LLM-Assisted Content Conditional Debiasing for Fair Text Embedding

Anonymous EMNLP submission

Abstract

Mitigating biases in machine learning models has become an increasing concern in Natural Language Processing (NLP), particularly in developing fair text embeddings, which are crucial yet challenging for real-world applications like search engines. In response, this paper proposes a novel method for learning fair text embeddings. First, we define a novel content-conditional equal distance (CCED) fairness for text embeddings, ensuring content-conditional independence between sensitive attributes and text embeddings. Building on CCED, we introduce a content-conditional debiasing (CCD) loss to ensure that embeddings of texts with different sensitive attributes but identical content maintain the same distance from the embedding of their corresponding neutral text. Additionally, we tackle the issue of insufficient training data by using Large Language Models (LLMs) with instructions to fairly augment texts into different sensitive groups. Our extensive evaluations show that our approach effectively enhances fairness while maintaining the utility of embeddings. Furthermore, our augmented dataset, combined with the CCED metric, serves as an new benchmark for evaluating fairness.

1 Introduction

Embedding text into dense representations is a widely used technique in modern NLP, powering applications such as sentiment analysis (Dang et al., 2020), recommendation systems (Zhang et al., 2016), and search engines (Palangi et al., 2016). However, the extensive use of these embeddings introduces inherent biases that can affect various applications (Packer et al., 2018; Baeza-Yates, 2018; Zerveas et al., 2022; Rabelo et al., 2022). For instance, search engines (Huang et al., 2020) preprocess all text contents and search queries into embeddings to optimize storage and enable efficient similarity matching. These inherent biases in text embeddings can influence the calculation of embedding similarity, impacting the filtering of numerous documents to find pertinent ones. Moreover, text embeddings are directly employed in other applications such as zero-shot classification (Yin et al., 2019; Radford et al., 2021) and clustering (John et al., 2023). Unfortunately, various forms of biases, including gender, racial, and religious biases, have been identified in text embeddings generated by pre-trained language models (PLMs), as reported in several studies (Bolukbasi et al., 2016; Nissim et al., 2020; Liang et al., 2020; May et al., 2019). Consequently, attaining fairness in text embedding models is crucial.

Recent debiasing techniques (Liang et al., 2020; Kaneko and Bollegala, 2021) for text embeddings use post-training to address biases, avoiding the inefficiency of retraining sentence encoders for each new bias. When removing bias, projection-based methods (Liang et al., 2020; Kaneko and Bollegala, 2021) reduce an embedding's projection onto each bias subspace. The distance-based method (Yang et al., 2023) constructs embeddings for sensitive groups and equalizes distances to text embeddings across these groups. Nevertheless, these methods persist in pursuing independence between sensitive attributes and text embeddings, which results in the complete removal of sensitive information. As a result, these approaches do not effectively find the *sweet spot* between fairness and utility trade-off (Zhao and Gordon, 2022; Deng et al., 2023; Zliobaite, 2015).

Recent studies (Mary et al., 2019; Deng et al., 2023; Pogodin et al., 2022) suggest that using datasets labeled with sensitive information to achieve conditional independence — specifically, conditioning on the content class to preserve semantic information within the text — provides a more effective approach to achieving fairness while preserving utility. Yet, the scarcity of text datasets with sensitive labels (Gallegos et al., 2023) limits the practical application of these findings. To

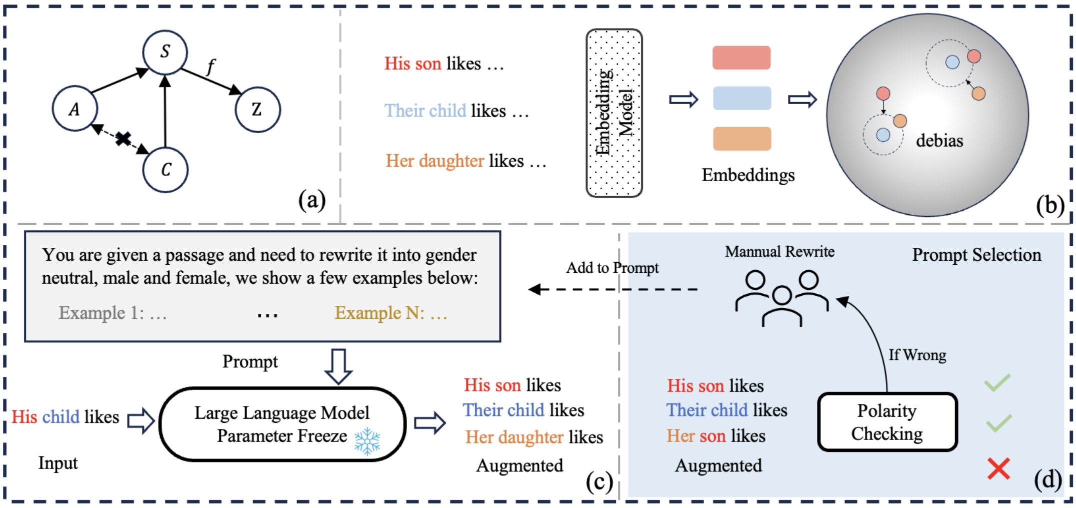

Figure 1: Pipleline of our method with **gender** as the sensitive attributes. (a) Graphical demonstration of the fairness issue. (b) The debiasing procedure achieves a content-conditioned equal distance to improve the fairness. (c) Overview of the data augmentation strategy, including the prompt template used to replace sensitive words with their equivalents from all sensitive groups. (d) Prompt search module: Augmented texts are sent to the demographic polarity checking block. Incorrectly augmented samples are then manually labeled and added to the prompts.

create such datasets, Counterfactual Data Augmentation (CDA) (Zhao et al., 2018) collects sensitive-related words and employs a rule-based method to augment the data, but this approach encounters challenges due to the need for an extensive list of words. Finally, while Large Language Models (LLMs) (Schick and Schütze, 2021; Shao et al., 2023) have offered new methods for data generation thanks to their rich contextual knowledge, yet they still struggle with inherent systematic biases (Yu et al., 2023).

