The Elephant in the Room: Exploring the Role of Neutral Words in Language Model Group-Agnostic Debiasing

Xinwei Guo[1] **Jiashi Gao**[1] **Junlei Zhou**[1]
Jiaxin Zhang[1] **Guanhua Chen**[1] **Xiangyu Zhao**[2] **Quanying Liu**[1]
Haiyan Wu[3] **Xin Yao**[4] **Xuetao Wei**[1†]
[1]Southern University of Science and Technology, [2]City University of Hong Kong
[3]University of Macau, [4]Lingnan University
guoxw2023@mail.sustech.edu.cn, weixt@sustech.edu.cn

Abstract

Large language models (LLMs) are increasingly integrated into our daily lives, raising significant ethical concerns, especially about perpetuating stereotypes. While group-specific debiasing methods have made progress, they often fail to address multiple biases simultaneously. In contrast, group-agnostic debiasing has the potential to mitigate a variety of biases at once, but remains underexplored. In this work, we investigate the role of neutral words—the group-agnostic component—in enhancing the group-agnostic debiasing process. We first reveal that neutral words are essential for preserving semantic modeling, and we propose ϵ-DPCE, a method that incorporates a neutral word semantics-based loss function to effectively alleviate the deterioration of the Language Modeling Score (LMS) during the debiasing process. Furthermore, by introducing the SCM-Projection method, we demonstrate that SCM-based debiasing eliminates stereotypes by indirectly disrupting the association between attribute and neutral words in the Stereotype Content Model (SCM) space. Our experiments show that neutral words, which often embed multi-group stereotypical objects, play a key role in contributing to the group-agnostic nature of SCM-based debiasing.

1 Introduction

From BERT to GPT and DeepSeek (Devlin et al., 2019; Achiam et al., 2023; Liu et al., 2024), large language models (LLMs) have become increasingly intelligent and cost-effective, seamlessly integrating into daily life. However, their widespread use has also intensified ethical concerns, particularly regarding interactions between LLMs and users. As LLMs expand the application scenarios of language models, the risk of amplifying and propagating stereotypes by their responses grows. To address this issue, numerous studies have focused on bias and stereotypes in language models (Gallegos et al., 2024; Doan et al., 2024).

Stereotypes can be regarded as over-generalized beliefs about a particular group of people, e.g., Asian Americans are good at math, or African Americans are athletic (Nadeem et al., 2021). Negative stereotypes can lead to biased attitudes towards specific groups, resulting in unfair treatment of individuals within those groups. Since text is a significant medium for expressing human stereotypes, LLMs pre-trained on extensive corpora—lacking thorough processing—may also learn and perpetuate these negative stereotypes. For language models, stereotypes exist in the form of spurious correlations, such as associations between "man" and "programmer", "woman" and "homemaker", due to their high co-occurrence (Bolukbasi et al., 2016). In existing word embedding debiasing practices (Kaneko and Bollegala, 2021; Yang et al., 2023; Omrani et al., 2023), stereotypes are embodied as the association between attribute words and neutral words (More details in Section 2.1).

Most current research on debiasing language models focuses on group-specific methods, which require the use of attribute words to define specific groups. However, identifying appropriate attribute words for all demographic groups is both cumbersome and challenging, making comprehensive debiasing difficult. While various group-specific debiasing techniques have been extensively studied—including Counterfactual Data Augmentation (CDA; Zmigrod et al. 2019), Dropout (Webster et al., 2020), Self-Debias (Schick et al., 2021)), and word embedding debiasing methods (Kaneko and Bollegala, 2021)—experiments by Meade et al. (2022) demonstrated that **these methods can only mitigate a single type of bias, such as gender bias, and are ineffective at addressing multiple biases simultaneously, such as those related to both gender and race**.

[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 20360–20371
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

The limitations of group-specific debiasing methods have spurred interest in exploring group-agnostic approaches. For instance, building upon the group-specific DPCE (Kaneko and Bollegala, 2021) framework, Omrani et al. (2023) proposed SCM-based debiasing, which eliminates group-agnostic stereotypes by incorporating the Stereotype Content Model (SCM) (Fiske et al., 2002). Compared with DPCE, which removes stereotypes by orthogonalizing the word embeddings of attribute words and neutral words, SCM-based debiasing achieves group-agnostic stereotype elimination by severing the connection between SCM-related and neutral words. Despite their consistent presence from group-specific DPCE to group-agnostic SCM-based debiasing, neutral words have yet to be thoroughly explored. **What role do neutral words play in these debiasing processes? If they do influence these processes, how exactly do they affect debiasing, particularly for SCM-based debiasing?**

To address the questions raised above, this work investigates the role of neutral words in group-agnostic debiasing of language models. We find that neutral words are indispensable and play a crucial role in both preserving potential semantic modeling and enhancing the debiasing performance to be group-agnostic. For the first role, we observe that DPCE (Kaneko and Bollegala, 2021) leads to a deterioration in the Language Modeling Score (LMS), a metric from StereoSet (Nadeem et al., 2021) that evaluates the overall language modeling capability of a language model. We attribute this issue to DPCE's failure to preserve potential semantic modeling beyond attribute words. Notably, we find that incorporating a loss function related to neutral word semantics can significantly mitigate the LMS deterioration problem.

For the second role, to better understand how neutral words function in SCM-based debiasing to eliminate group-agnostic stereotypes, we propose the Neutral-Attribute-SCM (NAS) framework and the SCM-Projection method. These tools help clarify the underlying mechanisms of SCM-based debiasing. Our findings confirm that SCM-based debiasing *indirectly* disrupts the association between attribute words and neutral words by repositioning neutral words within the SCM space. Finally, through experiments based on group-specific selection, we reveal that neutral words often include stereotypical objects associated with multiple demographic groups and contributes to the group-agnostic nature of SCM-based debiasing.

