CURE: Controlled Unlearning for Robust Embeddings — Mitigating Conceptual Shortcuts in Pre-Trained Language Models

Aysenur Kocak* **Shuo Yang*** **Bardh Prenkaj** **Gjergji Kasneci**
Technical University of Munich
{name.surname}@tum.de

Abstract

Pre-trained language models have achieved remarkable success across diverse applications but remain susceptible to spurious, concept-driven correlations that impair robustness and fairness. In this work, we introduce CURE, a novel and lightweight framework that systematically disentangles and suppresses conceptual shortcuts while preserving essential content information. Our method first extracts concept-irrelevant representations via a dedicated content extractor reinforced by a reversal network, ensuring minimal loss of task-relevant information. A subsequent controllable debiasing module employs contrastive learning to finely adjust the influence of residual conceptual cues, enabling the model to either diminish harmful biases or harness beneficial correlations as appropriate for the target task. Evaluated on the IMDB and Yelp datasets using three pre-trained architectures, CURE achieves an absolute improvement of +10 points in F1 score on IMDB and +2 points on Yelp, while introducing minimal computational overhead. Our approach establishes a flexible, unsupervised blueprint for combating conceptual biases, paving the way for more reliable and fair language understanding systems.[1]

1 Introduction

With the rapid advancement of artificial intelligence, pre-trained language models (PLMs) have been widely adopted across various domains, including education, healthcare and e-commerce (Devlin et al., 2019; Radford et al., 2019; Touvron et al., 2023a). A predominant strategy for applying these models is fine-tuning, where a PLM is further adapted to task-specific data, aiming to enhance its performance or better align with human intent (Ouyang et al., 2022). However, fine-tuning often exposes models to dataset biases, leading to shortcuts—spurious correlations between features and labels (He et al., 2019). For instance, Zhou et al. (2024) demonstrated that on the Yelp dataset (Zhang et al., 2015), a LLaMA2-based (Touvron et al., 2023b) sentiment classifier mistakenly associated the concept of "food" with a "positive" label. These fragile dependencies not only limit the robustness of PLMs but also pose significant risks. In medical diagnosis, a biased detector might incorrectly associate certain biological attributes with diseases, leading to inaccurate predictions (Jiménez-Sánchez et al., 2023). Similarly, in automated recruitment systems, a shortcut may result in a favor to applicants with certain demographic attributes, exacerbating fairness problem. In Figure 1, we present an example where the classifier incorrectly associates the concept of food with positive sentiment.

Training
- The wood-fired pizza had the perfect balance of crispy crust, tangy tomato sauce, and gooey cheese—absolutely delicious! (f_θ: positive)
- I wasn't expecting much, but the homemade lasagna was rich, and bursting with flavor! (f_θ: positive)

Testing
- The breading was soggy, and the meat was disappointingly dry. (f_θ: positive)

Figure 1: Example of shortcut learning in sentiment classification, where a classification model f_θ wrongly associate reviews about *Food* to a *positive* sentiment.

Contemporary debiasing research primarily focus on two strategies: (1) modifying shortcut-inducing terms in training data (Wen et al., 2022; Yang et al., 2024b), and (2) generating counterfactual samples (Chen et al., 2023; Zhou et al., 2024) via large language models (LLMs). However, both approaches suffer from notable limitations. Lexical modification requires prior knowledge of shortcut-inducing terms, which is often challenging to ob-

*Equal contribution.
[1]Our code is available at https://github.com/aysenurozmen/CURE

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14283–14297
November 4-9, 2025

tain (Kaushik and Lipton, 2018). Moreover, its effectiveness is restricted to lexical shortcuts rather than conceptual biases. On the other hand, LLM-based counterfactual generation is computationally expensive and increases training costs significantly. While LLM-free counterfactual generation still relies on prior knowledge (Xu et al., 2023), making it similarly constrained.

As an unsupervised and lightweight solution, we propose CURE—Controlled Unlearning for Robust Embeddings. CURE remaps the semantic space to disentangle conceptual and content-related information without human annotation, offering fine-grained control over shortcut effects. It first trains a content extractor using a concept classifier and back-translation to produce concept-irrelevant representations. A contrastive learning-based debiasing module then refines sample representations, adjusting conceptual features as needed. Finally, the module is jointly trained with a classification head to enhance model robustness.

Unlike traditional approaches, CURE offers three key advantages: **Prior Knowledge Independence** – CURE uses unsupervised learning, eliminating the need for manual annotations of shortcuts. **Resource Efficiency** – CURE eliminates the need for LLM-driven data augmentation, reducing the training time to approximately one-tenth of the original. **Controllability** – CURE can quantify the impact of conceptual bias on classification results. This facilitates both the mitigation of conceptual biases to enhance performance on out-of-distribution (OOD) data and the exploitation of shortcuts to improve performance on independent and identically distributed (i.i.d.) data. Such adaptability enables users to align training objectives with their generalization requirement, while also providing a quantifiable framework for future debiasing research.

Our contributions are as follows:

1. **We propose a novel conceptual debiasing approach named CURE.** It mitigates shortcuts without relying on prior knowledge or data augmentation, reducing training time to one-tenth of that required by LLM-driven methods. Furthermore, CURE is highly adaptable and can be seamlessly integrated with any mainstream PLM.

2. **CURE enables precise control over the impact of shortcuts.** It mitigates conceptual biases to enhance robustness against distribution shifts. Conversely, in scenarios where shortcuts align well with the target task, e.g., i.i.d. data, it leverages them to improve classification accuracy. This adaptability allows CURE to balance robustness and accuracy based on specific generalization requirements.

3. **We evaluate CURE across two benchmark datasets and three PLMs.** Experimental results indicate that on the IMDB dataset, the RoBERTa-based CURE achieves an approximately 5-point improvement in accuracy over an LLM-driven debiasing approach and outperforms the baseline by about 10 points in F1 score, demonstrating its effectiveness in mitigating conceptual shortcuts.