In this paper, we improve the text embeding fairness through defining fairness with theoretical analysis, a novel debiasing loss design, and an LLM-based data strategy for dataset generation. Our contributions include:

- Introducing `CCED` fairness for text embeddings, ensuring equal sensitive information and conditional independence between sensitive attributes and embeddings.
- Proposing `CCD` loss to achieve the desired `CCED` fairness by ensuring that texts with varied sensitive attributes but identical content have embeddings equidistant from their neutral counterparts.
- Employing LLMs to augment datasets fairly, representing diverse sensitive groups within the same content for effective training with `CCD`. Proposing polarity-guided prompting to ensure the LLM-generated data quality and minimize the potential biases from LLMs.
- Establishing `CCED` fairness as a benchmark for evaluating fairness in text embeddings.
- Extensive evaluations on debiasing benchmarks and downstream tasks demonstrate `CCD`'s effectiveness in promoting fairness while preserving utility.

2 Related Work

Debias Text Embedding: Bias in text embeddings (also known as sentence embedding) is a significant issue that arises when these models reflect or amplify societal stereotypes and prejudices found in their training data. To resolve the issue, (Liang et al., 2020) contextualizes predefined sets of bias attribute words to sentences and applies a hard-debias algorithm (Bolukbasi et al., 2016). Contextualized debiasing methods (Kaneko and Bollegala, 2021; Yang et al., 2023) apply token-level debiasing for all tokens in a sentence and can be applied at token- or sentence-levels (Kaneko and Bollegala, 2021) to debias pretrained contextualized embeddings. However, all the above methods aim to strictly achieve independence between text embedding and sensitive attributes, which may not balance fairness and utility well. While Shen et al. (2021, 2022) employ contrastive learning losses to mitigate biases in language representations for text classification, their approach relies on supervised data, which is often scarce and expensive to obtain, and primarily focuses on fairness in the subsequent task. Additionally, although (Leteno et al., 2023; Shen et al., 2022) observe that representational fairness and group fairness in subsequent tasks are either not correlated or only partially correlated, it is important to note that fairness in subsequent tasks and fairness in text embeddings are distinct areas, with the latter being crucial for various applications. A detailed discussion of these differences can be found in Appendix A.2. In this paper, we utilize LLMs to augment training data for learning fair text embeddings with proposed `CCD` loss.

LLMs for Dataset Generation: Leveraging the success of LLMs, researchers have begun using them to generate various forms of training data, such as tabular data (Borisov et al., 2022), relation triplets (Chia et al., 2022), sentence pairs (Schick and Schütze, 2021; Zhang et al., 2024), and instruction data (Shao et al., 2023; Wu et al., 2024). As we focus on obtaining data with sensitive attribute information, data generation for text classification would be the most similar one among those applications. Recent efforts in generating data for text classification (Meng et al., 2022; Ye et al., 2022; Wang et al., 2019) primarily employ simple class-conditional prompts while focusing on mitigating issues of low quality after generation. However, these efforts encounter the challenge of inherent systematic biases present in LLMs (Yu et al., 2023). While Yu et al. (2023) considers generated data bias, it focuses only on the diversity of topics and overlooks the inherent bias within words in a text (e.g. 'child' occurs more frequently with'mother'). In this paper, we instructs the LLM to only locate the gendered words and replace them with counterparts from other groups and propose polarity-guided prompt searching to minimize biases from LLMs and ensure the quality of augmented data.

3 Method

3.1 Problem Setting

This section outlines the problem of fairness in text embeddings. We define several key variables: $S \in \mathcal{D}$ represents the input text from the data distribution, C denotes the content of the text,[1] and $A = [a_1, \dots, a_{|A|}]$ represents the sensitive attributes (e.g. gender and age). The symbol n indicates neutral, meaning no sensitive information is present. A text with content C is considered neutral S_C^n if it contain no sensitive information, whereas text $S_C^{a_i}$ is associated with the sensitive attribute a_i if its sensitive polarity (Wang et al., 2023) is a_i, see Eq. (6). The text embedding model f processes a text into a d-dimensional embedding $Z \in \mathbb{R}^d$. The embedding of a neutral text encodes the content information C' (a well trained model $C' \approx C$), while the embedding of a sensitive text additionally encodes sensitive information. Words in the text related to the attribute a_i are denoted as X^{a_i}, and neutral words are denoted as X^n. For clarity, we provide detailed notations in Table 8 in Appendix.

Fairness Issue: Fig. 1 (a) shows there exists an association between attributes A and content variable C. If model f superficially treats A as a proxy for C,[2] it results in encoded C' being represented by A thus embedding Z will mainly contain sensitive information, which leads to issues of fairness.

Fairness Goal: Mitigating fairness is not trivial, as we need to address not only bias mitigation but also the protection of the model's representation ability. As shown in Fig. 1 (a), our method aims to (1) break the association between content C and the sensitive attribute A, and (2) preserve useful sensitive information in the text embedding. For example, in the case of a text about a father raising a child, its embedding should retain information about the father.

3.2 Content Conditional Debiasing

To break the superficial association, we propose to achieve conditional independence between sensitive attributes and content $A \perp C' \mid C$. The conditional independence allows prediction C' to depend on A but only through the content variable C, prohibiting abusing A as a *proxy* for C thus mitigating the fairness issue while preserving the utility. To protect utility, our objective is not to completely remove sensitive information but to ensure that text embeddings from different sensitive groups with identical content contain an equal amount of sensitive information.

3.2.1 Fairness Definition

Firstly, we propose a novel content conditional equal distance fairness for fair text embedding:

Definition 3.1. *(Content Conditional Equal Distance (CCED) Fairness.) Let S_C^n be a neutral text with content C. Assume $S_C^A = [S_C^{a_1}, S_C^{a_2}, ..., S_C^{a_{|A|}}]$*

[1]For instance, the texts 'he is a teacher' and 'she is a teacher' both convey the same content C = 'is a teacher'.

[2]For instance, raising children is frequently associated with women in the training corpus, resulting in the proxy effect.

being a set of texts from all sensitive groups with the same content C. Then, embedding model f is content conditioned equal distance fair with respect to attributes A, for any $a_i, a_j \in A$:*

$$\|f(S_C^{a_i}) - f(S_C^n)\| = \|f(S_C^{a_j}) - f(S_C^n)\|, \quad (1)$$

where $\|\cdot\|$ is L_2 norm.

