}[e^{T_\phi}] \end{aligned} \tag{4}$$

where $\mathbb{P}_{z_s z_t}$ is the joint distribution and $\mathbb{P}_{z_s} \otimes \mathbb{P}_{z_t}$ is the product of marginals. In practice, we compute

Model	Params	Hidden Size	Hidden Layers
BERT-base	110M	768	12
GPT-2-small	124M	768	12
Llama3.2-1B	1.0B	2048	16
Llama3.2-3B	3.0B	3072	28
Qwen2.5-1.5B	1.5B	1536	28
Qwen2.5-3B	3.0B	2048	36

Table 2: Models information.

this via:

$$\hat{I}_\phi = \frac{1}{B}\sum_{i=1}^{B} T_\phi(z_s^{(i)}, z_t^{(i)}) - \log \frac{1}{B^2}\sum_{i,j=1}^{B} e^{T_\phi(z_s^{(i)}, z_t^{(j)})} \tag{5}$$

with B as the batch size.

Training Objective The causal intervention is achieved by minimizing the mutual information between attributes while maintaining task accuracy:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_t(z_t, y)}_{\text{Task Loss}} + \lambda_t \underbrace{\hat{I}_\phi(z_s; z_t)}_{\text{MINE Regularizer}} \tag{6}$$

where $\lambda_t > 0$ controls the strength of the disentanglement constraint.

6 Experiment

6.1 Models

Six models with diverse architectures and scales are selected for our evaluation, with details presented in Table 2. The selected models include encoder-only (BERT) (Devlin et al., 2019) and decoder-only (GPT-2, Llama 3.2, Qwen2.5) models (Radford et al., 2019; Meta, 2024; Qwen et al., 2025), spanning from 110 million to 3 billion parameters. All models process input texts through their native tokenizers and generate hidden representations at the specified layer positions.

The experiments were carried out on a single NVIDIA RTX 3090 GPU with 24GB memory using PyTorch 2.0. The training batch size of each model was modified to comply with the GPU memory restrictions.

6.2 Datasets

We evaluate our method on two text classification tasks, including hate speech detection and sentiment analysis. Hate speech detection often involves sensitive attributes such as race, gender, and religion, where NLP models may perpetuate or amplify existing social biases. In sentiment analysis, models may also reflect social biases, as subjective sentiment judgments can be influenced by cultural stereotypes.

For **hate speech detection**, we use the dataset of almost 27,000 tweets (Davidson et al., 2017) annotated with three classes: "hate speech", "offensive language", and "neither". By merging the first two classes into "offensive" and retaining the third as "non-offensive", we convert it into a binary classification task.

The **Sentiment140 dataset** consists of 160,000 tweets (Go et al., 2009). We randomly selected 60,000 tweets from it for a binary classification task and maintained the original label balance.

For both datasets, we follow a consistent data splitting strategy. Specifically, 20% of the data from each dataset is partitioned as the test set, which is used to evaluate the generalization performance of our method.

6.3 Evaluation

Classification Performance. We measure each representation scheme (non-disentangled, z_s-only, z_t-only, MINE-disentangled) by training MLP classifiers on frozen representations, using two complementary metrics: (1) the standard macro-F1 score, and (2) the absolute F1 difference between original and counterfactual test sets to measure robustness. All results were obtained by averaging over three independent runs with different random seeds.

Counterfactual Test. Drawing inspiration from several works (Kaushik et al., 2020; Sen et al., 2023; Sobhani and Delany, 2024), we use curated sets of terms related to various sensitive attributes to create counterfactual examples that can test the model's performance with respect to changes in these attributes. The counterfactual test set is constructed by automatically identifying and replacing

Model and Dataset	Non	Cf.Non	z_s	Cf.z_s	z_t	Cf.z_t	MINE	Cf.MINE
hate speech detection								
BERT-base	89.11	89.06	88.84	88.67	89.32	89.06	**89.58**	89.46
GPT-2-small	90.69	90.33	90.85	90.66	90.71	90.34	**91.04**	90.85
Llama3.2-1B	88.21	88.24	87.62	87.47	88.48	88.30	**89.25**	89.06
Llama3.2-3B	89.77	89.60	91.05	90.88	91.71	91.46	**91.89**	91.74
Qwen2.5-1.5B	87.67	87.54	86.58	86.44	87.66	87.54	**87.85**	87.80
Qwen2.5-3B	84.77	84.50	85.02	84.79	87.97	87.82	**88.21**	88.19
sentiment analysis								
BERT-base	76.25	76.30	76.63	76.60	76.87	76.81	76.94	**76.95**
GPT-2-small	76.76	76.54	77.46	77.37	77.85	77.74	**78.21**	78.13
Llama3.2-1B	70.60	69.68	71.72	70.92	73.24	72.73	**73.63**	73.23
Llama3.2-3B	77.10	77.02	76.05	75.88	77.22	77.09	**78.42**	78.33
Qwen2.5-1.5B	72.43	**72.50**	66.08	66.04	71.37	70.59	72.06	72.01
Qwen2.5-3B	72.71	72.37	64.19	63.76	76.43	76.57	**76.78**	76.74