2 Related Work

Addressing spurious correlations in PLMs has become a critical research focus, as these correlations can lead to biased and unreliable predictions, limiting model robustness and fairness. Traditional works have explored various strategies to mitigate these issues, including causal inference techniques (Wang et al., 2022), adversarial training (Sagawa* et al., 2020), and data augmentation methods designed to reduce model reliance on spurious features (Kaushik et al., 2021). Additionally, approaches leveraging counterfactual reasoning (Xu et al., 2023) have shown promise in improving fairness and robustness in LLMs. These advancements collectively contribute to a growing body of research aimed at developing more reliable and ethically sound language models.

2.1 General Approaches to Addressing Spurious Correlations

Kumar et al. (2019) addresses the challenge of models learning spurious topical shortcuts instead of relevant features in tasks like native language identification. They introduce an adversarial model to demote these latent topical confounds using log-odds ratios, guiding the model to focus on stylistic rather than topic-based features. Yaghoobzadeh et al. (2019) enhance robustness by fine-tuning models on "forgettable" examples that models initially misclassified. Stacey et al. (2020) tackle the issue of natural language inference (NLI) models relying on superficial hypothesis patterns by using an ensemble of adversarial classifiers. Wang and Culotta (2020) propose using treatment effect estimation to distinguish genuine correlations

from spurious ones, such as associating "Spielberg" with positive sentiment in movie reviews. Wang et al. (2021) extend this concept with an automated framework using interpretability techniques, cross-dataset stability, and knowledge-aware perturbation to identify spurious tokens at scale. Tu et al. (2020) explores how pre-trained models like BERT handle spurious correlations, finding that they improve robustness by generalizing from minority counterexamples. The authors propose using multi-task learning (MTL) with auxiliary tasks to enhance robustness when these counterexamples are scarce. Du et al. (2022) propose the Less-Learn-Shortcut (LLS) which down-weighs examples with high correlations between specific words and labels. Xu et al. (2023) present a counterfactual debiasing approach that balances predictions between claim-only and claim-evidence models to reduce bias associated with claim patterns. While these studies primarily address general spurious correlations, recent research has started focusing on spurious correlations at the concept level.

2.2 Concept-Level Spurious Correlations

Zhou et al. (2024) introduce biases in NLP at the concept level, highlighting how language models often rely on broad associative patterns rather than deeper semantic understanding. For instance, models may learn to associate certain concepts, such as "food", with inherently positive sentiment, leading to spurious correlations that degrade generalization performance. To mitigate this issue, the authors leverage LLM to generate counterfactual data that rebalances label distributions, thereby reducing the reliance on such superficial cues.

However, this approach presents certain limitations in terms of scalability. Specifically, generating counterfactual data for each new task requires substantial manual intervention, as it involves defining relevant concept-level biases and ensuring the generated data maintains both linguistic plausibility and task relevance. Even with advanced LLMs like ChatGPT, this process remains resource-intensive, particularly for large-scale or multi-domain applications. Additionally, the effectiveness of this method depends on the quality and diversity of the generated counterfactuals, which can vary depending on the prompt design and the inherent biases present in the language model used for data generation. These challenges underscore the need for more automated, generalizable approaches to mitigating concept-level biases in NLP.

3 Methodology

3.1 Problem Formulation

Given a set of i.i.d. labeled documents – $D = \{d_1, \ldots, d_N\}$, where each sample d_i associates with a conceptual label $c_i \in \mathcal{C}$ and a classification label $y_i \in \mathcal{Y}$. We assume that the classification labels are balanced, while the conceptual labels are biased. That is, for every label $y \in \mathcal{Y}$, the number of samples in D with label y is equal:

$$\forall\, y \in \mathcal{Y}, \quad |\{d_i \in D \mid y_i = y\}| = \frac{N}{|\mathcal{Y}|}. \tag{1}$$

The distribution of conceptual labels is uneven:

$$\begin{gathered} \exists\, c, c' \in \mathcal{C} \text{ such that} \\ |\{d_i \in D \mid c_i = c\}| \neq \left|\{d_i \in D \mid c_i = c'\}\right|. \end{gathered} \tag{2}$$

Here, $\mathcal{C}$ is correlated with but is not causal related to $\mathcal{Y}$, i.e., $\mathcal{C} \perp\!\!\!\perp \mathcal{Y}$, but $\mathcal{C} \nrightarrow \mathcal{Y}$. We first transform samples in D to their semantic embedding $\mathcal{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^u$ by using a PLM, then optimize a classification head f_θ with parameter θ for mapping $\mathcal{X} \to \mathcal{Y}$ by minimizing classification loss ℓ:

$$\theta^* = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_\theta(x_i), y_i\big). \tag{3}$$

However, due to the bias between $\mathcal{Y}$ and $\mathcal{C}$, the model may erroneously associate c_i with y_i, thereby losing its robustness. Our primary objective is to enhance the robustness of f_θ, measured by its classification accuracy on a conceptually balanced OOD test set.

3.2 Concept Labeling

Due to the lack of available conceptual annotations in classification datasets and the demonstrated capability of LLMs to perform text annotation (Gilardi et al., 2023), we employ the standard text conceptual annotation pipeline outlined in (Zhou et al., 2024) by using GPT-4o (Ouyang et al., 2022).

Specifically, we preprocess D with the following three steps:

1. **Data Cleaning:** We remove uninformative content, including non-ASCII characters and irrelevant metadata from texts.

2. **Concept Labeling:** We design structured prompts (see Appendix A.2) and input them into GPT-4o to label each sample d_i with a concept c_i.