As shown in Fig. 1 (b), CCED fairness requires that texts with the same context from different sensitive groups have equal distance to their corresponding neutral text on the embedding space. This text embedding fairness definition has two merits: Equal sensitive information: The equal distance to the neutral embedding ensures an equitable encoding of sensitive information across diverse groups, allowing fair usage of sensitive information and preserving the utility of embeddings.
Content Conditional Independent: Echoing the methodologies in previous research (Hinton and Roweis, 2002; Yang et al., 2023), the conditional independence $A \perp C' \mid C$ can be represented as the CCED on the embedding space:

Assumption 3.2. *(Equal Probability) Within a content C, the likelihood $P(a_i|C)$ on all sensitive attributes $a_i \in A$ is uniform $P(a_1|C) = ... = P(a_A|C)$.*

Theorem 3.3. *When the equal probability assumption holds, achieving content conditioned equal distance fairness is equivalent to achieving conditional independence between sensitive attributes and content $A \perp C' \mid C$.*

Assumption 3.2 is true for a fair dataset that has balanced texts from all groups within content C (can be obtained through our data augmentation strategy in Section 3.3). Theorem 3.3 demonstrates the merit of CCED fairness (Definition 3.1) in achieving embedding fairness. Detailed proof can be found in Appendix A.5.

3.2.2 Content Conditional Debiasing Loss

Based on the defined CCED fairness, we design a loss function L_{bias} that aims to mitigate biases while preserving the representation ability of PLMs. For a sample pair $[S_C^{a_1}, ..., S_C^{a_{|A|}}, S_C^n]$:

$$L_{bias} = \sum_{i \in [A]} \sum_{j \neq i} |dist(f(S_C^{a_i}), f(S_C^n)) - dist(f(S_C^{a_j}), f(S_C^n))|, \quad (2)$$

where $dist(A, B) = \exp\left(-\frac{\|A-B\|^2}{2\rho^2}\right)$ measures the distance on the embedding manifold (Yang et al., 2023; Hinton and Roweis, 2002) (details in Appendix A.5), and ρ is selected as the variance of the distance over the training dataset for normalization. To further preserve the valuable information encoded in the model and achieve efficient debiasing, we design L_{rep} to enforce high similarity between the neutral texts' embeddings processed by the fine-tuned model f and those processed by the original model f^{org}:

$$L_{rep} = \|f(S^n) - f^{org}(S^n)\|. \quad (3)$$

Ensuring that neutral embeddings remain unchanged offers two benefits: preserving the model's representational capability and maintaining neutral embeddings as a consistent reference point in the debiasing loss, ensuring stable equal distance to embeddings with various sensitive attributes. Thus, the overall training objective is:

$$L_{all} = L_{bias} + \beta * L_{rep}, \quad (4)$$

where β is a hyper-parameter used to balance the two terms. An ablation study for setting β is detailed in Table 7.

3.3 LLM-Assisted Content Conditional Data Augmentation

We leverage the rich contextual knowledge of LLM with few-shot prompting to obtain a dataset that (1) fulfills the Assumption 3.2 to achieve our goal in Definition 3.1 as well as (2) avoids introducing inherent bias in LLM to augmented data. The data augmentation algorithm is shown in Alg. 1, followed by a detailed explanation below.

Augment Text into Different Sensitive Groups: As shown in Fig. 1 (c), our task description T instructs the LLM to only locate the gendered words and replace them with counterparts from other groups, leaving the other content unchanged thus avoiding fairness issues in text generation. Specifically, for sensitive words $X^A = [X^{a_i}, ..., X^{a_j}], a_i, a_j \in A$ in the text S, the LLM h substitutes X^A with words from different sensitive groups and neutral terms, thus obtaining augmented texts from all sensitive groups (as shown in Table 1):

$$h(S, T, P) = [S^{a_1}, ..., S^{a_{|A|}}, S^n], c \quad (5)$$

where c is the confidence score and P is the example prompts (detailed in Table 10 in Appendix). After augmentation, the dataset will have an equal

Algorithm 1 Data Augmentation Algorithm

Input: Dataset $\mathcal{D}$, Sensitive word lists V, Pretrained LLM h, Task Description T, Example Prompts P.

```
for k in 1, ..., K do        ▷ K = 10 in this work
    Block I: Augment Texts into Different Sensitive Groups
    for S ∈ D do
        h(S, T, P) → [S^{a_1}, ..., S^{a_{|A|}}, S^n], c
    end for
    if k = K then
        return Augmented Dataset D'
    end if
    Block II: Polarity Guided Prompt Searching
    for [S^{a_1}, ..., S^{a_{|A|}}, S^n] ∈ D' do
        Polarity Checking Eq.6
    end for
    Manually Augment the wrong augmentation with highest c and add to P.
end for
```

amount of texts from each sensitive group with identical content, meeting our equal probability Assumption 3.2.

Polarity-Guided Prompt Searching: To ensure the quality of augmented texts and the effectiveness of few-shot prompt tuning on LLMs, finding appropriate prompts P is crucial. We propose identifying difficult samples from incorrectly augmented texts to use as prompts. First, these incorrectly augmented samples are detected through a sensitive polarity check as described by (Wang et al., 2023) and illustrated in Fig. 1(d). By counting the occurrences of words in predefined sensitive word lists $V = [V^{a_i}, ..., V^{a_j}], a_i, a_j \in A$, the polarities of a series of sentences are determined as follows:

$$g(S) = \arg\max_{a_i \in A} occ(S, V^{a_i}), \tag{6}$$

where occ represents the number of times words from the list V^{a_i} appear in all augmented sentences S. For a properly augmented sentence S^{a_i}, its polarity should match the sensitive attribute a_i. If $g(S^{a_i}) \neq a_i$, the sentence is considered inaccurately augmented. Then we introduce our prompt searching strategy in Algorithm 1. In each iteration, the algorithm identifies the incorrectly augmented sample with the highest confidence c, manually augments it, and adds it to the example prompts P. This rule-guided prompt search is repeated K times (with $K = 10$) to prepare samples for the few-shot prompt tuning of de-biasing LLMs.

4 Experiments

In this paper, we take *gender* bias as an example due to its broad impact on society.