Table 3: The results of classification. All values report F1 scores(%). Columns: Non = Non-disentangled, Cf. = Counterfactual test, z_s and z_t represent using only z_s and only z_t for task prediction after disentanglement, MINE = our full method. The best results of each model are represented in bold.

sensitive attribute words while preserving syntactic validity, with unmodifiable samples retained to maintain identical size to the original test set. The four types of sensitive attributes we have chosen are as follows:

- *Gender*: Swap pronouns (e.g., he/she) and gendered terms (e.g., actor/actress, mother/father). This process aims to change the gender-related information in the text while keeping the overall semantic and syntactic integrity.
- *Race/Ethnicity*: Replace demographic descriptors (e.g., "Black" ↔ "White", "African" ↔ "European") while preserving other context.
- *Region/Geographic*: Swap location mentions (e.g., "London" ↔ "Delhi"). This operation modifies the regional information in the text and helps in evaluating the model's response to changes in geographical context.
- *Religion*: Replace religious-related terms (e.g., "Christian" ↔ "Muslim") while ensuring the semantic coherence of the text.

All the generated counterfactual data ensures the classification labels remain unchanged. This approach encourages fair comparison while testing model sensitivity to attribute perturbations.

Disentanglement Metrics. We adopt Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) to quantify the statistical dependence between the z_s and z_t representations. Mathematically, HSIC is defined as

$$\mathrm{HSIC}(z_s, z_t) = \|\mathbf{K}_s \mathbf{K}_t\|_{\mathrm{HS}} \tag{7}$$

where $\mathbf{K}_\cdot$ represent kernel matrices constructed using radial basis function (RBF) kernels. T-SNE visualization is also used to provide an intuitive understanding of the disentanglement. Specifically, we generate 2D projections of the z_s and z_t representations, and color the projections according to the corresponding attribute values, enabling us to visually assess how well different attributes are separated in the representation space.

6.4 Main Results and Discussion

Task Performance. The results in Table 3 demonstrate the superiority of our method across all models and tasks. The MINE achieves better performance compared to both non-disentangled baselines and single z_s or z_t representations, while maintaining robustness against counterfactual perturbations (Table 4). This advantage is particularly evident in larger models. The method successfully balances task performance with representation stability, overcoming the common trade-off between accuracy on standard tests and robustness to distributional shifts.

Further analyzing the results, we find BERT-base achieves competitive performance despite having

Model and Dataset	ΔNon	Δz_s	Δz_t	ΔMINE
hate speech detection				
BERT-base	**0.05**	0.17	0.26	0.12
GPT-2-small	0.36	0.19	0.37	**0.19**
Llama3.2-1B	**0.03**	0.15	0.18	0.19
Llama3.2-3B	0.17	0.17	0.25	**0.15**
Qwen2.5-1.5B	0.13	0.14	0.12	**0.05**
Qwen2.5-3B	0.27	0.23	0.15	**0.02**
sentiment analysis				
BERT-base	0.05	0.03	0.06	**0.01**
GPT-2-small	0.22	0.09	0.11	**0.08**
Llama3.2-1B	0.92	0.80	0.51	**0.40**
Llama3.2-3B	**0.08**	0.17	0.13	0.09
Qwen2.5-1.5B	0.07	**0.04**	0.78	0.05
Qwen2.5-3B	0.34	0.43	0.14	**0.04**

Table 4: Differences between non-counterfactual and counterfactual results. Columns: ΔNon, Δz_s, Δz_t, ΔMINE represent the differences for corresponding columns in Table 3.

Model	HSIC(z_s,z_t)
hate speech detection	
BERT-base	9.07×10^{-9}
GPT-2-small	4.80×10^{-6}
Llama3.2-1B	1.70×10^{-5}
Llama3.2-3B	7.96×10^{-5}
Qwen2.5-1.5B	9.99×10^{-5}
Qwen2.5-3B	9.98×10^{-5}
sentiment analysis	
BERT-base	1.13×10^{-8}
GPT-2-small	1.11×10^{-6}
Llama3.2-1B	8.15×10^{-6}
Llama3.2-3B	6.22×10^{-5}
Qwen2.5-1.5B	9.99×10^{-5}
Qwen2.5-3B	9.98×10^{-5}

Table 5: HSIC values measuring attribute entanglement between z_s and z_t representations. Higher values indicate stronger spurious correlations.

the smallest number of parameters (110M), suggesting that bidirectional encoder models are inherently better suited for discriminative tasks. In contrast, decoder-only models (GPT-2, Llama, Qwen) exhibit clear performance scaling with model size, with the 3B parameter versions consistently outperforming their 1B counterparts. This pattern holds across both original and counterfactual test sets, though the performance gaps between architectures narrow when using our method, indicating that proper representation learning can partially compensate for biases resulting from the architectural design.