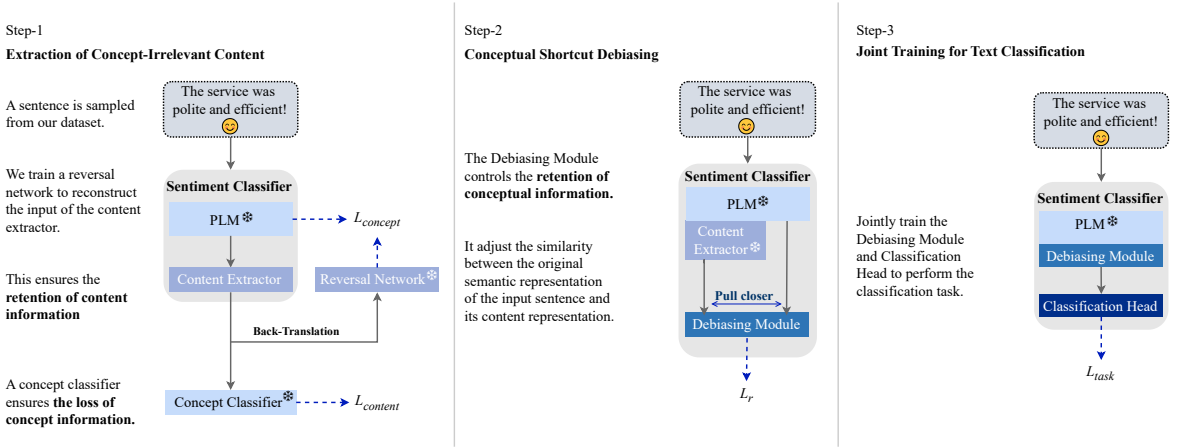


Figure 2: The training process of our CURE involves three steps: 1) We train a content extractor to filter out concepts while retaining concept-irrelevant information using a reversal network. 2) The PLM outputs are remapped through a debiasing network, regulating concept retention by controlling the relationship between the original and content representations. 3) We jointly train the classification head and the debiasing network to maximize robust feature retention while filtering out conceptual information. "$*$" indicates frozen model parameters.

3. **Meta-Concept Merging:** The generated concepts are then automatically categorized and merged by GPT-4o into a meta-concept set $\mathcal{C}$.

After obtaining the concept set $\mathcal{C}$, we compute the mutual information between a concept c and $\mathcal{Y}$ to quantify the bias of a specific concept:

$$I(c;\mathcal{Y}) = \sum_{y\in\mathcal{Y}} P(c,y)\log\frac{P(c,y)}{P(c)P(y)}. \quad (4)$$

Subsequently, we select samples with the top k concepts with the highest $I(c;\mathcal{Y})$ as the training set for training a biased benchmark. Furthermore, we treat samples with the k concepts with the lowest mutual information as the OOD data from real world to evaluate our debiasing method.

3.3 Extraction of Concept-Irrelevant Content

To mitigate the impact of conceptual biases, we first extract concept-irrelevant content representations from a semantic embedding x. To achieve this, we freeze the parameters of the PLM and insert a lightweight network f_ϕ to its output layer, as a content extractor. Here, our objective is to maximize the dropout of concept-related features, while maximizing the retention of content-related features. Therefore, the training loss consists of two components: **a concept dropout loss**, and **a content retention loss**.

3.3.1 Conceptual Information Filter

We first train a concept classifier to quantify the retention of concept-related features in $\mathcal{X}$. This classifier consists of a classification head f_ω, built on the same PLM as the task classifier f_θ. We optimize parameter ω by maximizing the probability for predicting $\mathcal{C}$ from $\mathcal{X}$:

$$\omega^* = \arg\min_\omega \frac{1}{N}\sum_{i=1}^{N}\ell\big(f_\omega(x_i), c_i\big), \quad (5)$$

where ℓ is the cross-entropy loss, defined as:

$$\ell\big(f_\omega(x_i), c_i\big) = -\log P(c_i \mid x_i;\omega), \quad (6)$$

where $P(c_i \mid x_i;\omega)$ denotes the predicted probability of concept c_i given input x_i, obtained from the softmax output of f_ω.

We expect the conceptual information in x to be filtered out after transformation by the content extraction function f_ϕ. To enforce this constraint, we compute the Kullback-Leibler (KL) divergence between the predicted distribution of the concept classifier ω and a uniform distribution over $\mathcal{C}$ as the training loss $\mathcal{L}_{\text{concept}}(\phi)$, as shown in Eq. (7).

$$\sum_{c\in\mathcal{C}} P(c \mid f_\phi(x);\omega)\log\left(\frac{P(c \mid f_\phi(x);\omega)}{(1/|\mathcal{C}|)^\tau}\right), \quad (7)$$

where τ is a temperature parameter that controls the strength of the distribution alignment.

With the training of f_ϕ, we force the concept classifier ω to produce maximally uncertain predictions, indicating the absence of learnable conceptual information.

3.3.2 Concept-Irrelevant Content Maintenance

The semantic features are often entangled with each other (Dai et al., 2019). As a result, although the content extractor f_ϕ solely aims at filtering out conceptual information, it is crucial to ensure that the concept-irrelevant information remain complete. Inspired by back-translation in machine translation (Sennrich et al., 2016), we construct a reversal network $\hat{\phi}$, with the same architecture as f_ϕ. $\hat{\phi}$ is designed to reconstruct x from $f_\phi(x)$, ensuring that the mapping function f_ϕ does not excessively lose the concept-irrelevant information. We first freeze ϕ, then use the following loss to train $\hat{\phi}$:

$$\mathcal{L}(\hat{\phi}) = \left\| f_{\hat{\phi}}\left(f_\phi(x)\right) - x \right\|_2^2. \tag{8}$$

Next, we freeze the parameters of $\hat{\phi}$. During the training of ϕ, we use it to remap $f_\phi(x)$ back to x, and compute the mean squared error between them as a content retention loss:

$$\mathcal{L}_{\text{content}}(\phi) = \|f_{\hat{\phi}}(f_\phi(x)) - x\|_2^2. \tag{9}$$

Finally, we combine $\mathcal{L}_{\text{content}}$ and $\mathcal{L}_{\text{concept}}$ with weighted summation to form the overall loss for training f_ϕ, as shown in eq. (10).

$$\mathcal{L}(\phi) = \mathcal{L}_{\text{concept}}(\phi) + \lambda \mathcal{L}_{\text{content}}(\phi), \tag{10}$$

where λ is the weighing to control the relative importance of the content retention.