Contribution The contributions of this work can be summarized as follows:

1. To the best of our knowledge, we are the first to identify and analyze the role of neutral words in the group-agnostic bias-mitigating process of language models. The neglect of neutral words has led to performance limitation in existing word embedding debiasing method, particularly in terms of preserving semantic modeling and enhancing the effectiveness of debiasing mechanisms. Our work provides crucial insights into the foundation for further optimizing debiasing techniques.

2. Building on the analyzed functionality of neutral words in preserving semantic modeling, we are able to interpretably propose the ϵ-DPCE, which significantly alleviates the deterioration in Language Modeling Score (LMS) that typically arises during the debiasing process in DPCE.

3. We take the first step toward analyzing and explaining the underlying mechanisms of SCM-based debiasing methods. By proposing the NAS framework and SCM-Projection method, we provide a new perspective on the mechanisms behind various debiasing strategies and find that multi-group stereotypical objects embedded in neutral words contribute to the group-agnostic nature of SCM-based debiasing.

2 Preliminaries

2.1 Word Embedding Debiasing

Stereotypes have been observed in both statistic and contextualized word embeddings (Bolukbasi et al., 2016; Kurita et al., 2019). Contextualized word embeddings, in particular, can more accurately reflect word meanings within specific contexts and have become prevalent in mainstream LLMs. Next, we will introduce the classic contextualized word embedding debiasing method known as DPCE (Kaneko and Bollegala, 2021) and demonstrate how to mitigate stereotypes associated with attribute words and neutral words.

Attribute Words Collecting attributes that represent demographic groups affected by stereotypes is crucial for DPCE. These attribute words can be adjectives, nouns, or other forms. For instance, researchers commonly use terms like “she”, “woman” and “her" to refer to the female group, while “he”, “man”, and “his" are used for the male group. In this paper, we define attribute words as $\mathcal{V}_{\text{attr}}$.

Neutral Words Neutral words, also known as target words, serve as the stereotypical object for a particular demographic group. We define neutral words as $\mathcal{V}_{\text{ntr}}$. Similar to attribute words, neutral words can be adjectives, nouns, or other forms. For example, various occupations can serve as gender-neutral words in the context of gender bias, including words like “doctor”, “nurse”, “programmer” and “dancer”.

DPCE Kaneko and Bollegala (2021) proposed Debiasing Pre-trained Contextualised Embeddings (DPCE), which eliminates the stereotype in language models by orthogonalizing the word embeddings of attribute words $\mathcal{V}_{\text{attr}}$ and neutral words $\mathcal{V}_{\text{ntr}}$. We retain some of the symbols used in (Kaneko and Bollegala, 2021). To obtain contextualized word embeddings, we first need to collect sentences that contain the word w and denote the set of sentences as $\Omega(w)$. Here, we define the set of sentences containing attribute words as $\mathcal{A} = \bigcup_{w \in \mathcal{V}_{\text{attr}}} \Omega(w)$ and the set of sentences containing neutral words as $\mathcal{N} = \bigcup_{w \in \mathcal{V}_{\text{ntr}}} \Omega(w)$. Then we input the sentence x, which contains the word w, into the pre-trained language model E with parameters θ to obtain the contextualized word embedding $E(w, x; \theta)$.

Kaneko and Bollegala (2021) required that the debiased word embedding $E(t, x; \theta)$ of a target word $t \in \mathcal{V}_{\text{ntr}}$ to be unrelated to a protected attribute a and they formalized this requirement as:

$$\mathcal{L}_{\text{bias}} = \sum_{t \in \mathcal{V}_{\text{ntr}}} \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_{\text{attr}}} \left(v(a)^{\top} E(t, x; \theta) \right)^2, \quad (1)$$

where $v(a)$ refer to the average contextualizd word embedding of the attribute a and $v(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E(a, x; \theta)$. At the same time, to maintain the language modeling ability of the model E, Kaneko and Bollegala (2021) proposed a regulariser aimed at minimizing changes to the hidden states associated with attribute words $\mathcal{V}_{\text{attr}}$ and formalized it as:

$$\mathcal{L}_{\text{reg}} = \sum_{x \in \mathcal{A}} \sum_{w \in x} \left|\left| E(w, x; \hat{\theta}) - E(w, x; \theta) \right|\right|^2 \quad (2)$$

In last, the overall training objective of DPCE can be expressed as the linear weighted sum of L_{bias} and L_{reg} as $\mathcal{L} = \alpha \mathcal{L}_{\text{bias}} + \beta \mathcal{L}_{\text{reg}}$. Coefficients $\alpha, \beta \in [0, 1]$ and $\alpha + \beta = 1$.

2.2 Stereotype Content Model

Fiske et al. (2002) proposed the stereotype content model (SCM) to characterize and differentiate the stereotypes associated with various groups in society. Specifically, SCM is represented as a two-dimensional space, with warmth and competence as its axes. To address the lack of comprehensive tools for text analysis related to the SCM, Nicolas et al. (2021) developed stereotype content dictionaries. These dictionaries comprise 28 different dictionaries containing a total of 14,449 words. The words included in these stereotype content dictionaries can be utilized to evaluate the warmth and competence dimensions of the SCM. More details in Appendix B.

2.3 Benchmark of Stereotypes: StereoSet

Researchers have progressively established stereotype assessment methods based on embedding association tests, ranging from the *word embedding association test (WEAT)* (Caliskan et al., 2017) to the *sentence embedding association test (SEAT)* (May et al., 2019) and the *contextualized embedding association test (CEAT)* (Guo and Caliskan, 2021). However, it is insufficient only to consider the effects of stereotype elimination when evaluating debiasing methods, as most of these methods inevitably alter the model parameters, potentially influencing the existing language modeling. Nadeem et al. (2021) developed StereoSet, a dataset that includes stereotypes related to occupation, gender, race, and religion, considering both language modeling abilities and stereotype evaluation. StereoSet includes three evaluation metrics: Language Model Score (LMS), Stereotype Score (SS), and Idealized CAT Score (ICAT).

