COBA: Counterbias Text Augmentation
for Mitigating Various Spurious Correlations via Semantic Triples

Kyohoon Jin[1,3*]   **Juhwan Choi**[2,4*]   **Jungmin Yun**[4]
Junho Lee[4]   **Soojin Jang**[3]   **Youngbin Kim**[3,4 †*]
[1]DATUMO   [2]AITRICS
[3]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
[4]Department of Artificial Intelligence, Chung-Ang University
kyohoon.jin@selectstar.ai, jhchoi@aitrics.com
{cocoro357, jhjo32, sujin0110, ybkim85}@cau.ac.kr

Abstract

Deep learning models often learn and exploit spurious correlations in training data, using these non-target features to inform their predictions. Such reliance leads to performance degradation and poor generalization on unseen data. To address these limitations, we introduce a more general form of counterfactual data augmentation, termed *counterbias* data augmentation, which simultaneously tackles multiple biases (e.g., gender bias, simplicity bias) and enhances out-of-distribution robustness. We present COBA: **Co**unter**B**ias **A**ugmentation, a unified framework that operates at the semantic triple level: first decomposing text into subject-predicate-object triples, then selectively modifying these triples to disrupt spurious correlations. By reconstructing the text from these adjusted triples, COBA generates *counterbias* data that mitigates spurious patterns. Through extensive experiments, we demonstrate that COBA not only improves downstream task performance, but also effectively reduces biases and strengthens out-of-distribution resilience, offering a versatile and robust solution to the challenges posed by spurious correlations.

1 Introduction

Despite deep learning's success across various domains, spurious correlations continue to pose significant challenges in training effective models (Ye et al., 2024). Spurious correlations are patterns that appear in datasets but do not represent genuine relationships, such as correlations with background or textures (Beery et al., 2018; Geirhos et al., 2019; Sagawa et al., 2020). This phenomenon is also prevalent in text data, where spurious correlations frequently emerge at the word-level. In such cases, certain words or phrases become associated with specific labels due to their co-occurrence in particular contexts. This association often fails to reflect

the actual meaning or intent, resulting in performance degradation in models (Wang et al., 2022; Joshi et al., 2022; Chew et al., 2024). Furthermore, spurious correlations are linked to various biases, including gender bias, and challenges related to out-of-distribution (OOD) robustness. Consequently, mitigating these correlations is crucial for enhancing deep learning models in a broader context (McMilin, 2022; Liusie et al., 2022; Ming et al., 2022; Zhou et al., 2024).

Early studies have focused on mitigating spurious correlations from a model-centric perspective by identifying spurious features. For instance, several approaches suggested reweighting data samples to mitigate spurious features; however, this strategy can inadvertently introduce new biases by overemphasizing irrelevant features (Han and Tsvetkov, 2021; Shi et al., 2023). Subsequently, recent studies have shifted the focus toward a data-centric approach, particularly in the field of natural language processing (Ye et al., 2024). In this direction, researchers have been exploring data manipulation techniques aimed at enhancing the generality and diversity of data distribution and reducing the impact of spurious correlations present in the original data, thereby improving model capabilities (Wang et al., 2024). A line of studies suggest that augmenting datasets with counterfactual data—entailing minimal modifications to the original sentences—can effectively mitigate spurious correlations (Kaushik et al., 2020; Udomcharoenchaikit et al., 2022; Chan et al., 2023). While early studies relied on human-annotated counterfactuals, more recent works propose automatically generating them through data augmentation, demonstrating their effectiveness in reducing spurious correlations (Zeng et al., 2020; Wang and Culotta, 2021; Wen et al., 2022; Treviso et al., 2023; Sachdeva et al., 2024). However, due to the minimal modifications, this approach may lack diversity, potentially leading to issues such as overfitting and

---

* * denotes euqal contribution.
†Corresponding author. Email: ybkim85@cau.ac.kr

subsequent performance degradation (Qiu et al., 2024).

In this study, we extend current research on counterfactual data augmentation to *counterbias data augmentation*, which simultaneously addresses various biases and challenges, such as gender bias, and out-of-distribution robustness. Although counterfactual data has been effective in mitigating spurious correlations, there remains significant potential for a unified approach that can concurrently tackle these diverse challenges. To explore this, we propose transforming the given text into a set of semantic triples using a large language model (LLM), with each triple encapsulating compressed information from the sentences. By generating counterfactual triples through modifications of the original triples and reconstructing text from these debiased triples using an LLM, we can create augmented *counterbias* data. This triple-level modification simplifies the generation of counterfactuals, as triples naturally contain the key elements of sentences. Additionally, with the support of LLMs in reconstructing text from triples, our framework can effectively diversify augmented text. *Counterbias data augmentation* differs from previous counterfactual data augmentation approaches, which aim to make minimal changes while flipping the original data's label.

Additionally, we conducted an analysis to identify principal words in various models using word importance measurements, revealing that each model has a distinct set of principal words. This finding suggests that counterbias data generated for a single model may not be effective for other models. To address this finding, we employ a majority-voting-based ensemble method to identify words that may contribute to spurious correlations. This approach is effective for augmenting counterbias data that can be universally applied across various models. Through experiments validating the effectiveness of our proposed framework, CoBA, we observed that it effectively alleviates various biases and challenges while also augmenting counterbias data applicable across different models.

Our main contributions are as follows:

- **A Unified Framework for Counterbias Augmentation:** We introduce CoBA, a novel approach that extends counterfactual data augmentation to *counterbias* data augmentation. Unlike prior methods that primarily focus on minimal label-flipping modifications, CoBA targets a broader range of biases and spurious correlations, improving both in-distribution performance and out-of-distribution robustness.

- **Insights into Spurious Correlations Across Models:** Through a detailed analysis of word importance, we reveal how spurious correlations vary significantly across different model architectures, underscoring the limitations of relying on a single model. This insight motivates our ensemble-based strategy to identify spurious correlations more reliably.

- **Empirical Validation and Practical Benefits:** Extensive experiments across tasks like sentiment analysis, natural language inference, and text style transfer show that CoBA consistently alleviates multiple biases and enhances model resilience to distribution shifts. These results highlight CoBA's versatility and its potential to construct more robust, fair, and generalizable deep learning models.

2 Related Work

Counterfactual data augmentation has been shown to mitigate spurious correlations effectively. An early study introduced the concept of counterfactual data by manipulating existing data to alter the label with minimal modifications (Kaushik et al., 2020). These counterfactual data have been demonstrated to be useful for mitigating spurious patterns and evaluating local decision boundaries of models (Gardner et al., 2020).

Since these studies relied on human annotators to generate counterfactual data, producing such data for various datasets was challenging. As a result, researchers began exploring automated methods for generating counterfactual data, particularly in data augmentation setups. In early explorations, predefined rules were applied to augment counterfactual data (Zmigrod et al., 2019; Wang and Culotta, 2021).

Beyond rule-based techniques, deep learning models have been employed to augment counterfactual data. For example, several studies have proposed leveraging well-trained classifiers to identify principal words (Wang et al., 2022; Wen et al., 2022; Bhan et al., 2023). Additionally, generating counterfactual data using deep learning models has been proven to be effective in diversifying the generated data (Wu et al., 2021; Treviso et al., 2023; Sun et al., 2024; Ding et al., 2024). Recently, researchers have also begun exploring the

use of LLMs for counterfactual data augmentation (Sachdeva et al., 2024; Chang et al., 2024; Li et al., 2024). While C2L similarly adopts a collective decision strategy over counterfactuals, it relies on a single model; in contrast, CoBA leverages multiple diverse models to more robustly identify causal features and mitigate spurious correlations (Choi et al., 2022).