Finally, we alternately train f_ϕ and $f_{\hat{\phi}}$ to ensure that $f_{\hat{\phi}}$ can effectively track the retention of concept-irrelevant information by f_ϕ. Here, ϕ and $\hat{\phi}$ minimizes conceptual information while maximizing the content information, forming an information bottleneck (Tishby et al., 1999).

After training, the classifier ϕ maximizes the retention of content while minimizing the retention of conceptual information to avoid conceptual shortcut in further training. Here, $f_\phi(x)$ is the content representation of x, denoted as x_{cont}.

3.4 Conceptual Shortcut Debiasing

Although x_{cont} can replace the original embedding x to mitigate the conceptual bias, we further argue that eliminating conceptual shortcuts is not always beneficial. Theoretically, we identify two special cases where preserving conceptual biases could be advantageous: (1) when the conceptual bias aligns with human intent, and (2) when the application scenario is constrained, where the optimization objective is limited to i.i.d. data.

When the imbalance of conceptual attributes aligns with human natural intent, the shortcuts should be enhanced. For example, in the movie review dataset IMDB (Maas et al., 2011), most reviews labeled by GPT-4o as containing the conceptual attribute of "humor" are positive. This observation is consistent with psychological studies on relation between language styles and sentiments, which suggest that humorous expression tends to be associated with positive emotion (Kuiper and Martin, 1993). Furthermore, for certain application scenario where the i.i.d. and OOD data distribution is identical, the real-world data hold the same distributional bias. For instance, in clinical medicine, a model trained on electronic health records collected from a specific hospital is often deployed to the same environment (Hur et al., 2022), classifying text with similar biases in training and inference. In such a case, reinforcing shortcuts can also improve classification performance in application.

To achieve flexible control over the shortcut exploitation, we introduce a lightweight feedforward network ψ on top of the frozen content extractor f_ϕ and the PLM. This network maps both the original embedding x and its content representation x_{cont} into a conceptually controlled semantic space $\mathcal{X}_{\text{CURE}} \subseteq \mathbb{R}^u$. We then employ contrastive learning to regulate their cosine similarity in this space. The training losses for removing the conceptual shortcut $\mathcal{L}_{\text{r}}(\psi)$ and enhancing the shortcut $\mathcal{L}_{\text{e}}(\psi)$ as follows:

$$\mathcal{L}_{\text{r}} = \max\left(0, 1 - \cos(f_\psi(x), f_\psi(x_{\text{cont}})) - \text{M}\right), \tag{11}$$

$$\mathcal{L}_{\text{e}} = \max\left(0, \cos(f_\psi(x), f_\psi(x_{\text{cont}})) - \text{M}\right), \tag{12}$$

where $\text{m} \in [0, 1]$ is a margin that controls the degree of conceptual information retention.

A smaller margin M enforces a stricter optimization objective. In the removal loss $\mathcal{L}_{\text{r}}$, decreasing M compels $f_\psi(x)$ and $f_\psi(x_{\text{cont}})$ to be nearly identical, ensuring the complete removal of conceptual information. Conversely, in the enhancement loss $\mathcal{L}_{\text{e}}$, a smaller M forces $f_\psi(x_{\text{cont}})$ and x_{cont} to be maximally separated, thereby amplifying the influence of conceptual features. By adjusting M, we can flexibly control the extent to which conceptual information is retained in $f_\psi(x)$.

Finally, we replace the original embedding x with $f_\psi(x)$, as the input to the classifier f_θ and jointly train f_θ and f_ψ using Equation (3). The trained model can flexibly adjust the extent of conceptual bias retention based on the training objective, making it either more robust or more *specialized*, as shown in Fig. 2. In terms of parameter efficiency, CURE introduces only a lightweight content extractor and feedforward network on top of the original classifier, ensuring minimal computational overhead.

4 Experiments

4.1 Experimental Setup

Dataset Description We used IMDB (Maas et al., 2011) and Yelp (Zhang et al., 2015) datasets. The IMDB movie review dataset is a binary sentiment analysis dataset, which consists of 50,000 positive or negative reviews from the Internet Movie Database. The Yelp dataset is provided by the Yelp Dataset Challenge, contains business reviews labeled with ratings ranging from 0 to 4 (Zhang et al., 2015). We used the version that was cleaned and organized by Dai et al. (2019).

Based on the concepts labeled in Section 3.2, we divided the samples in the each dataset into two groups for i.i.d. and OOD testing:

- **Group A** contains **imbalanced** concept distributions, where certain concepts are over-represented in one task-relevant category, but the overall number of samples across task-relevant labels remains equal. Samples in Group A will be separate to a biased training set and an i.i.d. test set.

- **Group B** contains **balanced** concept distributions, where each concept has an equal number of samples across the task-relevant categories. Samples in Group B will be used as the OOD test set.

Compared Methods and Hyperparameters As there are currently no model-based debiasing approaches, we primarily compare our method with FL (Lin et al., 2020), which optimizes loss computation with unbalanced data, and RAZOR (Yang et al., 2024b), which utilizes LLMs for data debiasing. The result is shown in Table 1.