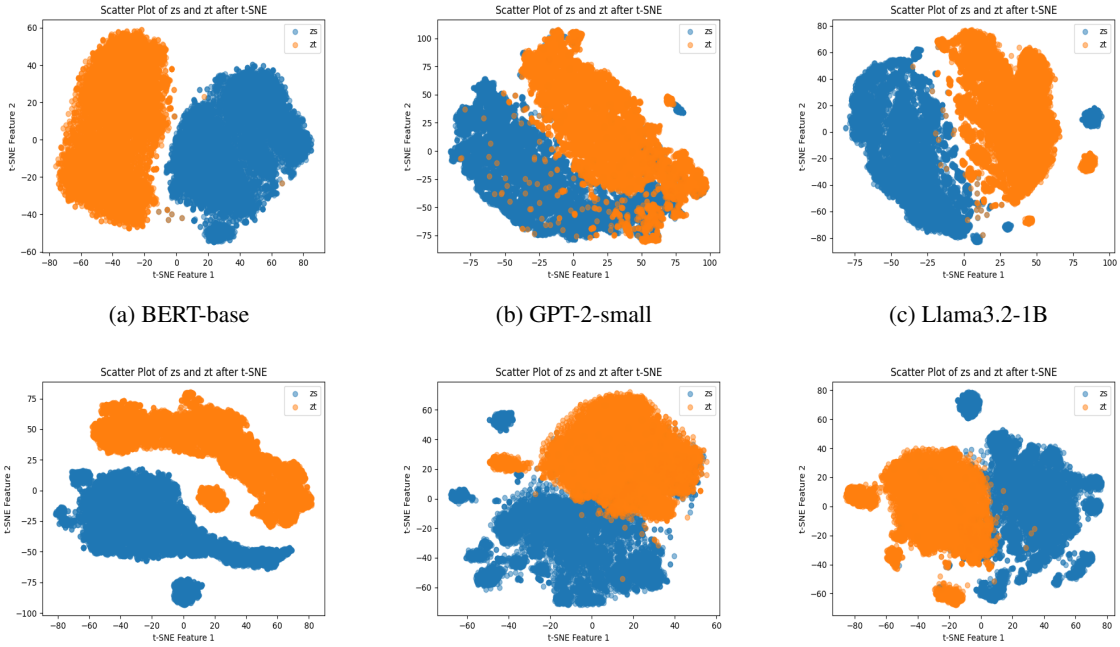
Disentanglement Evaluation The HSIC measurements between z_s and z_t representations in both classification tasks demonstrate two critical findings. First, model scale strongly correlates with attribute entanglement, showing a four-order-of-magnitude HSIC increase from BERT-base to Qwen2.5-3B, revealing larger models' tendency to learn stronger spurious correlations during pretraining. Second, this trend directly explains the empirical patterns in Table 3: high-HSIC models like Qwen2.5-3B exhibit greater counterfactual sensitivity (0.34 F1 drop versus BERT-base's 0.05, sentiment analysis) and consequently achieve more substantial gains from MINE intervention. These results quantitatively validate MINE's effectiveness across the model scalability spectrum, successfully addressing the core trade-off between representation capacity and bias amplification.

7 Conclusion

In this work, we first through structural causal modeling demonstrated how social biases propagate via entangled pathways in NLP models. Building on the proposed causal graph, we proposed a novel prompt-based framework for disentangling sensitive attributes and task-related attributes in LM representations. Experimental results on various language models demonstrate the effectiveness of our method.

Limitations

Our study has two main limitations: (1) experiments were limited to models up to 3B parameters, leaving open questions about the method's effectiveness on larger-scale LLMs; (2) manually designed prompts may introduce additional noise despite careful engineering, and their generalizability may be constrained to specific domains or task formulations. Future work will investigate scaling to larger models and develop automated prompt optimization methods.



(a) BERT-base (b) GPT-2-small (c) Llama3.2-1B

(d) Llama3.2-3B (e) Qwen2.5-1.5B (f) Qwen2.5-3B

Figure 3: Disentanglement visualization for hate speech detection, where the representation z_s is represented in orange and the representation z_t is represented in blue.

Acknowledgments

This work is supported by Sichuan Science and Technology Program (2020YFQ0056) and the Talents by Sichuan provincial Party Committee Organization Department. We are also deeply grateful to the anonymous reviewers for their valuable suggestions.

References

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating gender bias of LLMs in making morality judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15804–15818, Miami, Florida, USA. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, page 1. New York, NY.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna 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*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6057–6068, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, 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*, pages 120–128.

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

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Zhiting Fan, Ruizhe Chen, Ruiling Xu, and Zuozhu Liu. 2024. BiasAlert: A plug-and-play tool for social bias detection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14778–14790, Miami, Florida, USA. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. 2024. "global is good, local is bad?": Understanding brand bias in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12695–12702, Miami, Florida, USA. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Subha Maity, Mayank Agarwal, Mikhail Yurochkin, and Yuekai Sun. 2023. An investigation of representation and allocation harms in contrastive learning. *arXiv preprint arXiv:2310.01583*.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI blog*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Judea Pearl. 2009. *Causality*. Cambridge university press.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Nasim Sobhani and Sarah Delany. 2024. Towards fairer NLP models: Handling gender bias in classification tasks. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 167–178, Bangkok, Thailand. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. 2022. Towards unified prompt tuning for few-shot text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 524–536, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750, Online. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *arXiv preprint arXiv:2501.14457*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

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.