3 Methodology

3.1 Overview

In this paper, we aim to alleviate various biases and obstacles by mitigating spurious correlations through counterbias data augmentation. Specifically, given an original dataset $\mathcal{D}_{ori} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{ori}|}$, where x_i denotes the input text and y_i its corresponding label, our goal is to generate new pairs $(\hat{x}_i, \hat{y}_i)$, where $\hat{x}_i$ is the augmented counterbias text and $\hat{y}_i \neq y_i$. We define counterbias text as text that retains the spurious words and semantics of the original text but is assigned a different label to mitigate spurious correlations, similar to counterfactual text. This definition represents a relaxed form of counterfactual text, which refers to text with minimal differences from the original data but with different labels (Molnar, 2020). Unlike counterfactual data, counterbias data are not restricted to minimal changes; they may exhibit different syntactic structures and expressions, as long as they preserve the spurious words and semantics of the original text. This distinction allows counterbias data to introduce a wider variety of patterns, thereby amplifying the augmentation effect.

To accomplish this, we first decompose x_i into a set of semantic triples, denoted by $T^{(x_i)} = \{t_j^{(x_i)}\}_{j=1}^{|T^{(x_i)}|}$. Each triple $t_j^{(x_i)}$ consists of (subject, predicate, object) derived from a sentence in x_i[1]. This decomposition is carried out by an LLM using a designated prompt. Next, we modify each decomposed $t_j^{(x_i)}$ to mitigate spurious correlations at the triple level, resulting in modified triples $\hat{t}_j^{(x_i)} \in \hat{T}^{(x_i)}$. Specifically, we proceed as follows:

1) We identify two sets of words, W_s and W_p, where W_s is the set of spurious words that cause spurious correlations, and W_p is the set of principal words crucial for determining the label of x_i. To obtain $w_s \in W_s$ and $w_p \in W_p$, we use multiple classifiers with different backbones trained on $\mathcal{D}_{ori}$

---

[1] Note that the relationship between a triple and a sentence is not one-to-one; a sentence can contain multiple triples.

	LIME	IG	SV
SST-2	26.72% (83.9%)	18.53% (83.2%)	14.21% (85.3%)
IMDB	8.64% (81.4%)	7.00% (74.6%)	7.99% (81.3%)

Table 1: The ratio of duplication among the top-5 most principal words for each model. The number in parentheses indicates the degree of overlap between two or more models, but not every model.

and word importance measurement techniques.

2) We then generate $\hat{T}^{(x_i)}$ by modifying any $t_j^{(x_i)}$ that contains w_p, while retaining $t_j^{(x_i)}$ that contains w_s. This minimal, triple-level alteration differentiates the label yet preserves spurious words, producing what we term counterfactual triples.

3) Lastly, to introduce diverse patterns into the augmented data—thereby enhancing OOD robustness—we randomly permute the order of the triples in $\hat{T}^{(x_i)}$ and delete several triples as well.

Finally, we augment $\hat{x}_i$ by reconstructing counterbias text from the modified $\hat{T}^{(x_i)}$ using the LLM with a designated prompt. Since we have modified $T^{(x_i)}$ to $\hat{T}^{(x_i)}$ to retain $\hat{y}_i$ instead of y_i by modifying $t_j^{(x_i)}$ with w_p, the reconstructed $\hat{x}_i$ contains the label $\hat{y}_i \neq y_i$. This results in a counterbias-augmented dataset $\mathcal{D}_{cb}$, which is used to train a downstream task model by $\mathcal{D}_{ori} \cup \mathcal{D}_{cb}$. Figure 1 illustrates this overall procedure.

3.2 Analysis on Important Words

Before introducing CoBA in detail, we first present an analysis to investigate the differences in W_p across various models from two perspectives. Implementation details are provided in Appendix A.

3.2.1 Word-level Importance Analysis

First, we compared the word-level importance of each model. We measure word-level importance for each model using three different word importance measurement techniques: local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), integrated gradient (IG) (Sundararajan et al., 2017), and Shapley value (SV) (Rozemberczki et al., 2022).

We used these techniques to measure the top-5 important words for each model on the SST-2 and IMDB datasets. Afterward, we evaluated the ratio of duplicated words among the important words identified by each model. Specifically, we counted instances where all four models contained at least one duplicated word. Table 1 presents the results. The findings suggest that the number of words con-
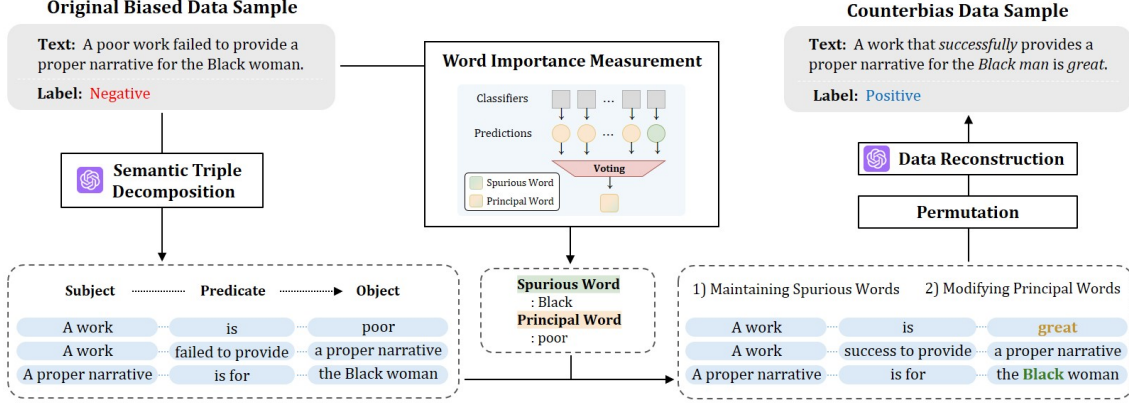
Figure 1: Overall procedure of CoBA.

SST-2	Noun	Verb	Adjective & Adverb	Others
BERT-base	**50.9**	10.1	18.1	*20.9*
BERT-large	**48.6**	2.3	*35.1*	14.0
RoBERTa-large	**67.9**	9.8	*15.6*	6.7
BART-base	**35.3**	*22.0*	21.0	21.6
T5-base	*25.5*	**40.1**	23.8	10.6
IMDB	Noun	Verb	Adjective & Adverb	Others
BERT-base	*22.6*	12.2	16.0	**49.2**
BERT-large	*28.4*	12.4	8.3	**50.8**
RoBERTa-large	**46.7**	16.0	*29.1*	8.2
BART-base	*26.8*	11.3	24.9	**37.0**
T5-base	6.0	18.5	*24.4*	**51.1**

Table 2: The ratio of POS tags among top-5 most important words for each model on SST-2 and IMDB. Bolded values represent the most frequent POS tag for each model and dataset, while italicized values represent the second most frequent POS tag.

sistently regarded as important across all models is small. Notably, this ratio was less than 10% of the total words in IMDB, which contains relatively longer text than SST-2. Although the models in this analysis share BERT-family architecture, they focus on different words when making predictions. However, when examining the overlap in important words between just two models at a time, we found that most cases exceeded 80%. This indicates that while each model has its own tendencies, there is still meaningful overlap in the patterns they recognize, suggesting they focus on sentence semantics in distinct yet related ways.

3.2.2 POS Tagging Analysis

To support the findings of the previous analysis, we conducted an additional analysis by performing POS tagging on top-5 important words identified from the analysis above. Table 2 presents the result of this analysis. The findings indicate that the important words identified by each model have dif-

ferent POS tags, revealing that each model focuses on different aspects of the given text. Appendix D.2 provides qualitative evaluation of this tendency.

These two analyses suggest that counterfactual data augmented by previous methods, which leveraged a single model to identify important words from the input text, may not be adequate for other models, diminishing the efficiency to be applied universally across various models. Inspired by this finding, we propose leveraging multiple models and using a majority-voting-based ensemble method to identify important words, including spurious and principal words.