In our training, we used a mini-batch size of 16, with the optimizer AdamW (Loshchilov and Hutter, 2019). The learning rate for the content extractor and reversal extractor was set to 0.0001, while that for the classification heads was set to 0.0003. The concept classifier head and task classifier head have identical structures and are based on the same PLM.

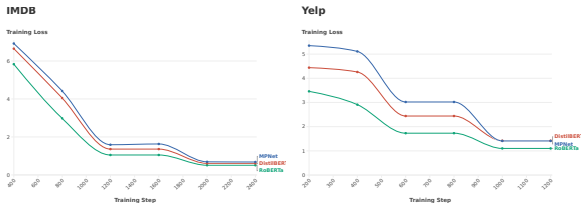


Figure 3: The convergence of the content extractor. We scale the loss values by a factor of 100 for clear comparison.

Computational Efficiency CURE is highly lightweight. Specifically, the content extractor used consists of two single linear layers with layer normalization and a single Transformer layer (Devlin et al., 2019), each with 768 neurons, resulting in a total of approximately 1.78M parameters. Our debiasing module comprises a SwiGLU layer (Shazeer, 2020) followed by a linear layer, with a total of approximately 1.18M parameters. We make a comparison with GPT-3.5-Turbo-based RAZOR in Table 2. Here, we calculated the average training and inference time per sample with a batch size of 16 on a single NVIDIA A100 Tensor Core-Graphics Processing Unit.

Case Study To better understand CURE's improvements, we analyze the model's attention patterns in sentiment classification tasks. Specifically, we randomly sampled a positive review from Yelp, using sentiment classifiers based on DistilBERT and CURE to classify it. After that, we studied their attention across different terms, which is measured by Shapley (Lundberg and Lee, 2017). The attribution visualizations in Table 3 and Table 4 highlight these differences.

The Convergence of the Content Extractor Since the content extractor ϕ is optimized by two training objectives simultaneously, i.e., $\mathcal{L}_{\text{content}}(\phi)$ and $\mathcal{L}_{\text{concept}}(\phi)$, we empirically demonstrated its convergence. The training curve of the content extractor is shown in Fig. 3.

4.2 Results and Discussions

CURE outperformed the baselines on nearly all metrics across both datasets. as shown in Table 1. The largest improvement comes from the Roberta model on the IMDB with the OOD test, with an ap-

Dataset		IMDB						Yelp					
Model		DistilBERT		MPNet		RoBERTa		DistilBERT		MPNet		RoBERTa	
		ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑
i.i.d.	Baseline	84.00	85.05	87.33	86.94	88.50	89.27	94.75	94.76	92.75	93.11	93.75	93.51
	FL	83.70	82.00	87.50	87.32	88.67	88.90	92.25	92.54	93.75	**95.17**	93.50	93.00
	RAZOR	83.25	83.00	87.00	86.50	85.33	83.19	**95.50**	**95.32**	93.50	94.83	92.50	93.00
	CURE	**85.50**	**85.48**	**88.83**	**88.78**	**89.67**	**89.77**	95.25	95.25	**95.00**	95.00	**94.75**	**94.63**
OOD	Baseline	81.67	82.20	79.33	80.19	78.83	74.85	89.75	90.44	89.00	88.30	89.25	89.64
	FL	81.33	82.25	79.00	76.75	79.33	76.70	90.25	89.53	90.25	89.40	89.00	89.52
	RAZOR	80.83	81.30	79.00	79.33	78.67	77.70	90.75	90.60	**90.75**	89.26	89.50	89.76
	CURE	**84.00**	**84.36**	**81.50**	**81.22**	**83.50**	**84.51**	**92.00**	**92.12**	**90.75**	**90.68**	**91.50**	**91.33**

Table 1: Accuracy and F1 on i.i.d. and OOD test on the IMDB and Yelp datasets. "Baseline" stands for PLMs fine-tuned solely on classification tasks. "ACC" stands for Accuracy; **Bolded** values indicate best performing; underlined the second-best.

Model	Scale ↓	Training ↓	Inference ↓
RoBERTa	125M	≈ 11ms	≈ 1ms
RAZOR	GPT-3.5-Turbo	> 600ms	≈ 1ms
CURE	127.96M	≈ 59ms	≈ 1ms

Table 2: Computational scale and the average training/inference time per sample. We take the Yelp dataset with RoBERTa as an example.

[CLS]	the	service	was	great	.	[SEP]

Table 3: SHAP-based token attribution visualization - DistilBERT. Red represents the contribution to positive sentiment. "[CLS]" and "[SEP]" are special tokens.

[CLS]	the	service	was	great	.	[SEP]

Table 4: SHAP-based token attribution visualization - CURE. Red represents the contribution to positive sentiment. "[CLS]" and "[SEP]" are special tokens.

proximate increase of 5 points in Accuracy and 10 points in F1 score. Compared to the i.i.d test, our model introduced a more significant improvement on the OOD test. We analyze that the benchmarks on the i.i.d. test have achieved relatively high accuracy, making it challenging to further improve their performances. Furthermore, we observe that CURE outperforms loss adjustment method FL and LLM-driven approach RAZOR. We attribute this to the fact that FL and RAZOR primarily address label- and word-level biases rather than conceptual biases. For semantic-level biases, these two methods lack mechanisms for regulating the semantic representations, making it challenging for them to improve the baselines. In contrast, CURE remaps the semantic space, enabling the controllable filtering of concept information that cause shortcuts, thereby enhancing robustness of baselines and boosting their OOD performances.

Our findings show that the baseline model tends to distribute attention across both sentiment-related and domain-specific words, while CURE prioritizes sentiment-expressive terms. Table 3 illustrates how the DistilBERT-based classifier assigns nearly equal importance to both "service" and "great", which indicates a reliance on topic-specific terms rather than sentiment indicators. In contrast, Table 4 shows that CURE places stronger emphasis on "great", which suggests it better captures the actual sentiment while reducing confounding biases.