Datasets: We utilize the News-commentary-v15 corpus (Tiedemann, 2012) as source samples to generate our training data with LLMs. For gender bias evaluation, we follow (Yang et al., 2023) to use SEAT (May et al., 2019), CrowS-Pairs (Nangia et al., 2020) and StereoSet-Intrasentence data (Nadeem et al., 2020). We additionally assess fairness on longer texts via the Bias-IR dataset (Krieg et al., 2023). To evaluate whether the biased models' representation ability is maintained, we follow (Kaneko and Bollegala, 2021; Yang et al., 2023) to select four small-scale subsequent tasks from the GLEU benchmark: Stanford Sentiment Treebank (SST-2 (Socher et al., 2013)), Microsoft Research Paraphrase Corpus (MRPC (Dolan and Brockett, 2005)), Recognizing Textual Entailment (RTE (Bentivogli et al., 2009)) and Winograd Schema Challenge (WNLI (Levesque et al., 2012)). More dataset information see Appendix A.3.

Backbone and Baseline Methods: For the selection of PLMs, we choose BERT-large-uncased (Devlin et al., 2018) and RoBERTa-base (Liu et al., 2019). To assess debiasing performance, we compare our algorithm with finetuning-based methods DPCE (Kaneko and Bollegala, 2021) and ADEPT-F (Yang et al., 2023). To assess the effectiveness of our data augmentation strategy, we compare our approach with CDA (Zhao et al., 2018).

LLM-Assisted Data Augmentation: We leverage ChatGPT (i.e., `gpt-3.5-tubo`) and Gemini (Team et al., 2023) to generate our training data. We obtained a dataset with texts of content C from all groups A and neutral. Using Gemini and ChatGPT for data augmentation resulted in datasets with 43,221 and 42,930 sample pairs, respectively. Examples of data augmented through our method are presented in Table 1, and the quality of the augmented dataset is assessed in Section 4.1.

Hyperparameters: We use Adam to optimize the objective function. During the debiasing training, our learning rate is 5e-5, batch size is 32, and β is 1. Our method requires training for only a single epoch and selecting the checkpoint with the

Gender	Generated Text
Male	But because Rumsfeld wanted to prove a point about transforming strategy. After championing the continuation of his hardline policy, his current strategy of negotiation is risky. He has been very vocal in voicing discontent with the rule of Kirchner and that of his husband and predecessor, Néstor Kirchner.
Neutral	But because the individual wanted to prove a point about transforming strategy. After championing the continuation of their hardline policy, the current strategy of negotiation is risky. They have been very vocal in voicing discontent with the rule of Kirchner and that of their spouse and predecessor, Néstor Kirchner.
Female	But because Rachel wanted to prove a point about transforming strategy. After championing the continuation of her hardline policy, her current strategy of negotiation is risky. She has been very vocal in voicing discontent with the rule of Kirchner and that of her wife and predecessor, Néstor Kirchner.

Table 1: We utilize LLM to augment text into three gender categories: Male, Female, and Neutral. Below are sample examples of the generated data, where words containing gender information are highlighted in colors: red for male, blue for neutral, and orange for female.

lowest validation loss (validate every 500 steps). The results for DPCE and ADEPT-F are obtained using the originally reported hyperparameters from the studies by (Kaneko and Bollegala, 2021; Yang et al., 2023). Consistent with these studies, we set the random seed to 42 to ensure a fair comparison. All experiments are conducted on an NVIDIA A100 GPU.

4.1 Augmentation Quality Checking

To demonstrate the quality of our augmented data on gender, we quantitatively assess the fairness of our augmented dataset using the union gender polarity accuracy metric, formulated as follows:

$$g_i^u = \left(g(S_i^n) = n \cap g(S_i^m) = a_m \cap g(S_i^f) = a_f\right)$$
$$Acc = \frac{\sum_{i=1}^N g_i^u}{N}, \tag{7}$$

where $[S_i^n, S_i^m, S_i^f]$ are the augmented texts for the i-th sample, N denotes the size of the augmented dataset, and $g(\cdot)$ is the polarity checking function as defined in Eq. (6). The union gender polarity accuracy metric measures the proportion of text triples (neutral, male, female) that are accurately augmented in alignment with their respective gender polarities. The results show both Gemini and GPT models achieve high accuracy, with Gemini and GPT reaching 83.4% and 82.2% respectively . This suggests that our data augmentation process has effectively produced a fair dataset. Incorporating polarity checking as a post-processing step further ensures the fairness of our augmented data.

4.2 Results and Analysis

We evaluate four models on all benchmarks, namely the original model (pre-trained with no explicit debiasing), the DPCE model, the ADEPT-F model, and our CCD.

Reducing Gender Biases: In Table 2 and Table 3, our experiments demonstrate that CCD with GPT and Gemini data strategies excels in debiasing, consistently outperforming baselines in the StereoSet and CrowS-Pairs datasets for both BERT and RoBERTa backbones. On SEAT, both CCD and DPCE achieve good performance, with CCD-Gemini achieving the best overall performance on SEAT across both backbones. Notably, our method attains a high ICAT score in the StereoSet dataset, indicating an excellent balance between performance and fairness. However, while DPCE maintains great fairness, it adversely affects its representation capability, as evidenced by a significantly lower LMS score in the StereoSet dataset.

Preserving Representation Ability: In Table 4 and Table 5, the GLUE results demonstrate that CCD-Gemini achieves the highest average performance with both BERT and RoBERTa backbones, suggesting that our CCD even enhances the model's representation capabilities. Conversely, DPCE, which strictly separate gender attributes from neutral text embeddings, harms the model's utility.

Bias in Information Retrieval: Since search engine performance is a crucial subsequent task of text embedding usage, we evaluate the bias in information retrieval using the Bias-IR dataset. For the BERT model, Table 4 shows that CCD-Gemini achieves the best fairness, with CCD-GPT ranking second. For the RoBERTa model, Table 5 demonstrates CCD-GPT achieves the best fairness, with CCD-Gemini ranking second. Overall, CCD with GPT and Gemini data strategies outperforms baselines in fairness across various fields, as well as in average fairness.