3.3 CoBA

In this section, we introduce the detailed procedure of our CoBA; **C**ounter**B**ias **A**ugmentation framework. CoBA consists of three major components: semantic triples decomposition, triple-level manipulation, and reconstruction of counterbias text. The overall procedure is illustrated in Figure 1.

3.3.1 Semantic Triple Decomposition

To augment the given x_i into $\hat{x}_i$, we first decompose x_i into a set of semantic triples $T^{(x_i)} = \{t_j^{(x_i)}\}_{j=1}^{|T^{(x_i)}|}$. Each $t_j^{(x_i)}$ represents a sentence or phrase from x_i, and follows the structure of (subject, predicate, object). For instance, given x_i as "I love In-N-Out. Their burger feels incredibly fresh", the desired $T^{(x_i)}$ is {(I, love, In-N-Out), (Their burger, feels, incredibly fresh)}. While various techniques exist for triple decomposition, they primarily focus on decomposing a *single* sentence into semantic triples, which differs from our purpose (Tan et al., 2019; Ye et al., 2021; Chen et al., 2021).

To effectively decompose text containing multiple sentences into semantic triples, we utilize

LLMs, which can perform various tasks when given proper instructions through prompts (Brown et al., 2020; Ouyang et al., 2022). We achieve this by designing a dedicated prompt p_{ext}[2] for an LLM $\mathcal{L}$. Consequently, the desired set of semantic triples $T^{(x_i)}$ is obtained by $T^{(x_i)} = \mathcal{L}(x_i, p_{ext})$.

3.3.2 Triple-level Manipulation

Since $T^{(x_i)}$ contains compressed information about the original x_i, we aim to modify this compressed $T^{(x_i)}$ to mitigate the underlying spurious correlations. The procedure is detailed as follows:

First, we employ a set of multiple classifiers $\mathcal{M}$, where each $m_i \in \mathcal{M}$ represents an individual classifier trained on $\mathcal{D}_{ori}$. After training $\mathcal{M}$, we measure word importance on x_i using each m_i and extract K important words, denoted as $W^{(m_i)}$. We then count the frequency of each word's appearance in the $W^{(m_i)}$. If a certain word appears in $W^{(m_i)}$ more than the threshold τ[3], indicating its importance across various models, we include it in W_p, the set of principal words crucial for determining the label of x_i. Words in $W^{(m_i)}$ that are not included in W_p are categorized into W_s, as they are important only for certain models and not universally significant, implying that such words may induce spurious correlations in the model. Additionally, arbitrary words designated by the engineer that are known to induce spurious correlations and biases can also be included in W_s in case of need.

Second, we modify $T^{(x_i)}$ to mitigate spurious correlations at the triple-level. Specifically, we first categorize each $t_j^{(x_i)}$ in $T^{(x_i)}$ as a spurious triple if $t_j^{(x_i)}$ contains a word from W_s. Other triples that contain a word from W_p are categorized as principal triples. After categorization, we obtain $\hat{T}^{(x_i)}$ by modifying only the principal triples while maintaining the spurious triples. In particular, we use $\mathcal{L}$ to alter the label of x_i by modifying the principal triples, which play a crucial role in determining the label. This process results in the generation of modified principal triple, $\hat{t}_j^{(x_i)} = \mathcal{L}(t_j^{(x_i)}, \hat{y}_i, p_{mod})$, where $\hat{y}_i \neq y_i$ denotes the desired label. This targeted manipulation preserves the spurious words and semantics of the original data while converting the label, thereby augmenting counterbias data.

Finally, to effectively leverage the flexibility of counterbias data, which allows for various changes

compared to the original data, such as different syntactic structures, we randomly permute the order of normal triples that are not categorized as spurious or principal triples within $\hat{T}^{(x_i)}$. Additionally, gender bias-inducing words are replaced with words of the different genders at the triple-level to mitigate gender bias. We used the WinoBias dataset (Zhao et al., 2018) to perform this modification. Furthermore, we randomly delete some of the normal triples with a small, predefined probability. Rearrangement and deletion of normal triples help introduce diverse patterns into the augmented data while preserving the core semantics. The completion of this process produces the final candidate set of triples for reconstruction, $\hat{T}^{(x_i)}$.

3.3.3 Reconstruction of Counterbias Text

Finally, we augment counterbias text $\hat{x}_i$ by reconstructing text given the processed $\hat{T}^{(x_i)}$. Specifically, we utilize $\mathcal{L}$ to achieve this, which is formulated as $\hat{x}_i = \mathcal{L}(\hat{T}^{(x_i)}, p_{rec})$. As a consequence, we obtain the counterbias data $(\hat{x}_i, \hat{y}_i)$. Note that we can easily generate multiple $\hat{x}_i$ using different configurations of decoding strategies for $\mathcal{L}$ or even different arrangements of $\hat{T}^{(x_i)}$. This is different from conventional counterfactual data augmentation, which is difficult to augment multiple data as they require minimal changes to the original data.

4 Experiments

4.1 Improvement on Task Performance

We first evaluated performance improvements in downstream tasks to determine if COBA effectively mitigates spurious correlations and outperforms conventional data augmentation methods, including counterfactual data augmentation. For this purpose, we primarily used natural language inference (NLI) and sentiment analysis tasks. We used SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) for NLI tasks and SST-2 (Socher et al., 2013) and IMDB (Maas et al., 2011) for sentiment analysis tasks, considering the configuration of previous studies (Wen et al., 2022; Wang et al., 2022; Treviso et al., 2023; Li et al., 2024). Implementation details and baseline methods are provided in Appendix B.

Table 3 demonstrates the result of the experiment. A key finding is that COBA outperformed other baselines, including counterfactual data augmentation methods, in most cases. Although counterfactual data augmentation methods effectively

---

[2]Please refer to Appendix H for full details on p_{ext} and other prompts.

[3]We simply set τ as $(|\mathcal{M}| + 1)/2$, where $|\mathcal{M}|$ denotes the number of classifier models. Note that $|\mathcal{M}|$ is an odd number.

	BERT	DeBERTaV3	T5	ModernBERT
Baseline w/o Augmentation	92.8 | 91.5 | 86.2 | 82.4	94.0 | 91.6 | 86.6 | 84.5	94.5 | 92.3 | 85.4 | 83.8	93.8 | 94.1 | 88.3 | 89.8
EDA (Wei and Zou, 2019)	93.1 | 90.8 | 86.8 | 80.6	93.1 | 91.6 | 86.2 | 81.5	92.9 | 91.6 | 88.8 | 82.6	92.4 | 91.4 | 89.7 | 86.9
BT (Sennrich et al., 2016)	93.2 | 91.4 | 87.7 | 83.1	93.5 | 92.2 | 84.1 | 82.8	89.3 | 88.2 | 88.1 | 83.4	94.8 | 93.9 | 89.7 | 85.7
C-BERT (Wu et al., 2019)	91.9 | 92.1 | 84.4 | 82.1	94.0 | 91.0 | 89.0 | 84.7	93.2 | 90.8 | 91.2 | 85.4	90.1 | 90.3 | 91.1 | 89.3
Human-CAD (Kaushik et al., 2020)	00.0 | 93.2 | 88.0 | 00.0	00.0 | 93.8 | <u>89.9</u> | 00.0	00.0 | 95.1 | 89.9 | 00.0	00.0 | 97.2 | 91.2 | 00.0
AutoCAD (Wen et al., 2022)	**94.9** | 92.8 | 88.0 | 89.8	96.4 | 93.3 | **90.1** | <u>91.3</u>	95.2 | 93.4 | 89.1 | **92.0**	95.4 | 95.7 | 91.6 | 92.4
GPT3Mix (Yoo et al., 2021)	93.2 | 93.9 | 00.0 | 00.0	95.2 | **94.1** | 00.0 | 00.0	95.3 | 93.9 | 00.0 | 00.0	96.7 | 96.3 | 00.0 | 00.0
AugGPT (Dai et al., 2023)	94.2 | 92.2 | **90.3** | 88.7	95.4 | 94.0 | 87.5 | 87.6	95.7 | <u>94.2</u> | 88.9 | 85.1	95.1 | 95.1 | <u>92.1</u> | 89.6
CoBA (*LLM-Identification*)	<u>94.6</u> | <u>94.4</u> | 89.9 | <u>90.6</u>	**96.7** | <u>94.0</u> | 88.2 | 90.6	<u>95.9</u> | 93.8 | **90.5** | 89.2	96.3 | 96.9 | **92.2** | **93.1**
CoBA	**94.9** | **95.4** | <u>90.1</u> | **91.1**	<u>96.5</u> | **94.1** | **90.1** | **91.7**	**96.2** | **95.3** | **90.5** | <u>91.3</u>	96.2 | 96.9 | 91.0 | <u>92.9</u>