CURE is lightweight and efficient, as shown in Table 2. Compared to the baseline, CURE holds only 2% additional parameters with a nearly identical inference time. Compared to RAZOR which is based on GPT-3.5-Turbo, CURE does not require participating of LLMs during training, which reduces the training time to approximately one-tenth of RAZOR's. Additionally, the time complexity of the debiasing module involved in inference is $\mathcal{O}(L \cdot H^2)$, where L represents the input length and H denotes the hidden state dimension, which aligns with that of the PLMs used (Vaswani et al., 2017). Therefore, the usage of CURE does not alter the time complexity of the baselines. This substantially reduces both computational and time costs that enhances the practicality and generalizability of CURE in real-world applications.

The content extractor used can converge under all conditions, as shown in Fig. 3. This not only provides an experimental foundation for CURE but also indicates that the two optimization objectives employed, i.e. $\mathcal{L}_{\text{content}}(\phi)$ and $\mathcal{L}_{\text{concept}}(\phi)$, are not in conflict. We argue that this finding supports that concept information is not entirely tangled with the semantic information in the latent space, thereby offering a theoretical basis for future work on feature disentanglement.

4.3 Ablation Study

The Effectiveness of Back-Translation To investigate the effect of the reversal network used in training, we conducted ablation experiments on the reversal network, as shown in Table 5.

	Yelp		IMDB	
	ACC ↑	F1 ↑	ACC ↑	F1 ↑
RoBERTa(w/o $\hat{\phi}$)	79.75	83.09	81.33	79.03
RoBERTa(w/ $\hat{\phi}$)	91.50	91.33	83.50	84.51
MPNet(w/o $\hat{\phi}$)	90.25	89.71	79.83	78.73
MPNet(w/ $\hat{\phi}$)	90.75	90.68	81.50	81.22
DistilBERT(w/o $\hat{\phi}$)	91.50	91.05	80.83	82.12
DistilBERT(w/ $\hat{\phi}$)	92.00	92.12	84.00	84.36

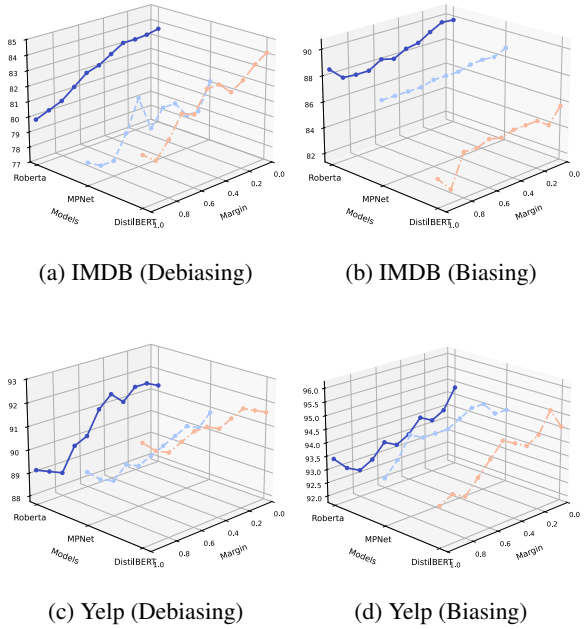
Table 5: Ablation study on the reversal network $\hat{\phi}$. "w/" and "w/o" represent "with" and "without", respectively.

We found that removing the reversal network results in a degradation in classification accuracy, as shown in Table 5. The most significant decline was observed with the RoBERTa model on the Yelp dataset, with a decrease of approximately 12 points in accuracy and 8 points in F1 score. Our further experiments revealed that the content extractor exhibited parameter sparsity in the absence of the reversal network.

Based on these observations, we hypothesize that, without control of content preservation, the content extractor attempts to map all inputs to similar representations, causing its output to become indistinguishable by the concept classifier and leading to the minimization of the loss $\mathcal{L}_{\text{concept}}$. In such a case, due to the information loss on robust features, the classifiers struggled to obtain sufficient effective features for learning, leading to a decline in performance.

The Controllability of Shortcuts We demonstrated how to weaken or enhance shortcuts by adjusting the value of the margin M in eq. (11) and eq. (11), as shown in Fig. 4. To ensure a fair comparison, all other training parameters were held constant in this experiment.

The margin has a controlling effect on the shortcut learning, as shown in Figure 4. We observed that with the increase of M increases, the performance of all three models on the two datasets exhibits a volatile decline. This suggests that a higher margin makes our method more permissive in enhancing or suppressing shortcut learning, leading to a corresponding decrease in performance on both i.i.d. and OOD data. Therefore, by adjusting M, CURE can quantitatively control the impact of shortcut learning on classification, providing a quantifiable benchmark for future debiasing research in theory.



(a) IMDB (Debiasing) (b) IMDB (Biasing)

(c) Yelp (Debiasing) (d) Yelp (Biasing)

Figure 4: The impact of the margin on classification accuracy. Fig. 4a and Fig. 4c show cases for reducing shortcuts on OOD test. Fig. 4b and Fig. 4d show cases for enhancing shortcuts on i.i.d. test.

5 Conclusion

In this work, we introduced CURE, a novel and lightweight framework for mitigating conceptual shortcuts in pre-trained language models. CURE enables fine-grained control over conceptual bias retention by systematically disentangling concept-relevant and content-relevant representations. It balances robustness and accuracy based on task requirements. Our experiments on IMDB and Yelp datasets demonstrate that CURE significantly improves out-of-distribution robustness, achieving up to 5-point accuracy gains and 10-points F1 gains over baselines while maintaining minimal computational overhead. Notably, CURE reduces training time by an order of magnitude compared to LLM-driven debiasing approaches, making it a scalable and efficient solution. These results highlight CURE, which reveals the potential of unsupervised conceptual debiasing in enhancing the reliability of language models while preserving critical task-relevant features.