CCED as Fairness Metric: We use our CCED fairness from Definition 3.1 to evaluate fairness. Specifically, we calculate the CCED gap for all methods on our Gemini-augmented dataset using the equation $\frac{1}{N}\sum_i^N |\|f(S_i^{a_i}) - f(S_i^n)\| - \|f(S_i^{a_j}) - f(S_i^n)\||$. Table 6 demonstrates that CCD achieves the best fairness on the CCED fairness metric and

Datasets	SEAT (0.00) the best							StereoSet:gender			StereoSet:all			CrowS-Pairs
Method	6	6-b	7	7-b	8	8-b	AVG (abs)↓	LMS↑	SS (50.00)	ICAT↑	LMS↑	SS(50.00)	ICAT↑	SS(50.00)
BERT	0.37	0.20	0.42	0.22	−0.26	0.71	0.36	86.34	59.66	69.66	84.16	58.24	70.29	55.73
DPCE	−0.21	0.27	0.44	0.07	0.25	0.21	0.24	81.19	56.72	65.41	64.06	52.96	60.26	52.29
ADEPT-F	0.83	−0.14	0.63	1.24	0.43	1.28	0.76	86.45	61.70	66.21	85.09	57.52	72.26	51.91
DPCE-Gemini	0.63	0.41	0.00	−0.01	0.19	0.17	**0.23**	82.63	60.68	64.98	64.08	54.91	57.78	51.53
ADEPT-F-Gemini	0.71	−0.23	0.21	0.92	0.35	0.99	0.57	86.80	61.72	66.44	85.47	58.50	71.71	51.91
CCD-CDA	0.16	0.03	0.43	0.38	0.47	0.22	0.29	80.34	**53.53**	74.69	79.10	53.46	73.62	46.95
CCD-GPT	0.35	−0.11	−0.17	−0.15	0.57	0.06	**0.23**	81.47	53.60	**75.60**	80.22	**52.83**	**75.97**	47.71
CCD-Gemini	0.47	−0.00	−0.02	−0.72	−0.30	0.07	0.26	82.91	54.93	74.72	82.97	55.00	74.67	**48.85**

Table 2: Comparison of debiasing performance on BERT. We test the debiased models on SEAT, CrowS-Pairs, and filtered StereoSet-Intrasentence, with the best and second best results in **bold** and underline respectively.

Datasets	SEAT (0.00) the best							StereoSet:gender			StereoSet:all			CrowS-Pairs
Method	6	6-b	7	7-b	8	8-b	AVG (abs)↓	LMS↑	SS (50.00)	ICAT↑	LMS↑	SS(50.00)	ICAT↑	SS(50.00)
RoBERTa	0.92	0.21	0.98	1.46	0.81	1.26	0.94	89.79	66.17	60.74	88.91	62.22	67.17	60.15
DPCE	0.40	0.11	0.73	0.98	0.03	0.75	0.50	82.93	61.80	64.11	61.30	55.14	54.99	54.79
ADEPT-F	1.23	−0.14	0.99	1.09	0.93	1.11	0.92	89.81	63.10	66.27	90.03	61.31	69.68	55.56
CCD-CDA	0.29	−0.07	0.87	0.94	0.58	0.85	0.60	88.52	60.29	70.29	88.88	59.12	72.66	50.57
CCD-GPT	0.40	0.08	0.41	0.85	0.57	0.63	0.49	87.21	59.51	**70.63**	88.33	57.61	**74.89**	48.66
CCD-Gemini	0.27	0.18	−0.13	0.82	0.08	0.81	**0.38**	81.35	**58.15**	68.10	84.68	**56.65**	73.41	**49.54**

Table 3: Comparison of debiasing performance on RoBERTa. We test the debiased models on SEAT, CrowS-Pairs, and filtered StereoSet-Intrasentence, with the best and second best results in **bold** and underline respectively.

Datasets	GLUE ↑					Bias-IR (Male Ratio, 0.50 the best)							
Method	SST-2↑	MRPC↑	RTE↑	WNLI↑	AVG↑	Appearance	Child	Cognitive	Domestic	Career	Physical	Relationship	AVG-DEV↓
BERT	92.9	84.6	72.5	38.0	72.0	0.71	0.50	0.75	0.46	0.75	0.68	0.61	0.16
DPCE	92.8	69.6	53.4	49.3	66.3	0.86	0.79	1.00	0.47	0.70	0.84	0.61	0.24
ADEPT-F	93.2	85.5	69.9	56.3	76.2	0.50	0.50	0.75	0.53	0.80	0.68	0.65	0.13
DPCE-Gemini	93.2	81.4	60.6	46.5	70.4	0.29	0.36	0.17	0.20	0.10	0.32	0.35	0.24
ADEPT-F-Gemini	92.7	81.4	71.5	56.3	75.5	0.71	0.43	0.83	0.53	0.65	0.74	0.65	0.17
CCD-CDA	92.8	**86.3**	65.3	50.7	73.8	0.79	0.79	0.83	0.80	0.70	0.79	0.83	0.29
CCD-GPT	**93.6**	85.1	70.4	56.3	76.4	0.78	0.78	0.50	0.73	0.50	0.63	0.52	0.13
CCD-Gemini	93.5	83.6	**72.9**	**56.3**	**76.6**	0.57	0.64	0.58	0.60	0.70	0.42	0.65	**0.11**

Table 4: Evaluation results on the GLUE dataset and the Bias-IR dataset with BERT, we calculate the average deviation to 0.5 for Bias-IR as AVG-DEV. The **bold** and underline represent the best and second-best respectively.

Datasets	GLUE ↑					Bias-IR (Male Ratio, 0.50 the best)							
Method	SST-2↑	MRPC↑	RTE↑	WNLI↑	AVG↑	Appearance	Child	Cognitive	Domestic	Career	Physical	Relationship	AVG-DEV↓
RoBERTa	93.8	88.2	70.8	56.3	76.9	0.28	0.28	0.66	0.40	0.60	0.42	0.70	0.16
DPCE	78.1	81.6	53.8	56.3	67.5	0.43	0.93	0.42	0.60	0.50	0.58	0.43	0.12
ADEPT-F	93.9	**89.2**	66.8	56.3	76.6	0.57	0.50	0.83	0.60	0.85	0.68	0.74	0.18
CCD-CDA	94.3	88.2	68.2	56.3	76.7	0.29	0.50	0.58	0.13	0.35	0.21	0.56	0.16
CCD-GPT	93.1	86.5	71.5	56.3	76.9	0.43	0.36	0.58	0.33	0.55	0.53	0.61	**0.09**
CCD-Gemini	**94.6**	86.5	**72.9**	56.3	**77.6**	0.43	0.50	0.67	0.53	0.65	0.58	0.69	0.10

Table 5: Evaluation results on the GLUE dataset and the Bias-IR dataset with RoBERTa, we calculate the average deviation to 0.5 for Bias-IR as AVG-DEV. The **bold** and underline represent the best and second-best respectively.