Table 3: Comparison of downstream task performance under different data augmentation strategies. The best performance in each group is in **bold**, and the second-best is in <u>underline</u>. Performances are reported as "SST-2 | IMDB | SNLI | MNLI" under each model. Note that Human-CAD only provides counterfactual datasets for IMDB and SNLI, and the official source code of GPT3Mix is limited to processing large datasets such as SNLI and MNLI, which are denoted as "~~00.0~~".

	IMDB	SNLI
w/o Augmentation	52.3	70.2
AutoCAD	86.1	75.6
CoBA	**87.2**	**75.8**

Table 4: The comparison of models on Human-CAD test set. For this experiment, we trained BERT-base.

mitigate spurious patterns, they limit data diversity by introducing minimal modifications when converting labels. On the other hand, conventional data augmentation methods, particularly LLM-based methods such as GPT3Mix and AugGPT exhibit the variation in augmented data; however, they do not take spurious correlations into account. By combining the advantages of mitigating spurious correlations and generating diverse augmented data, CoBA was able to outperform other baselines.

Additionally, we conducted a small ablation study, where LLMs identified W_s and W_p instead of using $\mathcal{M}$, a set of multiple well-trained classifiers. The results are presented as "CoBA (LLM-Identification)" in Table 3. While this approach showed remarkable performance compared to other baselines, its improvement was smaller than that of the original CoBA. This suggests that identifying spurious and principal words using LLM may be less effective than our majority-voting-based ensemble method with downstream task models. We hypothesize this phenomenon arises because a single LLM may not effectively capture W_s and W_p, given the difference in important words across models, as highlighted in Section 3.2.

4.2 Mitigation of Spurious Correlation

To verify that the effectiveness of the proposed method comes from mitigating spurious corre-

lations rather than just data augmentation, we adopted the Human-CAD test set, which provides human-annotated examples for assessing spurious correlation mitigation. For this evaluation, we trained BERT-base using a combination of $\mathcal{D}_{ori}$ and $\mathcal{D}_{cb}$ augmented by CoBA and AutoCAD.

As shown in Table 4, the model trained without $\mathcal{D}_{cb}$ exhibited significantly lower performance on the Human-CAD test set, indicating that the baseline model is vulnerable to spurious correlations. In contrast, models trained with AutoCAD and CoBA demonstrated more robust performance than the baseline. Furthermore, CoBA, our proposed method based on counterbias augmentation, outperformed existing counterfactual data augmentation methods, underscoring its effectiveness in mitigating spurious correlations.

4.3 Alleviation of Gender Bias

To verify CoBA's effectiveness in reducing biases by mitigating related spurious correlations, we conducted an experiment focused on gender bias reduction. For this experiment, we adopted the list of gender bias-related words from a previous study (Zhao et al., 2018). By incorporating words from this list into W_s, we aim to mitigate the underlying spurious correlations, thereby alleviating gender bias in the model. To quantify the gender bias, we used three benchmarks: StereoSet (SS) (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), and WikiBias (Zhong et al., 2021). For comparison, we established various baselines. BERT (Raw) refers to the original BERT model without any additional training, and BERT (IMDB) refers to the BERT model with additional pretraining on the IMDB training dataset. SentenceDebias is a baseline method that achieves debiasing at the em-

	SS	CrowS	WikiBias
BERT (Raw)	57.8	59.0	69.0
BERT (IMDB)	58.6	59.7	69.1
SentenceDebias (Liang et al., 2020)	53.8	58.1	72.1
Naive-masking (Thakur et al., 2023)	56.5	60.8	71.8
Random-phrase-masking (Thakur et al., 2023)	54.5	58.0	71.2
CoBA-PT	**51.4**	**52.0**	**75.4**

Table 5: Comparison of gender bias across methods, as measured by SS and CrowS. A score close to 50 for SS and CrowS, along with a higher score for WikiBias, indicates less gender bias in the model. Note that all models are based on `BERT-base`.

EDA	AutoCAD	AugGPT	CoBA
0.9957	0.9641	0.9658	0.9531

Table 6: Cosine similarity between the embedding vectors of x_i and $\hat{x}_i$ from each method.

	IMDB $\to$ SST-2	IMDB $\to$ Yelp
Baseline	63.2	61.2
EDA	66.2	58.2
AutoCAD	80.1	73.6
AugGPT	83.0	71.6
ReAct	84.5	75.3
CoBA	83.2	74.0

Table 7: AUROC(%) performance of the models in OOD scenario. "IMDB $\to$ SST-2" indicates a scenario where a model trained on IMDB is tested on SST-2, and "IMDB $\to$ Yelp" means the model trained on IMDB is tested on Yelp. For ReAct, we follow the reported performance from previous work (Baran et al., 2023).

Baseline	AugGPT	CoBA
72.10 & 73.59	74.50 & 75.21	75.70 & 76.08

Table 8: BLEU-4 scores (Papineni et al., 2002) for the informal-to-formal text style transfer task on the GYAFC dataset. We used for `T5-Base` (Raffel et al., 2020) and `Flan-T5-Base` (Chung et al., 2024) for this experiment.

bedding level (Liang et al., 2020). Naive-masking and Random-phrase-masking are methods based on word-level substitution (Thakur et al., 2023). Lastly, CoBA-PT refers to the BERT model with additional pretraining on the IMDB training dataset, combined with the augmented data generated by CoBA.

Table 5 shows the result of the experiment. The model trained with a combination of original and CoBA-augmented counterbias data achieved a score closest to the ideal 50 on SS and CrowS-Pairs, as well as highest performance on WikiBias, outperforming other baselines. Unlike other strategies, such as masking gender-related pronouns to neutral pronouns, our CoBA focuses on augmenting data with representations of the different gender, leading to a more balanced introduction of gender-related representations. As a result, CoBA contributed to mitigating spurious correlations, thereby alleviating gender bias in the model.

4.4 OOD Robustness with Diverse Augmented Data

Unlike counterfactual text augmentation, which introduces minimal modifications to alter labels, counterbias text augmentation has no such restrictions, allowing for a wider range of lexical and semantic expressions. This flexibility plays a crucial role in enhancing model performance through data augmentation (Cegin et al., 2024). To validate this, we randomly sampled 100 augmented

data from IMDB using each method and measured the difference between the original and augmented data by calculating the cosine similarity produced by `BERT-base`. The results in Table 6 indicate that CoBA introduces meaningful differences in the data while preserving core semantics.

To further support the effectiveness of this diversification in augmented data, we conducted an evaluation in an OOD scenario. For this experiment, we trained a model on IMDB but tested it on SST-2 and Yelp (Zhang et al., 2015). The results of this evaluation are presented in Table 7. This evaluation suggests that our CoBA exhibits remarkable improvement in OOD robustness. While the performance gain is slightly lower than that of ReAct (Sun et al., 2021) baseline, it is important to note that ReAct is a strategy that solely focused on enhancing OOD robustness. In contrast, our CoBA offers various benefits such as mitigation of spurious correlation and other biases. In conclusion, we validated that CoBA jointly offers numerous benefits to the model, from the mitigation of spurious correlation to the improvement of OOD robustness.