Limitations

While CURE demonstrates strong performance and computational efficiency, we acknowledge the following limitations. First, due to computational constraints, we were unable to include large-scale comparisons against debiasing baselines such as RAZOR (Yang et al., 2024b) or Focal Loss (Lin et al., 2020) on newer model architectures such as LLaMA3-1B (Meta, 2024) and Qwen-2.5 (Yang et al., 2024a), as well as the GYAFC dataset. While we conducted additional experiments on both settings to evaluate the generalization ability of CURE (see Appendix A.4), these evaluations were limited to comparisons with standard fine-tuned baselines. A complete benchmarking against other debiasing approaches in these settings is left to future work. Second, although CURE itself does not rely on LLM-driven data augmentation during training, we utilized large language models for a one-time concept annotation step during data preprocessing, following prior work (Zhou et al., 2024). This step does not incur additional inference cost and could be replaced with human-annotated concepts in future applications to reduce reliance on external models. However, we did evaluate the plausibility of these annotations through a human study (see Appendix A.3), confirming their quality for use in downstream evaluations.

Despite these limitations, CURE remains a scalable and adaptable framework for mitigating conceptual biases in NLP models, paving the way for more robust and generalizable language understanding systems.

References

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

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.

Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. 2022. Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. *arXiv preprint arXiv:2205.12593*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. 2022. Unifying heterogeneous electronic health records systems via text-based code embedding. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 183–203. PMLR.

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. 2023. Detecting shortcuts in medical images - a case study in chest x-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Darsh Kaushik, Abdullah Faiz Ur Rahman Khilji, Utkarsh Sinha, and Partha Pakray. 2021. CNLP-NITS @ LongSumm 2021: TextRank variant for generating long summaries. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 103–109, Online. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Nicholas A Kuiper and Rod A Martin. 1993. Humor and self-concept. *Humor*, 6(3):251–270.

Sachin Kumar, Shuly Wintner, Noah A Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. *arXiv preprint arXiv:1909.00453*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Meta. 2024. Llama 3.2-1b model card. Accessed: January 23, 2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Noam M. Shazeer. 2020. Glu variants improve transformer. *ArXiv*, abs/2002.05202.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. *arXiv preprint arXiv:2004.07790*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*.

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. AutoCAD: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. 2019. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2024b. Razor: Sharpening knowledge by cutting bias with unsupervised text rewriting. *Preprint*, arXiv:2412.07675.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

We provide further details on the implementation of our method as well as additional experimental results. This appendix is structured as follows:

- In Section A.1, we formalize our training procedure and present the full algorithm.
- In Section A.2, we describe the prompt design used for concept labeling and clustering.
- In Section A.3, we report details of the annotation quality evaluation and human study setup.
- In Section A.4, we provide supplementary results, including additional baselines and experiments demonstrating the robustness of CURE.
- In Section A.5, we show sentiment distributions within the imbalanced concept groups across datasets.

A.1 Training Algorithm

Algorithm 1 The Training Algorithm of CURE

Require: A biased dataset D, a corresponding label set $\mathcal{Y}$, a corresponding concept set $\mathcal{C}$.
Ensure: A robust classifier θ.

Train a concept classifier ω using (5).
for $d \in D$ **do**
    Input d to a PLM, obtain output x.
    Freeze a content extractor ϕ and train a reversal extractor $\hat{\phi}$ using (8).
    Freeze $\hat{\phi}$ and train ϕ using (10).
end for
for $d \in D$ **do**
    Freeze ϕ and train a debiasing module ψ using (11).
end for
Train a task classifier θ using (3).

A.2 Prompt Design for Concept Labeling

Here is a given movie review:

{review}

Identify the main concept discussed in this review using only ONE WORD. Your response should be ONE-WORD for each review (e.g., acting, plot, cinematography).

Examples:
1. Review: "Seen 'Back to the Future'? This movie, 'Tangents' (aka 'Time Chasers'), tries a similar time-travel concept but fails to hit the mark. Made in 1994, it looks and feels like it's from the 80s. The cast includes an unappealing leading man, a cliché-ridden leading lady, a cartoonish villain, and henchmen with questionable jobs. The plot is hard to follow, so I'd recommend watching it with Mystery Science Theater 3000 for entertainment. On its own: 3 stars. With MST3K: 8 stars."
Concept: plot

2. Review: "And you thought your significant other's family was weird? Wedding Slashers will make you think twice about ever saying 'I do.' It is reminiscent of past horror titles such as 'Deadly Friend' and 'Friday the 13th.' It is a classic slasher film that features characters with names like 'Sock Monkey' and 'The Mortician.' You may laugh at first but trust me, these guys will freak you out. This is a quencher for the blood-thirsty horror/slasher fan that needs to see gore, gore and more gore. It's not all slash and gash either - Wedding Slashers is chock-full-of one-liners and will give you more than just a chuckle. You're going to need to see this one to believe it."
Concept: genre

Now, classify the given review and provide the main concept using only ONE WORD:

Table 6: Prompt P_a is used to label the IMDB dataset for concept annotation. The placeholder **{review}** represents the input movie review. The example reviews are also sourced from the IMDB dataset.

Here is a list of extracted concepts from movie reviews: **{concepts}**
Analyze these concepts and suggest an appropriate number of clusters and one-word cluster names to group them. Cluster names should not overlap, should be distinctive.

Table 7: Prompt P_b is used to refine the concept clusters, and it returns final concept list $\mathcal{C}$. The placeholder **{concepts}** represents the concepts that are generated using P_a.

Given concept: **{concept}**
Predefined Concept List: **{concept labels}**
Provide the concept from the predefined list that is closest to the given concept. Return nothing else.