Method	CCED ↓
BERT	0.339
DPCE	0.212
ADEPT-F	0.324
CCD-CDA	0.081
CCD-GPT	**0.056**
CCD-Gemini	0.077

(a) CCED on BERT.

Method	CCED ↓
RoBERTa	0.438
DPCE	0.177
ADEPT-F	0.159
CCD-CDA	0.166
CCD-GPT	0.143
CCD-Gemini	**0.052**

(b) CCED on RoBERTa.

Table 6: Debiasing performance in terms of CCED.

DPCE being the fairest baseline. The CCED results align well with the results on other benchmarks in Table 2 and Table 3, indicating that CCED serves as an new benchmark for text embedding fairness.

Comparision of Data Strategy: To demonstrate the effectiveness of our proposed data strategy, we conduct comparisons with CDA as shown in Table 2 to Table 5. Integrating our debiasing loss with all data strategies results in improved fairness. However, CDA consistently performs worse than GPT and Gemini on fairness due to its limited sensitive word list. This highlights the superiority of our LLM-based augmentation method in leveraging the rich contextual knowledge of LLMs. For the use of different LLMs, both ChatGPT and Gemini achieve strong performance.

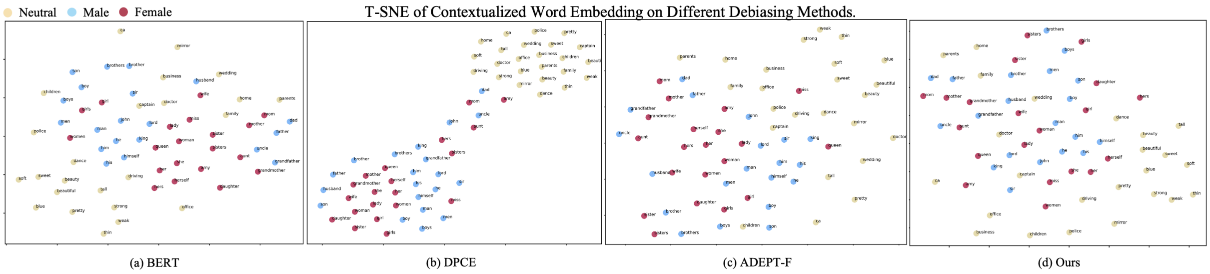

Figure 2: T-SNE plots of embeddings that are processed by different methods. Our approach maintains embedding positions similar to BERT while mixing male and female embeddings thus achieving fairness.

Method	β	LMS	SS	ICAT
CCD-Gemini	0.0	64.37	**51.03**	63.02
	0.5	73.67	53.69	68.22
	1.0	82.91	54.93	**74.72**
	1.5	**84.28**	57.64	71.39

Table 7: Influence of β on StereoSet dataset with BERT.

Baseline with augmented data: In this section, we study of baseline methods with our Gemini augmented data and denote as DPCE-Gemini and ADEPT-F-Gemini . Table 2 shows that our augmented dataset marginally improves fairness in certain metrics, though the overall performance remains similar to that of the original dataset. We arrive at the same conclusion: our CCD surpasses these baseline approaches. Regarding representation capability and BiasIR performance, the results are reported in Table 4. We observed that DPCE experienced an improvement in GLUE average performance, while ADEPT-F showed a slight decline. Despite these variations, both DPCE-Gemini and ADEPT-F-Gemini still exhibit a significant performance gap compared to CCD methods, as detailed in Table 4. To summarize, even with our augmented dataset, our CCD still outperforms baseline methods.

Influence of β: We perform the ablation study of β on CCD-Gemini using the StereoSet dataset on BERT, known for its comprehensive evaluation metrics that assess performance (LMS), fairness (SS), and the trade-off between them (ICAT). We highlight that increasing β amplifies the impact of the L_{rep}, as detailed in Eq. 4, ensuring that neutral embeddings remain unchanged. This provides two key benefits: preserving the model's representational capability and maintaining neutral embeddings as a consistent reference point in the debiasing loss. We vary β from 0 to 1.5, with the results presented in the Table 7. As β increased, we observed an increase in the LMS score from 64.37 to 84.28, indicating improved model utility. However, the fairness score decreased from 57.64 to 51.03, suggesting a shift towards prioritizing utility over fairness. Setting $\beta = 1$ resulted in the optimal ICAT score, balancing fairness and utility.

Embedding Visualization: (1) Fairness Improvement: Fig. 2.a shows the T-SNE of the original BERT model, where male (blue dots) and female (red dots) embeddings form distinct clusters, indicating fairness issues (Peltonen et al., 2023). In contrast, baseline methods and our CCD mix male and female embeddings, thus improving fairness. (2) Utility Preservation: DPCE (Fig. 2.b) separates gendered (blue and red) and neutral (yellow) embeddings, completely removing sensitive information. This disrupts the original embedding geometry and significantly reduces performance (Tables 2 and 4). ADEPT (Fig. 2.c) also causes a performance drop and worsens fairness, as shown in Tables 2 and 4. Notably, our approach (Fig. 2.d) maintains an embedding geometry similar to BERT while mixing male and female embeddings, achieving fairness without compromising utility.

5 Conclusion

In conclusion, we introduce CCED fairness for text embeddings, ensuring conditional independence and equal sensitive information between attributes and embeddings. We propose the CCD loss to achieve this fairness by ensuring that texts with varied sensitive attributes but identical content have equidistant embeddings from their neutral counterparts. By employing LLMs to fairly augment datasets, we achieve effective training with CCD. We establish CCED fairness as a benchmark for evaluating text embeddings fairness. Extensive evaluations on debiasing benchmarks and downstream tasks demonstrate CCD's effectiveness in promoting fairness while preserving utility.