4.5 Extension to Generation Tasks

In this paper, we proposed CoBA, which involves decomposing the given text into semantic triples, selecting spurious and principal triples, applying bias-mitigation techniques, and then reconstructing the augmented text. This approach is applicable

	Original [Label: Neutral]
Premise	A **woman** talks on a cellphone while sitting in front of blue railings that are in front of the ocean.
Hypothesis	**She** talks to her boyfriend about plans that night.
	Human-CAD [Label: Contradiction]
Premise	A **woman** talks on a cellphone while sitting in front of blue railings that are in front of the ocean.
Hypothesis	**He** has a conversation on her phone outdoors.
	CoBA [Label: Contradiction]
Premise	A **man** is sitting in front of blue railing while talking on a cellphone, with the railing positioned in front of the ocean.
Hypothesis	**A man** talks to his boyfriend while he is in a new car.

Table 9: The comparison of augmented data generated by Human-CAD and CoBA on SNLI.

not only to classification tasks but also to text generation tasks. To verify CoBA's effectiveness in text generation, we conducted an experiment applying CoBA to a text style transfer task using the GYAFC dataset (Rao and Tetreault, 2018).

For applying CoBA to the text style transfer task, we first decomposed the given x_i into $T^{(x_i)}$. Next, we utilized $\mathcal{L}$ to identify principal triples in both formal and informal sentences[4]. Additionally, we included the gender bias alleviation scheme introduced in Section 4.3. After permuting the order of normal triples that are not principal, we reconstructed the augmented text.

The results of this experiment, displayed in Table 8, show that CoBA exhibited a remarkable performance improvement compared to the AugGPT baseline. This underscores CoBA's extensibility to text generation tasks based on the triple-level modifications. We plan to investigate the strategies for identifying spurious patterns in text generation tasks in future work.

4.6 Qualitative Analysis

Table 9 compares Human-CAD and CoBA for data augmentation in the SNLI dataset. In this example, the relationship between the original premise and hypothesis is neutral, while the augmented pairs exhibit a contradiction. In the Human-CAD example, the premise remains unchanged, and the label change is introduced by altering the gender in the hypothesis. Conversely, CoBA modifies

---

[4]Note that our purpose in this experiment is to verify the usefulness of CoBA, a triple-based augmentation method, in text generation tasks, rather than to mitigate spurious correlations in these tasks. Accordingly, we did not identify spurious triples. We leave the identification and mitigation of spurious correlations in text generation tasks as future work.

"outdoor" to "car" to change the label to a contradiction, leading to a contextually meaningful change without human annotation. Additionally, CoBA's modification of the premise enhances the diversity of augmented data, likely contributing to the performance improvements shown in Table 3. Additional qualitative analysis results can be found in Appendix D.1.

5 Conclusion

We introduced *counterbias data augmentation* as a more general and flexible extension of counterfactual data augmentation, capable of addressing multiple forms of bias and improving out-of-distribution robustness. Through an analysis of word importance across different models, we highlighted the limitations of using a single model to identify spurious correlations. Building on these insights, we developed CoBA, a framework that leverages LLMs to decompose text into semantic triples and apply triple-level modifications guided by a majority-voting-based ensemble. This approach enabled us to effectively mitigate spurious correlations and alleviate biases.

Our extensive experiments demonstrated CoBA's versatility and effectiveness across various tasks, including sentiment analysis, natural language inference, and text style transfer. Unlike existing counterfactual methods that emphasize minimal label-flipping modifications, CoBA allows for more diverse and semantically rich augmentations, leading to broader improvements in both accuracy and robustness.

Another potential strength of CoBA is its explainability through semantic triples. Since CoBA generates triples as an intermediate step during augmentation, it provides a clearer view of LLM behavior in text augmentation. Additionally, analyzing and comparing triples extracted from different LLMs can deepen our understanding of their mechanisms. Future research could further explore this direction to enhance LLM explainability.

Looking ahead, we plan to extend CoBA to more complex text generation scenarios, further exploring the framework's potential to mitigate spurious patterns in generated text. We envision future work refining the decomposition and reconstruction steps, optimizing the balance between information preservation and bias mitigation, and generalizing our approach to a wider range of application domains and model architectures.

Limitation

We discuss several limitations of our approach. While the CoBA framework demonstrates promising results in mitigating spurious correlations and various biases, it is important to acknowledge the inherent challenges and trade-offs associated with our methodology. In this study, we acknowledge several limitations:

Potential Information Loss in Semantic Triples. Our triple-based approach to text representation, although effective in capturing core semantic relationships, inherently entails some degree of information loss. Decomposing text into subject-predicate-object triples simplifies complex linguistic structures and potentially discards contextual nuances that could be relevant for certain tasks. This simplification, while advantageous for text manipulation and reconstruction, may inadvertently introduce new patterns that manifest as alternative forms of spurious correlations.

Trade-off Between Semantic Preservation and Flexible Augmentation. The intentional sacrifice of some semantic details is a design choice to foster diversity in the augmented data and mitigate spurious correlations. However, this compromise may not be ideal for all domains and tasks, especially where fine-grained information is essential. Furthermore, the robustness of our majority-voting ensemble method for identifying spurious correlations, while validated in our tested scenarios, may vary when applied to different types of biases or domains beyond the scope of our current evaluation.

Accuracy of Triple-level Modifications. We acknowledge that our study did not perform an extensive evaluation regarding the accuracy of triple decomposition and reconstruction through LLMs. Instead, we focused on empirical validation of performance improvement and bias mitigation through augmented counterbias data, thereby validating the effectiveness of CoBA. Nonetheless, it is important to recognize that the accuracy of these triple-level modifications may vary, and any inaccuracies in decomposition or reconstruction could potentially affect the downstream task performance. Future work should include a more granular evaluation of the fidelity of the extracted semantic triples and the consistency of the reconstructed text, as well as the development of robust metrics to quantify the precision of these modifications.

Limited Scope of Addressed Biases. Although our study is one of the first studies to alleviate multiple biases through counterbias data augmentation, our analysis primarily focuses on specific biases within the datasets examined. Although our framework has demonstrated success in these contexts, its generalizability to other types of biases remains to be fully explored. For instance, future studies in this direction could incorporate a validation on racial bias and concept bias (Field et al., 2021; Zhou et al., 2024).

Limited Usage of POS Tagging. Our approach utilizes LLMs for triple decomposition. While lightweight POS taggers could offer computational efficiency, they operate at the sentence level and struggle with complex structures. We considered comparing these methods but found that lightweight approaches often failed to extract meaningful triples from long or intricate sentences. In contrast, LLMs capture higher-order semantics beyond sentence boundaries. However, reducing computational overhead remains crucial. Future work could explore hybrid approaches, leveraging lightweight methods when feasible and applying LLMs selectively for complex cases to balance efficiency and accuracy in CoBA.

Limitations of Majority-Voting Strategy. While the majority-voting strategy employed in CoBA has shown strong performance in aggregating outputs from diverse models, it may overlook signals from particularly robust individual models in scenarios where model robustness is imbalanced. Although such extreme cases were not observed in our experiments, we recognize this as a potential limitation, particularly in other domains or under different configurations. Future work may explore adaptive voting mechanisms or reliability-weighted aggregation strategies to better account for model-specific performance differences.

Ethics Statement

In this study, the replacement of gender-related words in CoBA was based on a binary conceptualization of gender, excluding non-binary identities and other complex gender expressions. This limitation arose because the gender-related word list we used from a previous study was restricted to binary genders (Zhao et al., 2018). While this approach helped systematize the experiments, it inevitably

overlooked the nuances of gender diversity. As a result, the framework's binary focus presents an ethical limitation. We fully recognize the importance of addressing this issue and hope future research will expand counterbias data augmentation to inclusively represent non-binary genders.