Language Modeling Score (LMS) An ideal language model should achieve a Language Model Score (LMS) of 100, indicating that the model consistently favours the more meaningful associations for each target term in the dataset.

Stereotype Score (SS) The Stereotype Score (SS) for an ideal language model is 50, indicating that for each target term, the model shows no preference for either stereotypical or anti-stereotypical associations.

Idealized CAT Score (ICAT) ICAT is a comprehensive evaluation of the language model, taking into account both its semantic modeling ability and the degree of stereotypes. It can be calculated by $\text{ICAT} = \text{LMS} \times \frac{\min(\text{SS}, 100-\text{SS})}{50}$.

In comparison to evaluating language models using a variety of tasks, like GLUE (Wang et al., 2018), StereoSet offers a more integrated approach by combining stereotype testing with language modeling tests within the same case. This integration allows for a more accurate and direct reflection of how eliminating stereotypes affects the language model performance.

3 Role 1: Neutral Words Protect Potential Semantic Modeling in Debiasing

3.1 Problem Discussion

Eliminating stereotypes within the word embedding space constitutes a critical component of language model debiasing, and DPCE (Kaneko and Bollegala, 2021) is a prevalent method for tackling this challenge. The process of acquiring the optimal debiased model using DPCE can be summarized as follows. Initially, the dataset is partitioned into a training set and a development (dev) set. The language model is then trained on the training set and its performance is evaluated on the dev set. Ultimately, the model exhibiting the minimal loss on the dev set is selected as the optimal debiased model. **However, there is a gap between the expected performance and the actual performance of the optimal debiased model obtained by the above process.**

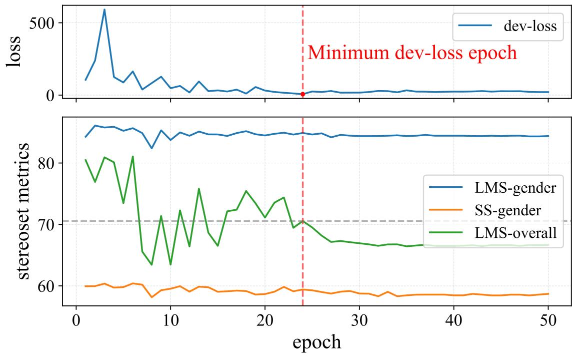


Figure 1: The curves of dev-loss, LMS-gender, SS-gender, and LMS-overall as the number of DPCE training epochs increases. The LMS-overall has dropped significantly.

To explain the performance gap more intuitively and clearly, we conducted gender-related stereotype elimination experiments on the `bert-large-uncased` (Devlin et al., 2019). Attribute words and neutral words used in the experiments are available in Appendix A. Besides, the debiased model was evaluated on StereoSet after each epoch to gather performance data.

In Figure 1, we plotted the curves of dev-loss, LMS-gender, SS-gender and LMS-overall during the training of the `bert-large-uncased`. LMS-gender represents the LMS performance on gender-related stereotype test cases, while LMS-overall reflects the LMS result after testing all test cases. As the number of training epochs increases, the dev-loss tends to decrease; however, the performance of the debiased model on StereoSet does not improve as expected. Although SS, which represents the degree of stereotype, continues to decrease, the performance of LMS-overall also deteriorates. It should be noted that **LMS-overall decreased seriously, but LMS-gender has not been affected significantly, which may mislead previous researchers.**

A low LMS-overall suggests that the language modeling ability of the `bert-large-uncased` has been compromised. The specific manifestation of this degradation in StereoSet is that when faced the fill-in-the-blank question like "Girls tend to be more __ than boys", `bert-large-uncased` might incorrectly select the nonsensical option "fish" instead of more suitable choices like "soft" or "determined". If the optimal debiased model is selected according to the minimum dev-loss, the actual LMS-overall of the debiased `bert-large-uncased` drops from **over 80** to **below 70**, as shown in Figure 1. This demonstrates that there is a significant gap between the expected performance and the actual performance of the optimal debiased model.

The reason for the above problems is that the $\mathcal{L}_{\text{reg}}$ in Formula 2 focuses solely on protecting the semantic modeling related to attribute words, neglecting the protection of other potential semantic modelings. Specifically, since the goal of $\mathcal{L}_{\text{bias}}$ is to make the hidden states corresponding to the attribute word and the neutral word orthogonal and thus unrelated, the hidden state vectors for these two groups of words will inevitably change after backpropagation adjusts the model parameters. However, the purpose of $\mathcal{L}_{\text{reg}}$ is to ensure that the hidden states associated with attribute words change as little as possible before and after training. Therefore, in the process of minimizing $\mathcal{L}_{\text{bias}}$, the hidden states associated with neutral words experience a more substantial alteration, which leads to the destruction of potential semantic modelings related to neutral words and a sharp drop in the LMS-overall performance on StereoSet.

3.2 Solution

To bridge the gap between the expected and the actual performance of the debiased model, we propose the ϵ-DPCE, which protects the potential semantic modeling through the use of the loss function $\mathcal{L}_{\text{ntr}}$.