6 Limitaions

In this study, we utilize gender bias to demonstrate the efficacy of our method. As our approach constitutes a general pipeline, we plan to extend our methodology to address other types of biases (e.g., race, age) in the future. Moreover, we discuss the application of our method in a binary gender setting, which generally does not reflect the real world where gender (and other biases) may not be strictly binary. Fortunately, our method is readily extensible to any number of dimensions. We consider this extension as part of our future work.

7 Ethical Consideration

Our work pioneers in mitigating biases in text embeddings, crucial for fairness and inclusivity in NLP applications. We introduce a method that ensures fair representation by achieving conditional independence between sensitive attributes and text embeddings, aiming to reduce societal biases. Employing LLMs for data augmentation represents ethical advancement in tackling inherent biases, moving towards equitable technology and inspiring future bias-aware research. Our contribution significantly advances AI fairness by validating a method that minimizes bias in text embeddings, promoting inclusivity in machine learning.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

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

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *arXiv preprint arXiv:2203.09101*.

Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. 2023. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, pages 158–169. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Renjun Hu, Charu C Aggarwal, Shuai Ma, and Jinpeng Huai. 2016. An embedding approach to anomaly detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 385–396. IEEE.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Jeen Mary John, Olamilekan Shobayo, and Bayode Ogunleye. 2023. An exploration of clustering algorithms for customer segmentation in the uk retail market. *Analytics*, 2(4):809–823.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 444–448.

Thibaud Leteno, Antoine Gourru, Charlotte Laclau, Rémi Emonet, and Christophe Gravier. 2023. Fair text classification with wasserstein independence. *arXiv preprint arXiv:2311.12689*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

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*.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Ben Packer, Yoni Halpern, Mario Guajardo-Cspedes, and Margaret Mitchell. 2018. Text embedding models contain bias. here's why that matters. *Google Developers*.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Jaakko Peltonen, Wen Xu, Timo Nummenmaa, and Jyrki Nummenmaa. 2023. Fair neighbor embedding. In *International Conference on Machine Learning*, pages 27564–27584. PMLR.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J Sutherland, Victor Veitch, and Arthur Gretton. 2022. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. *arXiv preprint arXiv:2302.00618*.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *CoRR*, abs/2109.10645.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Rui Wang, Pengyu Cheng, and Ricardo Henao. 2023. Toward fairness in text generation via mutual information minimization based on importance sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 4473–4485. PMLR.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362.

Han Zhao and Geoffrey J Gordon. 2022. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

A Algorithm Details

A.1 Notation

		Basic Variables
L	$\triangleq$	loss function
f,f^{ori}	$\triangleq$	finetuned and original text embedding model.
h	$\triangleq$	Large language model.
θ_p	$\triangleq$	Few-shot prompts that used to empower a LLM.
A, a_i	$\triangleq$	Sensitive attribute set and i-th sensitive attribute.
S^{a_i},S^n	$\triangleq$	Text that relate to sensitive attribute a_i and neutral text.
C,C'	$\triangleq$	Content variable and predicted content.
X^{a_i}, X^n	$\triangleq$	words from group a_i and neutral words in a text.
V^{a_i}	$\triangleq$	words list that contains all collected words related to attribute a_i.

Table 8: Main notations used in this paper.

A.2 The significance of text embedding fairness and its distinction from subsequent task fairness

Recently (Shen et al., 2021, 2022) apply contrastive learning losses to mitigate biases in language representations for text classification and (Leteno et al., 2023; Shen et al., 2022) find a representational fairness and subsequent task group fairness are not, or only partially, correlated. However, subsequent tasks and text embedding fairness represent two distinct areas that are both important and need to be distinguish:

The importance of embedding fairness: Recent efforts, as highlighted in the introduction of our paper, emphasize the significance of text embedding fairness. The fairness of embeddings is essential due to their widespread application across various systems. For instance, Search Engine (Huang et al., 2020), preprocess all content—including documents, videos, and audios—into embeddings to save on storage. When a search query is submitted, it is converted into an embedding to retrieve the most relevant results, especially during the recall phase, where embedding similarity is used to filter through numerous documents to find pertinent ones. Moreover, embeddings are directly used in other applications such as zero-shot classification (Yin et al., 2019; Radford et al., 2021), clustering (John et al., 2023), and Anomaly Detection (Hu et al., 2016), among others. Given the critical role that embeddings play in these and additional applications, addressing fairness issues within the embeddings themselves is undeniably crucial.

Difference between embedding fairness and subsequent task group fairness: This paper focuses on the intrinsic fairness of text embeddings. However, the group fairness of subsequent tasks extends beyond this, incorporating additional modules that take embeddings as input for predictions, which are influenced by other sources of bias. For instance, in a medical report dataset where only females are depicted as having a cold, even if the embedding captures information about gender equally (as defined in Definition 3.1), subsequent modules in the system might still incorrectly associate women with having colds. As a result, it is important to distinguish the difference between the fairness of subsequent tasks and the intrinsic fairness of embeddings.

What we explored and can explore in the future: In this paper, we focus on text embedding fairness and studied its influence on information retrieval tasks, as shown in Table 4 and Table 5 in our paper. Creating fair text embeddings directly improves the fairness of information retrieval. While group fairness of subsequent tasks falls outside the scope of this paper, exploring the relationship between embedding fairness and group fairness in future work could be valuable. This exploration would involve selecting an appropriate metric (Mehrabi et al., 2021) for representation fairness and disentangle the fairness of subsequent task modules and embedding intrinsic fairness.

Considering the widespread use of embeddings, differences between group fairness and embedding fairness, we believe the fairness of text embeddings is indeed an important research topic in itself.

A.3 Dataset Details

We generated training data using the News-Commentary-v15 corpus (Tiedemann, 2012) focusing on gender bias. By employing Gemini and ChatGPT for data augmentation, we obtained datasets comprising 43,221 and 42,930 sample pairs, respectively. Each pair contains texts with identical content from male, female, and neutral perspectives. We use last 1000 data as validation set and the remaining data as training set.