Acknowledgement

References

Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. 2023. Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of ACL 2023 Student Research Workshop*, pages 119–129.

Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of ECCV*, pages 456–473.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023. Tigtec: Token importance guided text counterfactuals. In *Proceedings of ECML-PKDD (Part 3)*, pages 496–512.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of ACL*, pages 214–217.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, pages 1877–1901.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Effects of diversity incentives on sample diversity and downstream model performance in llm-based text augmentation. In *Proceedings of ACL*, pages 13148–13171.

Robin Chan, Afra Amini, and Mennatallah El-Assady. 2023. Which spurious correlations impact reasoning in nli models? a visual interactive diagnosis through data-constrained counterfactuals. In *Proceedings of ACL (Demo Track)*, pages 463–470.

Mingshan Chang, Min Yang, Qingshan Jiang, and Ruifeng Xu. 2024. Counterfactual-enhanced information bottleneck for aspect-based sentiment analysis. In *Proceedings of AAAI*, pages 17736–17744.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of ACL*, pages 1474–1483.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of EACL*, pages 1013–1025.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10526–10534.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Bowen Ding, Qingkai Min, Shengkun Ma, Yingjie Li, Linyi Yang, and Yue Zhang. 2024. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1112–1140.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. In *Proceedings of ACL*, pages 1905–1925.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*, pages 1307–1323.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased

towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of ICLR*.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of EMNLP*, pages 4398–4409.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of ICLR*.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of EMNLP*, pages 9804–9817.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of EMNLP (Demo Track)*, pages 175–184.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of LREC-COLING*, pages 13201–13221.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of ACL*, pages 5502–5515.

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

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of AACL*, pages 78–84.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Andrew 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 ACL*, pages 142–150.

Emily McMilin. 2022. Selection bias induced spurious correlations in large language models. In *Proceedings of ICML 2022 Workshop on Spurious Correlations, Invariance and Stability*.

Yifei Ming, Hang Yin, and Yixuan Li. 2022. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of AAAI*, pages 10051–10059.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu.com.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of EMNLP*, pages 1953–1967.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-08-09.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Xiaoqi Qiu, Yongjie Wang, Xu Guo, Zhiwei Zeng, Yu Yue, Yuhong Feng, and Chunyan Miao. 2024. Paircfr: Enhancing model training on paired counterfactually augmented data through contrastive learning. In *Proceedings of ACL*, pages 11955–11971.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL*, pages 129–140.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of KDD*, pages 1135–1144.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The shapley value in machine learning. In *Proceedings of IJCAI*, pages 5572–5579.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2024. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of EACL*, pages 1876–1898.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *Proceedings of ICLR*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of NeurIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*, pages 86–96.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. 2023. How re-sampling helps for long-tail learning? In *Proceedings of NeurIPS*, pages 75669–75687.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. In *Proceedings of NeurIPS*, pages 144–157.

Yuewen Sun, Erli Wang, Biwei Huang, Chaochao Lu, Lu Feng, Changyin Sun, and Kun Zhang. 2024. Acamda: Improving data efficiency in reinforcement learning through guided counterfactual data augmentation. In *Proceedings of AAAI*, pages 15193–15201.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of ICML*, pages 3319–3328.

Zhen Tan, Xiang Zhao, Wei Wang, and Weidong Xiao. 2019. Jointly extracting multiple triplets with multi-layer translation constraints. In *Proceedings of AAAI*, pages 7080–7087.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of ACL*, pages 340–351.

Marcos Treviso, Alexis Ross, Nuno M Guerreiro, and André FT Martins. 2023. Crest: A joint framework for rationalization and counterfactual text generation. In *Proceedings of ACL*, pages 15109–15126.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Proceedings of EMNLP*, pages 11308–11321.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Findings of NAACL*, pages 1719–1729.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of EMNLP*, pages 4798–4818.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of EMNLP*, pages 3431–3440.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of AAAI*, pages 14024–14031.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP*, pages 6382–6388.

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. Autocad: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of EMNLP*, pages 2302–2317.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP (Demo Track)*, pages 38–45.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of ACL*, pages 6707–6723.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Proceedings of ICCS*, pages 84–95.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Contrastive triple extraction with generative transformer. In *Proceedings of AAAI*, pages 14257–14265.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of EMNLP*, pages 2225–2239.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of EMNLP*, pages 7270–7280.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of NeurIPS*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*, pages 15–20.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. Wikibias: Detecting multi-span subjective biases in language. In *Findings of EMNLP*, pages 1799–1814.

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 ACL*, pages 478–492.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of ACL*, pages 1651–1661.

A   Experimental Setup for Word-level Importance Analysis

For experiments on important word analysis in Section 3.2, we employed four models—BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) to identify the top-importance words. For training the classifier, we set the batch size to 32, the initial learning rate of the AdamW optimizer (Loshchilov and Hutter, 2019) to 5e-5, the maximum token length to 300, and the maximum training epochs to 15. We selected the best checkpoint based on the accuracy of the validation set to extract the word importance of each model.

Additionally, the second analysis involves performing POS tagging using NLTK (Bird and Loper, 2004) and comparing the tendencies of each model concerning the POS tags of important words. Specifically, we used pre-trained English Punkt tokenizer (Kiss and Strunk, 2006). For this analysis, we employed four different models: BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). We trained these models on SST-2 (Socher et al., 2013) and IMDB (Maas et al., 2011) datasets.

B   Experimental Setup for Task Performance

We employed five models: BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and BART-base (Lewis et al., 2020) to identify w_s and w_p. In this experiment, we set the number of top-k words to 5. We used GPT-4o-mini (OpenAI, 2024) for triple decomposition and text reconstruction. The generated counterfactual data was combined with the original dataset and used to train BERT-base, DeBERTaV3-base (He et al., 2023), T5-base (Raffel et al., 2020) and ModernBERT (Warner et al., 2024) classifiers. For training the classifiers, we set the batch size to 32, the initial learning rate of the AdamW optimizer (Loshchilov and Hutter, 2019) to 5e-5, and the maximum training epochs to 10. The best checkpoint was selected based on validation set accuracy. All experiments were conducted using the Transformers library (Wolf et al., 2020).

For evaluating task performance, the experiments in this paper utilize datasets from two domains: sentiment analysis and SNLI. Specifically, for sentiment analysis, the SST-2 (Stanford Sentiment Treebank) and IMDB datasets are employed, both of which belong to the movie review domain and feature binary classification labels, positive or negative sentiment. The IMDB dataset (Maas et al., 2011), consists of approximately 50,000 reviews, while SST-2 (Socher et al., 2013) contains around 67,000 sentences. For NLI tasks, the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets are used. These datasets include three classification labels: contradiction, neutral, and entailment. SNLI comprises around 570,000 sentence pairs derived from image captions. MNLI contains approximately 433,000 sentence pairs spanning multiple domains, making it a more diverse and

challenging benchmark. The experiments leverage the Datasets library (Lhoest et al., 2021) to access and preprocess these datasets, ensuring consistency and ease of implementation across different models.

The baseline methods for comparison are as follows:

- EDA (Wei and Zou, 2019): A rule-based augmentation technique that modifies sentences through word-level modification. In this study, the modification ratio was set to 20%.

- Back-translation (Sennrich et al., 2016): An augmentation technique that translates the original sentence into a pivot language and then back-translates it into the source language.

- C-BERT (Wu et al., 2019): A strategy that leverages the contextual capabilities of the BERT model by filling in masked tokens.

- Human-CAD (Kaushik et al., 2020): This baseline uses the Human-CAD dataset, which was created by employing human annotators to generate counterfactual data from a subset of the SNLI and IMDB datasets. Specifically, we trained a model using a combination of the Human-CAD dataset and the original dataset.