ϵ-DPCE In our experiments, we found that the potential semantic modeling is closely related to the neutral words. Therefore, we first enhanced the DPCE by incorporating the loss function $\mathcal{L}_{\text{ntr}}$, defined as:

$$\mathcal{L}_{\text{ntr}} = \sum_{x \in \mathcal{N}} \sum_{w \in x} \left|\left| E(w, x; \hat{\theta}) - E(w, x; \theta) \right|\right|^2. \tag{3}$$

However, conflicts among $\mathcal{L}_{\text{bias}}$, $\mathcal{L}_{\text{ntr}}$, and $\mathcal{L}_{\text{reg}}$ (Section 3.1) complicate the training of the debiased model. To address these conflicts, we employ the ϵ-constraint method within the DPCE framework, which is a well-established approach for solving multi-objective optimization problems. Specifically, we constraint the $\mathcal{L}_{\text{ntr}}$ of model with parameter θ less than ϵ. Specifically, we constrain the $\mathcal{L}_{\text{ntr}}$ value of the model with parameter θ to be less than ϵ, and formulated it as:

$$\begin{aligned} &\min \mathcal{L}(\theta) \\ &\text{s.t. } \mathcal{L}_{\text{ntr}}(\theta) \leq \epsilon, \end{aligned} \tag{4}$$

where $\mathcal{L}(\theta) = \alpha\mathcal{L}_{\text{bias}}(\theta) + \beta\mathcal{L}_{\text{reg}}(\theta)$. Similar to experiments described in Section 3.1, we utilize the ϵ-DPCE to debias the `bert-large-uncased` and depict the results in Figure 2.

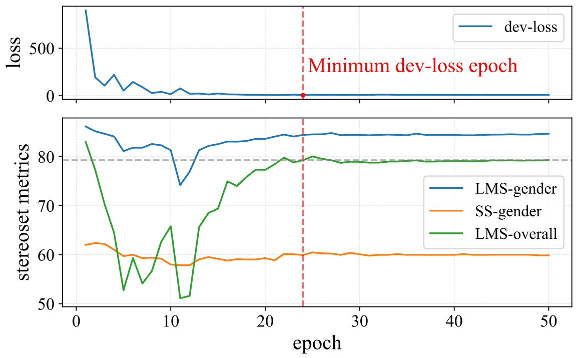


Figure 2: The curves of dev-loss, LMS-gender, SS-gender, and LMS-overall as the number of ϵ-DPCE training epochs increases.

As can be seen in Figure 2, as the dev-loss decreases, the performance of the debiased model on StereoSet tends to be stable. In particular, the LMS-overall value of the debiased model with minimum dev-loss is close to 80, which is significantly better than the DPCE. To provide a more straightforward comparison between DPCE and ϵ-DPCE, we have summarized the performance metrics on StereoSet in Table 1. According to Table 1, it is acceptable to consider both the last epoch debiased model and the debiased model with minimal development loss as optimal choices for the ϵ-DPCE. In addition to `bert-large-uncased`, we also conducted experiments on gender-related stereotype elimination using `Llama-3.2-1B`, with results provided in Appendix C.

	Min dev loss		The last epoch	
	DPCE	ϵ-DPCE	DPCE	ϵ-DPCE
$\text{LMS}_{\text{gender}} \uparrow$	84.89	84.45	84.39	84.70
$\text{SS}_{\text{gender}} \to 50$	59.41	59.95	58.70	59.87
$\text{ICAT}_{\text{gender}} \uparrow$	68.91	67.65	69.71	67.98
$\text{LMS}_{\text{overall}} \uparrow$	70.56	**79.30**	66.66	**79.30**
$\text{SS}_{\text{overall}} \to 50$	53.38	54.00	52.51	54.06
$\text{ICAT}_{\text{overall}} \uparrow$	65.78	72.96	63.31	72.86

Table 1: Experimental results of debiased models correspond to the last epoch and the epoch with minimal dev-loss in ϵ-DPCE.

4 Role 2: Neutral Words Contribute to the Group-Agnostic Nature of SCM-Based Debiasing

To further investigate why SCM-based debiasing can eliminate group-agnostic stereotype, we divide the question into two sub-questions: ❶ How does SCM-based debiasing eliminate stereotypes? ❷ How does SCM-based debiasing achieve group-agnostic stereotype elimination? Exploring the answers to the above questions is helpful in improving the interpretability of the SCM debiasing method and provides better guidance for its application in stereotype elimination practice.

4.1 How Does SCM-Based Debiasing Eliminate Stereotypes?

As discussed in Section 2.1, stereotypes in language models often manifest through associations between attribute words and neutral words, such as “she-nurse” and “he-doctor”, along with other similar spurious correlations. Therefore, it is straightforward to understand that severing the link between attribute words and neutral words can help eliminate these stereotypes. However, an important question arises: **Why does cutting the association between SCM words and neutral words also contribute to the reduction of stereotypes?**

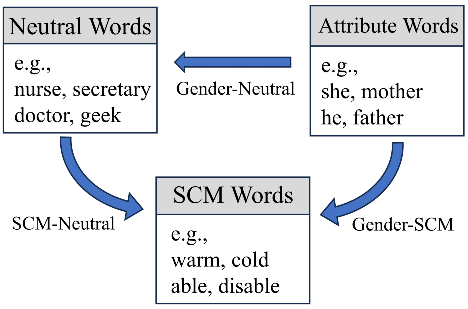


Figure 3: The overview of the Neutral-Attribute-SCM (NAS) framework. There are three debiasing solutions within the NAS framework: Gender-Neutral, SCM-Neutral, and Gender-SCM.

Neutral-Attribute-SCM (NAS) To solve the above puzzles, we design the Neutral-Atrribute-SCM (NAS) framework as shown in Figure 3. Arrows in Figure 3 represent three potential debiasing solutions: Gender-Neutral, SCM-Neutral and Gender-SCM. For convenience, we denote the method that eliminates stereotypes by removing the association between gender attribute words and neutral words as Gender-Neutral. Likewise, we can define SCM-Neutral and Gender-SCM. Inspired by the NAS framework and SCM space, we conjecture that the effectiveness of the SCM-Neutral in eliminating stereotypes arises from ***indirectly* disrupting the association between gender attribute words and neutral words by altering the mapping positions of neutral words in the SCM space.**

We verify the above conjecture in two ways. First, we take gender stereotypes as an example and conducted the SCM-Projection experiments. Taking the original model as a reference, the distribution changes of neutral words in the SCM space of the SCM-Neutral debiased model should be similar to that of Gender-Neutral. Second, the effectiveness of Gender-Neutral and SCM-Neutral in eliminating stereotypes has been verified (Kaneko and Bollegala, 2021; Omrani et al., 2023). However, if the conjecture is true, the Gender-SCM should also be effective in debiasing because it cuts off the association between attribute words and neutral words by altering the mapping positions of gender attribute words in the SCM space.