Table 8: Prompt P_c is used to assign the final concept from $\mathcal{C}$ to each movie review. The placeholder **{concept}** represents the extracted concept from a movie review. The placeholder **{concept labels}** refers to $\mathcal{C}$, the predefined concept list generated using P_b.

A.3 Annotation Quality Evaluation

To measure the quality of GPT-4o's concept annotations, we conducted a human evaluation using crowdsourcing. We randomly selected 10 annotations from each dataset (Yelp and IMDB), and each annotation was rated by seven independent annotators using Qualtrics [2]. The annotators assessed how accurately each concept reflected the associated text using a 5-point Likert scale (Likert, 1932), where 1 = *Not accurately at all*, 2 = *Slightly accurately*, 3 = *Moderately accurately*, 4 = *Very accurately*, and 5 = *Extremely accurately*. The average ratings were 4.31 for Yelp and 3.81 for IMDB. We define the *agreement rate* as the proportion of ratings above 3, which reached 100% for Yelp and 70% for IMDB. These results indicate that GPT-4o's concept annotations are largely considered plausible and can be reliably used in downstream tasks.

Dataset	Mean Rating	Agreement Rate (%)
Yelp	4.31	100
IMDB	3.81	70

Table 9: Human evaluation of GPT-4o's concept annotations.

A.4 Supplementary Experiments

To further demonstrate the generalization capability of our method, we conducted additional evaluations in two settings: (1) on the GYAFC dataset—the largest text-style analysis dataset (Rao and Tetreault, 2018), and (2) by applying CURE to the LLaMA3-1B model (Meta, 2024).

Evaluation on GYAFC. We evaluated CURE on the GYAFC dataset, and the results show consistent performance gains across all model backbones (see Table 10):

Model	ACC	F1
DistilBERT	75.62	0.68
DistilBERT (w/ CURE)	**81.87**	**0.79**
MPNet	75.94	0.71
MPNet (w/ CURE)	**77.19**	**0.73**
RoBERTa	77.81	0.72
RoBERTa (w/ CURE)	**87.50**	**0.87**

Table 10: Classification results on the GYAFC dataset.

Experiments with LLaMA3. We further replicated our experiments using LLaMA3-1B to evaluate the applicability of CURE to recent foundation models. Experimental results using LLaMA3-1B are shown in Table 11.

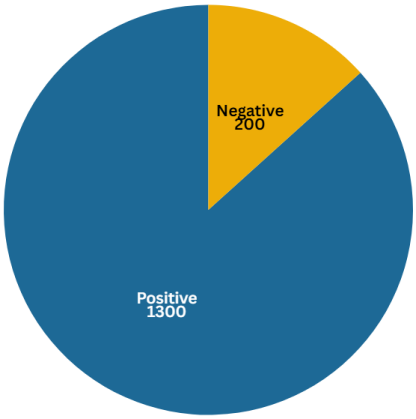
These results confirm that CURE remains effective even on recent and larger models, yielding a 3-point accuracy improvement on IMDB and consistent gains on Yelp, comparable to trends observed with smaller-scale PLMs. Due to computational resource limitations, these additional experiments could not be extended and, therefore, were not included in the main body of the study. Nevertheless, we report them here in the appendix to highlight the broader applicability and robustness of CURE across both new datasets and emerging model architectures.

[2] https://www.qualtrics.com/

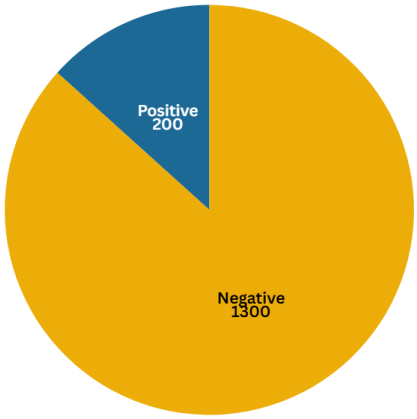
Dataset	IMDB		Yelp	
Model	ACC ↑	F1 ↑	ACC ↑	F1 ↑
LLaMA3	70.67	65.00	93.25	93.00
LLaMA3 (w/ CURE)	**73.83**	**75.00**	**93.50**	**94.00**

Table 11: Accuracy and F1 results on IMDB and Yelp using LLaMA3-1B. "LLaMA3" refers to the base model without debiasing. **Bold** indicates the best result.

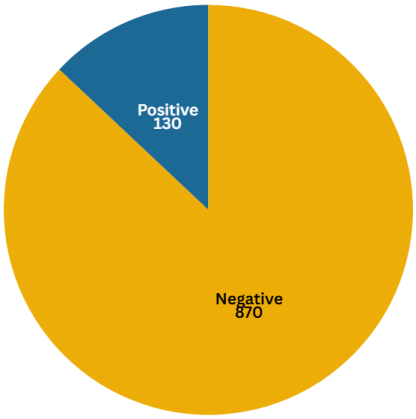
A.5 Sentiment Distributions in the Imbalanced Groups



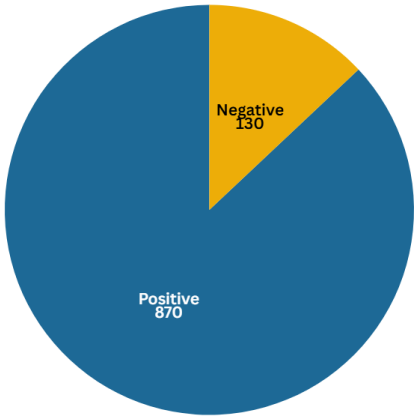
(a) Sentiment distribution in the concept "Emotion"



(b) Sentiment distribution in the concept "Story"



(c) Sentiment distribution in the concept "Experience"



(d) Sentiment distribution in the concept "Service"

Figure 5: Sentiment distributions in the imbalanced groups of the IMDB and Yelp datasets.