- AutoCAD (Wen et al., 2022): A counterfactual data augmentation method that uses a text-infilling model.

- GPT3Mix (Yoo et al., 2021): An LLM-based augmentation technique using few-shot examples and the assignment of soft label predicted by the LLM. We used `GPT-4o-mini` for a fair comparison.

- AugGPT (Dai et al., 2023): An augmentation approach based on ChatGPT, where LLMs are prompted to generate paraphrases of original sentences. We used `GPT-4o-mini` for a fair comparison.

C Ablation Study on CoBA

To verify the effectiveness of the proposed method, we conducted an ablation study using `BERT-base` trained with CoBA on IMDB dataset. For this experiment, we systematically removed one or more components of our proposed approach to assess the contribution of each component.

Gender Bias Word Change	Triple Shuffling	
		94.9
	✓	95.7
✓		92.9
✓	✓	95.4

Table 10: Ablation study results for CoBA. '✓' indicates that the specified method was applied, while blank indicates that the specified method was not performed. The baseline model at the top refers to the method where the obtained triples are directly reconstructed without any modifications.

# of Models	SST-2 Accuracy	IMDb Accuracy
1	91.8	93.5
3	91.8	93.1
5	**94.9**	**95.4**
7	94.0	94.1

Table 11: Performance based on the number of models used in the ensemble.

The results of this experiment is presented in Table 10. We observed that solely replacing gender-related words led to a decline in performance. However, when shuffling was introduced, performance not only improved beyond the baseline but also surpassed our proposed CoBA, highlighting the significant impact of augmented data diversity on model performance. Nevertheless, it is important to note that CoBA is designed to simultaneously enhance model performance and mitigate underlying biases, providing multiple advantages for model training. Additionally, the modular structure of CoBA enables engineers to selectively utilize its components. For example, if the training dataset is confirmed to be free of significant biases, engineers can exclude the replacement procedure and use only shuffling, thereby maximizing the performance gains from data augmentation.

Also, We investigated the effect of the number of models used in the ensemble on downstream performance. The goal was to determine how model diversity impacts the robustness of generated counterfactuals. As shown in Table 11, increasing the ensemble size to five improves accuracy, while performance slightly saturates or even drops beyond that. This suggests that our choice of using five diverse models is empirically justified and not arbitrary.

In addition, we conducted an ablation study on

Top-k Words	SST-2 Accuracy	IMDb Accuracy
1	90.7	92.2
3	91.8	91.7
5	**94.9**	**95.4**
10	94.1	94.9
20	93.5	92.8

Table 12: Performance based on number of top-k words selected per model.

the number of top-ranked words selected from each model when generating counterfactuals. As shown in Table 12, selecting the top-5 words per model yields the highest accuracy across both datasets. Choosing too few limits diversity, while too many introduces noise, validating our design choice of top-5 selection as a balanced and effective configuration.

D   Additional Analysis

D.1   Qualitative Analysis on COBA

In addition to the qualitative analysis in Section 4.6, we provide further examples. Table 13 compares augmented data generated by Human-CAD and COBA using the IMDB dataset. The original data and its Human-CAD augmentations, modified by human annotators, demonstrate that while the overall syntactic structure of the text is preserved, key expressions are systematically altered. Notably, negative expressions are explicitly transformed. For instance, the phrase "time line is screwed up" was modified to "time line is accurate." However, it should be noted that the most crucial phrase determining sentiment in this sentence is "not great."

Whereas Human-CAD modifies most expressions that contribute to the label, COBA instead preserves less critical phrases like "time line is screwed up" and focuses on altering key sentiment-related information, such as changing "Overall, it is not great" to "Overall, it is great." This targeted modification ensures that the augmented data remains close to the local decision boundary, thereby enhancing model performance (Gardner et al., 2020). Moreover, by shuffling the order of triples before reconstructing the text, COBA generates more syntactically diverse expressions compared to Human-CAD.

D.2   Qualitative Analysis on POS Tagging Analysis

Table 2 presents the differences in POS tag distributions across various models. To examine how each model assigns POS tags to principal words, we provide examples of principal words selected by three models using IG (Sundararajan et al., 2017) in Table 14. A notable finding is that the BERT-base model identified the proper noun "Nancy" as the most important word. This aligns with prior research indicating that certain proper nouns are often associated with specific labels, leading to spurious correlations (Wang and Culotta, 2020). Analyzing sentences from the IMDB dataset containing "Nancy", we found that approximately 70

Typically, words like "enjoy" or "entertaining" are considered crucial for sentiment analysis. However, not all models prioritized these words. For example, the RoBERTa-base model identified "succeeded" as a principal word. While "succeeded" can imply positivity, it carries little significance in the given sentence. This variation suggests that each model exhibits different tendencies toward spurious correlations, leading to inconsistencies in POS tagging for important words and demonstrating divergent patterns across models.

D.3   Discussion Regarding POS Tagging Analysis

In our POS tagging analysis in Section 3.2.2, we acknowledge that dataset characteristics significantly impact the observed distributions of POS tags. Specifically, we found that the IMDB dataset exhibited a notably high frequency of the "Others" category, which can be attributed to the nature of the dataset itself. Unlike structured datasets such as SST-2, IMDB consists of lengthy internet comments, which often contain abbreviations, informal language, and exaggerated expressions. These linguistic traits contribute to a greater proportion of words that do not fall neatly into standard noun, verb, adjective, or adverb categories, thereby inflating the "Others" category.

However, despite this dataset-induced variation, we also observed model-specific differences in POS tagging tendencies. For instance, while all models predominantly identified nouns as the most frequent POS category in the SST-2 dataset, RoBERTa exhibited a relatively higher proportion of nouns in IMDB compared to other models, whereas BART demonstrated a more evenly distributed POS ra-

Original [Label: Negative]
Penn takes the time to develop his characters, and we almost care about them.
However there are some real problems with the story here,
we see no real motivation for the evil brother's behavior, and the time line is screwed up.
Supposedly set in 1963, the music is late 60s/early 70s.
The references and dialogue is 70s/80s.
The potential for a powerful climax presents itself, and Penn allows it to slip away.
But even with all these difficulties it is worth the watch, but not great.
Human-CAD [Label: Positive]
Penn takes the time to develop his characters, and we truly care about them.
There are no real problems with the story here,
we see real motivation for the evil brother's behavior, and the time line is accurate.
Set in 1963, the music is early 60s. The references and dialogue are 60s too.
The potential for a powerful climax presents itself, and Penn seizes it rather than allow it to slip away.
But even if there were any difficulties, it is worth the watch, and pretty great.
CoBA [Label: Positive]
The timeline is screwed up, and the references and dialogue are 70s/80s,
even though the story is supposedly set in 1963.
Despite these issues, there are some interesting elements in the story.
Penn takes time to develop his characters,
and the potential for a powerful climax presents itself, which Penn delivers effectively.
Overall, it is great, although the music is from the late 60s/early 70s.
It is definitely worth the watch, even though we see no motivation
for the evil brother's behavior.
Nevertheless, we genuinely care about the characters.

Table 13: The comparison of augmented data generated by Human-CAD and CoBA on IMDB.

Original
I really enjoyed this movie.
It succeeded in doing something that few movies do now;
it provided family values while entertaining me.
Nancy Drew is a heroine for all generations
and a role model for young girls to look up to.
BERT-base
I really enjoyed this **movie**.
It succeeded in doing something that few movies do now;
it provided family values while **entertaining** me.
Nancy Drew is a heroine for all generations
and a role model for young girls to look up to.
DistilBERT
I really **enjoyed** this movie.
It succeeded in doing something that few movies do now;
it provided **family** values while **entertaining** me.
Nancy Drew is a heroine for all generations
and a role model for young girls to look up to.
RoBERTa-base
I really **enjoyed** this movie.
It **succeeded** in doing something that few movies do now;
it provided **family** values while entertaining me.
Nancy Drew is a heroine for all generations
and a role model for young girls to look up to.