SCM-Projection We design the SCM-Projection method to map attribute words and neutral words into the SCM space. Given a to-be-mapped word w, a SCM word set $\mathcal{V}_{\text{scm}}$ and the language model E with parameter θ, we define the distance $S(w, \mathcal{V}_{\text{scm}})$ between w and $\mathcal{V}_{\text{scm}}$ as the inner product of their word embedding vectors. Similar to the $\mathcal{L}_{\text{bias}}$, we complete the distance calculation between the w and $\mathcal{V}_{\text{scm}}$ by:

$$S(w, \mathcal{V}_{\text{scm}}) = \sum_{t \in \mathcal{V}_{\text{scm}}} \sum_{x \in \Omega(t)} \left(v(w)^{\top} E(t, x; \theta) \right)^2, \quad (5)$$

where the $E(t, x; \theta)$ refers to the last hidden state. In the SCM space, we represent the four SCM word sets as $\mathcal{V}_{\text{competent}}$, $\mathcal{V}_{\text{incompetent}}$, $\mathcal{V}_{\text{warm}}$ and $\mathcal{V}_{\text{cold}}$, respectively.

Then, we can calculate the distances between the word w and these four SCM word sets as follows: $s_{c_1} = S(w, \mathcal{V}_{\text{competent}})$, $s_{c_2} = S(w, \mathcal{V}_{\text{incompetent}})$, $s_{w_1} = S(w, \mathcal{V}_{\text{warm}})$ and $s_{w_2} = S(w, \mathcal{V}_{\text{cold}})$. Afterwards, we can obtain the uncalibrated position (d_x, d_y) of word w in the SCM space by:

$$d_x = 1 - \frac{2s_{c_2}}{s_{c_1} + s_{c_2}} \quad (6)$$

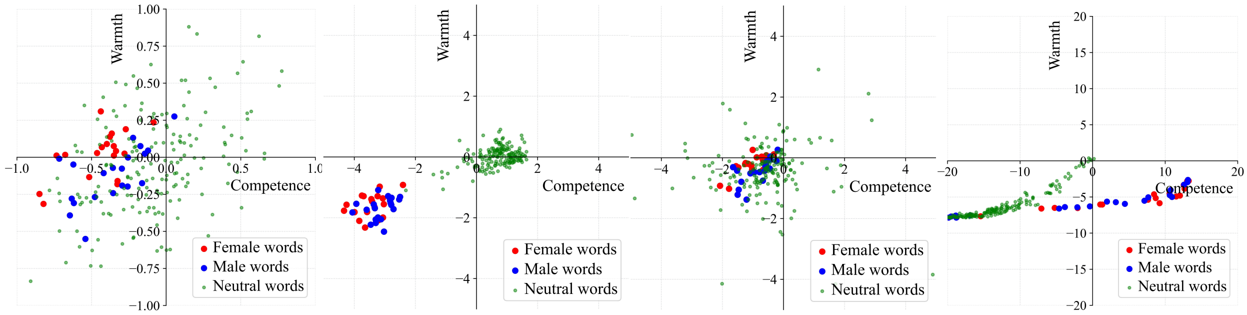
$$d_y = 1 - \frac{2s_{w_2}}{s_{w_1} + s_{w_2}} \quad (7)$$

Calibration However, due to possible deviations in word embeddings, d_x and d_y need to be calibrated to be more accurate. For calibration, we need four anchors X_{pos}, X_{neg}, Y_{pos} and Y_{neg}. Let $X_{\text{pos}} = \frac{1}{|\mathcal{V}_{\text{competent}}|} \sum_{w \in \mathcal{V}_{\text{competent}}} S(w, \mathcal{V}_{\text{competent}})$, $X_{\text{neg}} = \frac{1}{|\mathcal{V}_{\text{incompetent}}|} \sum_{w \in \mathcal{V}_{\text{incompetent}}} S(w, \mathcal{V}_{\text{incompetent}})$, $Y_{\text{pos}} = \frac{1}{|\mathcal{V}_{\text{warm}}|} \sum_{w \in \mathcal{V}_{\text{warm}}} S(w, \mathcal{V}_{\text{warm}})$, and $Y_{\text{neg}} = \frac{1}{|\mathcal{V}_{\text{cold}}|} \sum_{w \in \mathcal{V}_{\text{cold}}} S(w, \mathcal{V}_{\text{cold}})$. With the help of anchors, we can get the calibrated position (p_x, p_y) of word w by:

$$p_x = \frac{2d_x - (X_{\text{pos}} + X_{\text{neg}})}{X_{\text{pos}} - X_{\text{neg}}} \quad (8)$$

$$p_y = \frac{2d_y - (Y_{\text{pos}} + Y_{\text{neg}})}{Y_{\text{pos}} - Y_{\text{neg}}} \quad (9)$$

Based on the SCM-Projection method, we debiased the `bert-large-uncased` with ϵ-DPCE and plotted the mapping distributions in Figure 4. After undergoing Gender-Neutral and SCM-Neutral debiasing, the mapping distribution of female and male words becomes closer to each other and exhibits



(a) Original (b) Gender-Neutral (c) SCM-Neutral (d) Gender-SCM

Figure 4: The projection distributions of attribute words and neutral words in the SCM space. These four subfigures correspond to the original model and three models that have been debiased using three debiasing methods.

greater concentration compared to the original distribution. Therefore, we can summarize the underlying mechanism of the debiasing methods: **overlapping the mapping positions of female words and male words in the SCM space prevents them from forming stereotypical associations due to their individual proximity to neutral words.**

As shown in Figure 4(c), since SCM-Neutral does not involve attribute words, it minimizes individual associations with female or male words by concentrating the projections of neutral words toward the origin. In other words, **SCM-based debiasing (SCM-Neutral) *indirectly* disrupts the association between attribute words and neutral words by repositioning neutral words within the SCM space.** Furthermore, the experimental data of Figure 4 is summarized in Table 2, revealing that Gender-SCM debiased model achieves notable stereotype elimination on StereoSet, consistent with the distinct separation between neutral words and gender attribute words observed in Figure 4(d).