For the bias evaluation dataset, we provide detailed statistics in Table 9. Our augmented dataset sets a new benchmark, featuring an extensive dataset size that enhances the robustness and comprehensiveness of bias assessment.

Evaluation Data	**Level**	**Data Size**
Sentence Encoder Association Test (SEAT)	Text	5172
CrowS-Pairs	Text	1508
StereoType Analysis	Text	8497
Gender-Bias-IR	Query-Doc	236
CCD-GPT (ours)	Text	42,930
CCD-Gemini (ours)	Text	43,221

Table 9: Dataset Statistics on various bias evaluation benchmarks.

A.4 Data Augmentation Prompts

The prompt template can be found in Figure 1. To provide a clearer demonstration, we also list the examples we used. Notably, to save computational costs, we have shortened the examples and merged the selected 10 examples into 8, as shown in the Table 10.

A.5 Ommited Proofs

In this section, we give a detailed proof of Theorem 3.3.

Proof. Firstly, we establish the conditional independence $A \perp C' \mid C$ for any $a_i, a_j \in A$:

$$P(C' \mid A = a_i, C) = P(C' \mid A = a_j, C) \tag{8}$$

where C' represents the content embedding. Assuming equal probabilities for different sensitive attributes $P(a_1 \mid C) = \cdots = P(a_A \mid C)$, we can rewrite Eq. (8) as:

$$\begin{aligned} P(C' \mid A = a_i, C)P(a_i \mid C) &= P(C' \mid A = a_j, C)P(a_j \mid C) \\ P(C', a_i \mid C) &= P(C', a_j \mid C) \end{aligned} \tag{9}$$

According to Section 3.1, $f(S_C^{a_i})$ encodes both content and sensitive information, allowing us to obtain:

$$P(f(S_C^{a_i}) \mid C) = P(f(S_C^{a_j}) \mid C) \tag{10}$$

Because a fair and well-trained embedding model f can effectively extract the content C from the neutral text S_C^n without introducing bias, we can approximate Eq. (10) as:

$$P(f(S_C^{a_i}) \mid f(S_C^n)) = P(f(S_C^{a_j}) \mid f(S_C^n)) \tag{11}$$

Following (Hinton and Roweis, 2002; Yang et al., 2023), the conditional probability $P(f(S_C^{a_i}) \mid f(S_C^n))$ can be represented as the similarity between $S_C^{a_i}$ and $f(S_C^n)$, and can be modeled using a Gaussian distribution. We thus measuring $P(f(S_C^{a_i}) \mid f(S_C^n))$ by calculating:

$$P(f(S_C^{a_i}) \mid f(S_C^n)) = \frac{\exp\left(-\frac{\|f(S_C^{a_i}) - f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_i \in A} \exp\left(-\frac{\|f(S_C^{a_i}) - f(S_C^n)\|^2}{2\rho^2}\right)} \tag{12}$$

where ρ controls falloff of the P with respect to distance and is set by hand. Eq. (12) can be interpreted as follows: (1) Consider setting a Gaussian distribution with a covariance matrix equal to ρ times the identity matrix at the embedding of a neutral text S_C (with content C), which is denoted as $f(S_C^n)$. Then, a text with the same content but containing sensitive information a_i appears in the distribution with a probability proportional to $\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)$, represented as the numerator. (2) The denominator aggregates the aforementioned probabilities across all sensitive groups $a_i \in A$ and serves as the normalization factor. Then we combine Eq. (11) and Eq. (12) and obtain:

$$
\begin{aligned}
\frac{\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_i\in A}\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)} &= \frac{\exp\left(-\frac{\|f(S_C^{a_j})-f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_j\in A}\exp\left(-\frac{\|f(S_C^{a_j})-f(S_C^n)\|^2}{2\rho^2}\right)} \\
\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right) &= \exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right) \\
\|f(S_C^{a_i})-f(S_C^n)\|^2 &= \|f(S_C^{a_j})-f(S_C^n)\|^2 \qquad (13)
\end{aligned}
$$

Thus we obtain the Theorem 3.3. As a result, achieving conditional independence between sensitive attributes and content embeddings is equivalent to achieving content-conditioned equal distance. □

Example	Original passage	Neutral passage	Male passage	Female passage
Example 1	The high popularity of the current president (Socialist Michelle Bachelet, Chile's first female chief executive)	The high popularity of the current president (A Socialist, Chile's first chief executive)	The high popularity of the current president (Socialist Mike Bachelet, Chile's first male chief executive)	The current president (Socialist Michelle Bachelet, Chile's first female chief executive)
Example 2	Rwanda has the highest female legislators in the world.	Rwanda has the highest legislators in the world.	Rwanda has the highest male legislators in the world.	Rwanda has the highest female legislators in the world.
Example 3	When a kid arrived, accompanied by a doting father, the prophet's son.	When a kid arrived, accompanied by a doting parent, the prophet's child.	When a kid arrived, accompanied by a doting father, the prophet's son.	When a kid arrived, accompanied by a doting mother, the prophet's daughter.
Example 4	wizards Hunt people, poor paternal nutrition.	People Hunt people, poor nutrition.	wizards Hunt people, poor paternal nutrition.	Witch Hunt people, poor maternal nutrition.
Example 5	Bruni's life path become opera divo, barman and actress.	A people's life path become opera performer, bar staff and acting.	Michael's life path become opera diva, barwoman and actor.	Bruni's life path become opera divo, barman and actress.
Example 6	Ally is marchioness, Bride for Sarkozy.	they are noble, partner of someone.	Alexandria is marquis, Groom for Sara.	Ally is marchioness, Bride for Sarkozy.
Example 7	Mike embarked on a fascinating experiment with sons.	Leader embarked on a fascinating experiment with offsprings.	Mike embarked on a fascinating experiment with sons.	Merkel embarked on a fascinating experiment with daughters.
Example 8	Orban and Tomy appointed a police as his secretary, most strong-minded male Democrat.	They appointed a police as their secretary, most strong-minded Democrat.	Orban and Tomy appointed a police as his secretary, most strong-minded male Democrat.	Olivia and Michelle appointed a police as her secretary, most strong-minded female Democrat.

Table 10: Task template and prompt examples for gender-neutral, male, and female passages.