Table 14: Principal words for each models, BERT-base, DistilBERT, and RoBERTa-base.

tio across categories. This suggests that, although dataset properties play a crucial role in shaping POS distributions, model architectures also influence how certain word types are emphasized.

To further clarify the composition of the "Others" category, we conducted an in-depth analysis of the IMDB dataset. We found that determiners, numerals, and WH-words (e.g., "who," "what," "where") appeared in similar proportions within this category. Since no single POS type overwhelmingly dominated among these, we originally grouped them under "Others" to maintain clarity. In the revised version of this work, we explicitly define which POS types are included in this category to enhance interpretability.

E Preliminary Evaluation on Racial Bias.

To assess CoBA's effectiveness beyond gender and spurious correlation biases, we conducted a preliminary experiment on the Measuring Hate Speech Corpus, which includes racially biased language. We evaluated the downstream classification accuracy of a DeBERTaV3 model under various augmentation strategies. As shown in Table 15, CoBA outperformed both traditional and LLM-based baselines, indicating its potential applicability to racial bias mitigation as well.

Method	Accuracy (%)
Baseline w/o Augmentation	91.1
EDA	90.7
AugGPT	92.2
CoBA	**92.6**

Table 15: Performance on Measuring Hate Speech Corpus using DeBERTaV3.

Method	Variance (%)
EDA	73.6
C-BERT	72.1
GPT3Mix	65.1
AugGPT	64.9
CoBA	**64.3**

Table 16: Proportion of variance explained by the top 50 principal components (lower is better).

F Diversity Analysis Using PCA.

To further evaluate how triple permutation affects the diversity of augmented data, we performed a principal component analysis (PCA) on the embeddings of generated samples from each augmentation method. We used OpenAI's `text-embedding-3-small` model to obtain sentence-level representations and measured the proportion of variance explained by the top 50 principal components. A lower explained variance suggests higher dispersion in the embedding space and thus greater data diversity. As shown in Table 16, CoBA achieved the most diverse outputs among the compared methods, including both traditional and LLM-based baselines.

G Cost Analysis

In our experiments, we used GPT-4o-mini, which is priced at 0.15 USD per million input tokens and 0.60 USD per million output tokens. While our approach involves two stages—semantic triple decomposition and sentence reconstruction—we conducted a detailed cost analysis using the actual prompts and the tiktoken tokenizer [5]. On average, the decomposition step consumes approximately 250 tokens, resulting in a cost of roughly 0.0005 USD per sample.

When applied to a dataset like IMDB, the total cost adds up to only around 2 USD, which is com-

parable to other LLM-based augmentation methods such as AugGPT and GPT3Mix. Although we acknowledge that our method does incur a slightly higher cost (about 2 USD more), we believe that this is a reasonable trade-off given the stronger bias mitigation and robustness improvements demonstrated in our results.

---

[5] https://github.com/openai/tiktoken

H Prompt Design

H.1 Prompt for Semantic Triple Decomposition (Sentiment Analysis)

System: You are a chatbot used for data augmentation. I will provide two paragraphs or internet comments for natural language understanding (NLU) tasks or sentiment analysis tasks.
User: Please create semantic triples for the following sentence.
Triple consists of three elements: subject, predicate, and object.

Here is an example of a sentence and its corresponding Semantic Triplet: A few people in a restaurant setting, one of them is drinking orange juice.

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice

Here is another example of a sentence and its corresponding Semantic Triplet: A poor work that failed to provide a proper narrative for the black woman.
1. A work | is | poor
2. A work | failed to provide | a proper narrative
3. A proper narrative | is for | the black woman

Please provide no answers other than the semantic triplet. Output only the semantic triplet.

Here is a paragraph you should make a semantic triplet:
Content

H.2 Prompt for Reconstructing Triples into Sentences (Sentiment Analysis)

System: You are a chatbot used for data augmentation. Your job is reconstructing the selected triples into a sentence or paragraph.
User: Please create sentences for the following Triples.
Here is an example of a Semantic Triples and its corresponding reconstructed text:

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice
Output format:
A few people in a restaurant setting, one of them is drinking orange juice.

Here is another example of a Semantic Triples and its corresponding reconstructed text:
2. I | am | a student
1. I | am | a professor
Output format:
I am a student and also a professor.

Please provide no answers other than the reconstructed text. Output only the reconstructed text. And don't consider the number of sentences in the input text.

Please follow the order of the inputs strictly as they are written. Do not consider the numbers provided in the inputs.
For example:
2. I | am | a student
1. I | am | a professor
Output format:
I am a student and also a professor.
In this case, even though the sequence numbered "2" comes first numerically, ignore the numbers and generate the output starting with "I | am | a student" as shown in the example.

Here is a Semantic Triples you should make a text:
Content

H.3 Prompt for Semantic Triple Decomposition (Natural Language Inference)

System: You are a chatbot used for data augmentation. I will provide two paragraphs or internet comments for natural language understanding tasks. This natural language understanding task has a label of entailment, contradiction, or neutral.
User: You should creating semantic triples from the following paragraph, and select the most important semantic triples. Your task is to receive two sentences along with the label for a natural language understanding task corresponding to those sentences. For each sentence, you need to create semantic triples.
Here is an example of two input sentence and label:

sent1: A woman is walking across the street eating a banana, while a man is following with his briefcase.
sent2: An actress and her favorite assistant talk a walk in the city.
label: neutral

Here is an output example of semantic triples:

sent1:
1-1. A woman | is walking | across the street
1-2. A woman | is eating | a banana
1-3. A man | is following | a woman
1-4. A man | is carrying | a briefcase

sent2:
2-1. An actress | is walking | in the city
2-2. An actress | is with | her favorite assistant
2-3. An actress and her favorite assistant | are talking | while walking

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice

Here is an another example of two input sentence and label:

sent1: Two women, holding food carryout containers, hug.
sent2: Two groups of rival gang members flipped each other off.
label: contradiction

Here is an output example of above example:

sent1:
1-1. Two women | are holding | food carryout containers
1-2. Two women | hug | each other

sent2:
2-1. Two groups of rival gang members | flipped | each other off

Please provide no answers other than the semantic triplet. Output only the semantic triples.

Here is a paragraph you should make a semantic triplet:
Content

H.4 Prompt for Reconstructing Triples into Sentences (Natural Language Inference)

System: You are a chatbot used for data augmentation. I will provide triples for natural language understanding tasks. This natural language understanding task has a label of entailment, contradiction, or neutral.
User: You should reconstruct the semantic triples into a sentence or paragraph. Don't change other triplet. Then reconstruct the semantic triples into a sentence or paragraph.
Here is an example of two input triples and label:

sent1:
1-1. An older woman | sits | at a small table
1-2. An older woman | has | orange juice
1-3. Employees | are smiling | in the background
1-4. Employees | are wearing | bright colored shirts

sent2:
2-1. A girl | flips | a burger

label: contradiction

Here is example of output:

reconstructed sent1:
An older woman sits at a small table with a glass of orange juice, while employees in bright-colored shirts smile in the background.
reconstructed sent2:
A girl flips a burger.

Here is another example of two input triples and label:

sent1:
1-1. The school | is having | a special event
1-2. The special event | is to show | American culture
1-3. American culture | deals with | other cultures in parties

sent2:
2-1. A school | is hosting | an event

Here is example of output:

reconstructed sent1:
The school is having a special event in order to show the american culture on how other cultures are dealt with in parties.
reconstructed sent2:
A school is hosting an event.

Please follow the example format exactly and only output the necessary graph triplets. Do not start with conversational phrases like "Here's" or "Sure."

Here is an semantic triples you should reconstruct:
Content