	Gender			Overall		
	LMS	SS	ICAT	LMS	SS	ICAT
Original	86.54	63.24	63.63	**84.40**	58.83	69.50
Gender-Neutral	84.45	59.95	67.65	79.30	54.00	72.96
SCM-Neutral	85.52	61.06	66.60	84.34	55.59	74.91
Gender-SCM	84.98	**58.44**	70.64	82.84	55.21	74.21

Table 2: Evaluation results of the original model and three debiased models that have been debiased using three debiasing methods.

4.2 How Does SCM-Based Debiasing Achieve Group-Agnostic Stereotype Elimination?

Since SCM-Neutral does not rely on attribute words and experiments revealed that Gender-Neutral debiasing also slightly enhanced the language model's performance on religion-related stereotypes, we speculate that multi-group stereotypical objects included in neutral words contribute to the group-agnostic nature of SCM-based debiasing. To verify the above assertion, we divide the neutral words into two categories: *intersectional* neutral words and *group-specific* neutral words. A typical example of intersectional neutral words is "CEO", whose stereotype is the white male that involves both racial and gender groups. Therefore, when using SCM-Neutral to eliminate the stereotype of the neutral word "CEO", both race-related and gender-related stereotypes will be alleviated. In contrast, group-specific neutral words correspond only to stereotypes of a single demographic group. Next, we validate our assertion by removing group-specific neutral words and assessing the corresponding performance .

Group-Specific Selection As outlined in Algorithm 1, the process of selecting group-specific neutral words for two groups begins with calculating the stereotype correlation strength $\mathcal{L}(\mathcal{A}, t)$ between the group attribute words $\mathcal{A}$ and neutral word t. Next, neutral words are sorted in descending order based on their stereotype correlation strength, and the top k neutral words are selected. Finally, the difference set ($\mathcal{N}_{\text{gender}}$ or $\mathcal{N}_{\text{religion}}$) of two neutral word sets for each group is computed as the group-specific neutral word set. To measure the strength of the stereotype association between the corresponding groups of neutral word t and attribute set

		LMS↑				SS→ 50				ICAT↑			
		Original	G-S	R-S	Full	Original	G-S	R-S	Full	Original	G-S	R-S	Full
StereoSet	Gender	86.54	86.10	85.09	85.52	63.24	62.38	62.20	61.06	63.63	64.78	64.34	66.60
	Religion	84.27	84.00	82.88	83.12	59.94	58.36	60.89	59.28	67.51	69.97	64.83	67.70
	Overall	84.40	83.86	83.04	84.34	58.83	57.16	57.68	55.59	69.50	71.86	70.29	74.91

* G-S and R-S correspond to cases that remove gender-specific words and religion-specific words from the neutral words set, respectively.

Table 3: Experimental results correspond to the performance of four models: the unprocessed original model, the model debiased after removing the gender-specific neutral words (G-S), the model debiased after removing the religion-specific neutral words (R-S) and the model debiased with the full set of neutral words (Full).

$\mathcal{A}$, we design the function $\mathcal{L}(\mathcal{A}, t)$ based on $\mathcal{L}_{\text{bias}}$. The calculation of $\mathcal{L}(\mathcal{A}, t)$ is as follows:

$$\mathcal{L}(\mathcal{A}, t) = \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_a} \left(\boldsymbol{v}(a)^\top E(t, x; \boldsymbol{\theta}) \right)^2, \qquad (10)$$

where $v(a)$ refer to the average contextualizd word embedding of the attribute a and $v(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E(a, x; \theta)$.

Algorithm 1 Group-Specific Selection

```
**Input:** $\mathcal{A}_{\text{gender}}, \mathcal{A}_{\text{religion}}, \mathcal{N}, k$
**Output:** $\mathcal{N}_{\text{gender}}, \mathcal{N}_{\text{religion}}$
$\mathcal{D}_{\text{gender}} \leftarrow \{\}, \mathcal{D}_{\text{religion}} \leftarrow \{\}$.
for each $t \in \mathcal{N}$ do
    $\mathcal{D}_{\text{gender}}[t] \leftarrow \mathcal{L}(\mathcal{A}_{\text{gender}}, t)$;
    $\mathcal{D}_{\text{religion}}[t] \leftarrow \mathcal{L}(\mathcal{A}_{\text{religion}}, t)$;
end for
                                        ▷ Sort in descending order.
$\text{sort\_values}(\mathcal{D}_{\text{gender}})$
$\text{sort\_values}(\mathcal{D}_{\text{religion}})$
$\mathcal{N}_{\text{gender}} \leftarrow \mathcal{D}_{\text{gender}}.\text{keys}[: k]$
$\mathcal{N}_{\text{religion}} \leftarrow \mathcal{D}_{\text{religion}}.\text{keys}[: k]$
                                        ▷ Calculate the difference set.
$\mathcal{N}_{\text{gender}} \leftarrow \mathcal{N}_{\text{gender}} - (\mathcal{N}_{\text{gender}} \cap \mathcal{N}_{\text{religion}})$
$\mathcal{N}_{\text{religion}} \leftarrow \mathcal{N}_{\text{religion}} - (\mathcal{N}_{\text{religion}} \cap \mathcal{N}_{\text{gender}})$.
return $\mathcal{N}_{\text{gender}}, \mathcal{N}_{\text{religion}}$
```

To verify the assertion, we first acquire the gender-specific and religion-specific neutral words using Algorithm 1. Subsequently, we separately remove these gender-specific and religion-specific neutral words from the complete set of neutral words. Finally, we train the debiased model using ϵ-DPCE along with the set of processed neutral words, respectively. We summarize the results of above experiments in Table 3. The inclusion of group-specific neutral words directly influences the effectiveness of stereotype elimination for the respective demographic group. This demonstrates that **neutral words encompassing stereotypical objects from multiple groups contribute to the group-agnostic nature of SCM-Neutral (SCM-based debiasing).** The lower SS score of G-S compared to the original model is attributed to the existence of intersection neutral words.

5 Related Work

There are several existing methods for debiasing language models, including Counterfactual Data Augmentation (CDA; Zmigrod et al. 2019), Dropout (Webster et al., 2020), Self-Debias (Schick et al., 2021), and word embedding debiasing methods (Kaneko and Bollegala, 2021; Yang et al., 2023). Our work closely follows the word embedding method DPCE (Kaneko and Bollegala, 2021), which is more plain and more interpretable for subsequent analysis. Omrani et al. (2023) was the first to propose SCM-based group-agnostic debiasing method. Originating from social psychology, Stereotype Content Model (SCM; Fiske et al., 2002) offers valuable insights for group-agnostic debiasing.

Compared with DPCE proposed by Kaneko and Bollegala (2021), we have developed the ϵ-DPCE, which bridges the gap between the expected and the actual performance of debiased model. On the other hand, compared with SCM-based debiasing method proposed by Omrani et al. (2023), we explore further by explaining both why and how SCM-based debiasing achieve group-agnostic debiasing.

6 Conclusion

In this work, we explored the role of neutral words in group-agnostic debiasing of language models from two perspectives. First, we demonstrated that incorporating a loss function grounded in neutral word semantics into DPCE effectively preserved potential semantic modeling, which led to the development of ϵ-DPCE. Second, through our proposed SCM-Projection method, we revealed that the underlying mechanism of SCM-based debiasing is repositioning the neutral words in the SCM space. Our analysis further showed that multi-group stereotypical objects embedded within neutral words contributed significantly to the group-agnostic properties of SCM-based debiasing. Overall, our findings highlight the essential role of neutral words in group-agnostic debiasing and offer valuable insights into their effective integration into bias mitigation methods.

7 Limitations

This work has explored the role of neutral words in group-agnostic debiasing of language models. As no multilingual stereotype benchmark equivalent to StereoSet was available, our stereotype evaluations of language models were limited to English sentences. Furthermore, due to the limitation inherited from StereoSet, gender-related stereotypes in our experiments and analysis are binary and do not involve non-binary gender.

Acknowledgments

This work was supported in part by Key Program of Guangdong Province under Grant 2021QN02X166, and in part by the National Natural Science Foundation of China (Project No. 72031003). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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

Thang Viet Doan, Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness definitions in language models explained. *arXiv preprint arXiv:2407.18454*.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878–902.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings* *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371. Association for Computational Linguistics.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

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.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Words Used in Our Experiments

A.1 Attribute Words

Female she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts, niece, nieces.

Male he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews.

A.2 SCM Words

Warm social, warm, popular , good, right , kind, loyal, pleasant, friendly, funny, moral, fair, sympathetic, sensitive, cooperative, innocent, liked, responsible, genuine, polite, trustworthy, reliable, caring, helpful, thoughtful.

Cold antisocial, cold, unpopular, bad, wrong, mean, treacherous, unpleasant, unfriendly, boring, immoral, unfair, unsympathetic, insensitive, selfish, criminal, disliked, irresponsible, fake, rude, untrustworthy, unreliable, uncaring, unhelpful, inconsiderate.

Competent able, bright, brilliant, competent, capable, wise, rational, practical, dominant, dependent, confident, active, efficient, ambitious, determined, critical, secure, daring, educated, aggressive, motivated, intelligent, graceful, creative, energetic.

Incompetent unable, stupid, dumb, incompetent, incapable, unwise, irrational, impractical, submissive, independent, insecure, inactive, inefficient, lazy, doubtful, naive, vulnerable, cautious, uneducated, docile, unmotivated, unintelligent, clumsy, unimaginative, lethargic.

A.3 Neutral Words

aerobics, adventurer, apparel, aggressive, assistant, tycoon, baker, warrior, bathing, ambitious, beautiful, trucker, beauty, welder, blonde, strong, bookkeeper, terrorist, ca, soldier, cashier, astronomer, chatty, sniper, cheerleader, skipper, cheerleading, banker, clerk, baseball, cocktail, sergeant, cooking, bodyguard, counselor, boss, crafting, boxer, cute, cabbie, dancer, captain, educator, cardiologist, emotional, carpenter, flirt, ceo, flirtatious, chairperson, flower, chancellor, gossip, chef, graceful, colonel, hairdresser, commander, hairdryer, conductor, homemaker, police, hooker, custodian, housekeeper, dentist, housekeepers, detective, housework, diplomat, hula, doctor, indoor, driving, jealousy, drummer, jewelry, economist, kawaii, electrician, laundering, engineer, librarian, engineering, librarians, entrepreneur, lotion, financier, lovely, firefighter, marvelous, footballer, mirror, gambler, moisturizer, gamer, nanny, gangster, neat, geek, nurse, geeks, nursery, gentle, nurses, guitarist, nurturing, industrialist, parenting, inventor, passive, investigator, pink, laborer, pretty, lawyer, receptionist, leader, ribbon, lieutenant, romance, lifeguard, romantic, magistrate, secretary, manager, selfie, marshal, server, mathematician, sew, mechanic, sewing, muscle, shopping, muscular, smoothie, owner, soft, philosopher, softball, physicist, stylist, pilot, submissive, plumber, sweet, politician, tailor, president, tall, professor, teacher, programmer, thin, rugby, violinist, sailor, waiter, science, weak, scientist, yoga, sculptor, hysterical, blue, makeup, football, executive, management, professional, corporation, salary, office, business, career, home, parents, children, family, cousins, marriage, wedding, relatives, math, algebra, geometry, calculus, equations, computation, numbers, addition, poetry, art, dance, literature, novel, symphony, drama, sculpture, science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy, Shakespeare.

B Stereotype Content Model (SCM)

Stereotype Content Model (SCM) divides stereotypes into four types: High Warmth-High Competence (HW-HC), Low Warmth-High Competence (LW-HC), High Warmth-Low Competence (HW-LC) and Low Warmth-Low Competence (LW-LC), corresponding to four quadrants respectively. For any given group, SCM can illustrate its stereotypical position in the SCM space based on evaluations of these two dimensions. For example, the American middle class is the representative group of HW-HC, and the homeless individuals are the representatives of LW-LC. It should be noted that the hypothesis that the two dimensions of warmth and competence determine the distribution of outgroups has been verified through cross-cultural studies in 17 different countries and regions.

Existing studies on applying the SCM to language model debiasing have made some progress. Fraser et al. (2021) first demonstrated the feasibility of projecting stereotypes into the SCM space and proposed a method for defining the axes of warmth and competence in semantic embedding space. However, Fraser et al. (2021) primarily focused on static word embeddings and did not explore effective debiasing solutions. In contrast, Ungless et al. (2022) confirmed that the SCM holds for contextualized word embeddings and implemented debiasing by fine-tuning the BERT following the DPCE. Omrani et al. (2023) proposed an SCM-based debiasing method and tested its effectiveness on both statistic word embeddings and contextualized word embeddings. More importantly, Omrani et al. (2023) proposed and discovered that the SCM-based debiasing method is social-group-agnostic. **However, the underlying reason why the SCM-based debiasing method can eliminate social-group-agnostic stereotypes remains unexplored.**

C Gender-related stereoype elimination experiments using `Llama-3.2-1B`

C.1 Debiasing by DPCE

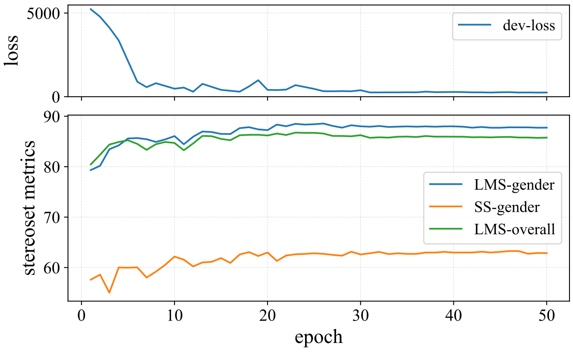


Figure 5: As the number of DPCE training epochs increases, the corresponding dev-loss, LMS-gender, SS-gender, and LMS-overall curves of `Llama-3.2-1B`.

Unlike `bert-large-uncased` in Figure 1, LMS-overall value of `Llama-3.2-1B` did not continue to decline; instead, it dropped to around 80 at the beginning of training (compared to the original value of 93.31 in Table 4).

C.2 Debiasing by ϵ-DPCE

As shown in Figure 6, although `Llama-3.2-1B`'s LMS-overall also exhibited a significant decline during the early stages of training with ϵ-DPCE, it improved more effectively in the later stages compared to DPCE in Figure 5, with both the LMS-overall and LMS-gender values showing better results.

To improve the generalizability of findings, we used `Llama-3.2-1B` (released on September 25, 2024) as the experimental subject and conducted the same experiments as in Section 3. In the Table 4, we present the original metric data for `Llama-3.2-1B`, along with the test results for the model after applying DPCE and ϵ-DPCE debiasing. These results are based on two optimal model selection strategies: one uses the model corresponding to the epoch with the minimum development loss (Min dev-loss epoch), and the other selects the model from the final epoch (The last epoch).

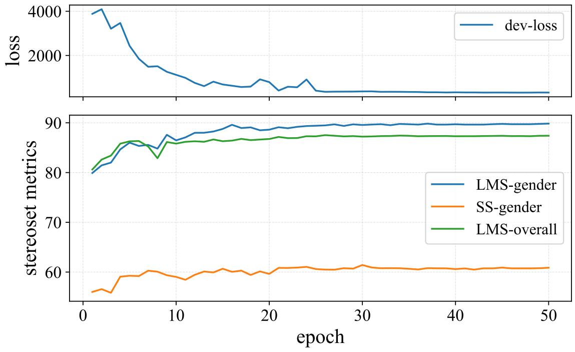


Figure 6: As the number of ϵ-DPCE training epochs increases, the corresponding dev-loss, LMS-gender, SS-gender, and LMS-overall curves of `Llama-3.2-1B`.

	Origin	Min dev loss		The last epoch	
		DPCE	ϵ-DPCE	DPCE	ϵ-DPCE
$\text{LMS}_{\text{gender}} \uparrow$	93.88	87.73	**89.72**	87.74	**89.84**
$\text{SS}_{\text{gender}} \rightarrow 50$	71.83	62.89	60.75	62.86	60.87
$\text{ICAT}_{\text{gender}} \uparrow$	52.89	65.11	**70.43**	65.18	**70.30**
$\text{LMS}_{\text{overall}} \uparrow$	93.31	85.72	**87.34**	85.76	**87.40**
$\text{SS}_{\text{overall}} \rightarrow 50$	64.78	59.10	59.84	59.08	59.89
$\text{ICAT}_{\text{overall}} \uparrow$	65.73	70.13	**70.15**	**70.19**	70.11

Table 4: Experimental results of debiased `Llama-3.2-1B` correspond to the last epoch and the epoch with minimal dev-loss in ϵ-DPCE.

According to the Table 4, we can observe a similar phenomenon as noted in the Section 3: compared to DPCE, ϵ-DPCE can reduce the SS (stereotype score) and better preserve the performance of the LMS (language modelling score).