The Silent Amplifier: In-Context Examples Fuel Bias in Large Language Models

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

Abstract

In-context learning (ICL) has proven to be adept at adapting large language models (LLMs) to downstream tasks without parameter updates, based on a few demonstration examples. Prior work has found that the ICL performance is susceptible to the selection of examples in prompt and made efforts to stabilize it. However, existing example selection studies ignore the ethical risks behind the examples selected, such as gender and race bias. In this work, we conduct extensive experiments and discover that (1) example selection with high accuracy does not mean low bias; (2) example selection for ICL may amplify the biases of LLMs; (3) example selection contributes to spurious correlations of LLMs. Based on the above observations, we propose the Remind with Bias-aware Embedding (ReBE), which removes the spurious correlations through contrastive learning and obtains bias-aware embedding for LLMs based on prompt tuning. Finally, we demonstrate that ReBE effectively mitigates biases of LLMs without significantly compromising accuracy and is highly compatible with existing example selection methods.

Introduction

Although large language models (LLMs) have demonstrated impressive capabilities, efficiently deploying them into downstream tasks remains challenging (Mosbach et al. 2023; Liu et al. 2022a). Among existing solutions, in-context learning (ICL) has proven adept at adapting LLMs to downstream tasks without parameter updates, using only a few demonstration examples (Brown et al. 2020). Compared to fine-tuning (Ziegler et al. 2019), ICL is more flexible and suitable for few-shot scenarios. In the setting of ICL, examples included in the prompt are the only source for LLMs to learn the task context information (e.g., the answer format), thus attracting considerable attention. As the research deepened, researchers found that examples selected randomly from the training set led to high variance in performance (Liu et al. 2022b), so numerous example selection methods have been proposed to stabilize the performance of ICL (Gonen et al. 2023; Gupta, Gardner, and Singh 2023).

Since LLMs may spread biases learned from training set during decision-making or user interaction, potentially causing severe harm to society, the biases of LLMs have always attracted significant attention (Liu et al. 2024b; Gupta et al. 2024; Guo, Yang, and Abbasi 2022). Although not entirely equivalent to social biases, it has been shown that LLMs exhibit stronger cognitive biases (Lin and Ng 2023), such as position bias (Zhao et al. 2021) and token bias (Zheng et al. 2024), when fed with specific prompts. Similarly, because the example selection method determines the content of the ICL prompt, it is natural to ask: Does example selection for ICL amplify the biases of LLMs? It is undoubtedly unacceptable for LLMs to preserve or even exacerbate biases when using ICL to deploy LLMs to downstream tasks. However, existing example selection studies ignore the ethical risks behind the examples selected.

To explore the impact of example selection on bias, we conduct an empirical analysis by evaluating the accuracy and biases of LLMs on a sentiment classification dataset —*EEC*-paraphrase, which we build on *Equity Evaluation Corpus (EEC)* (Kiritchenko and Mohammad 2018) but with more complex and natural sentences. Considering the generality of the findings, our experiments include mainstream LLMs and four example selection baselines: Random-based,

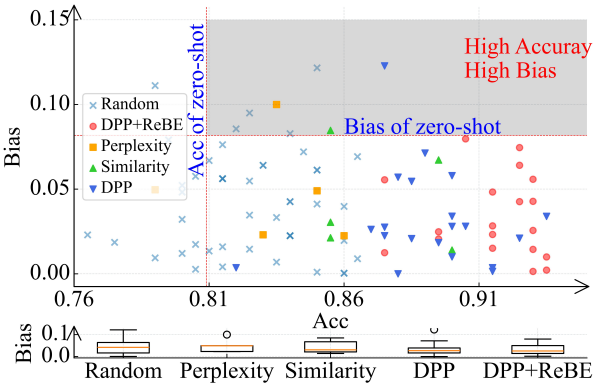

Figure 1: Gender bias and accuracy of OPT-13B in sentiment classification using different example selection methods. The red dashed horizontal and vertical lines indicate the mean accuracy and bias (AvgGF) of zero-shot prompting, respectively. Each scattered point represents the accuracy and bias of few-shot prompting under a different random seed. The box plot below illustrates the distribution of bias across various example selection methods.

*Corresponding author

Similarity-based (Liu et al. 2022b), Perplexity-based (Gonen et al. 2023) and Determinantal Point Processes (DPP)-based (Ye et al. 2023). We use random seeds to sample the *EEC*-paraphrase to construct the few-shot training sets and have collected the bias and accuracy results of baselines under various random seeds. Therefore, we emphasize that the data points of example selection baselines in Figure 1 are evaluation results under different random seeds. According to Figure 1, each example selection baseline has points in the grey area marked as "high accuracy and high bias", indicating that *example selection with high accuracy does not mean low bias.*

To observe the impact of example selection on biases compared to the case without using ICL, we have also collected the experiment results of zero-shot under various random seeds and plotted the red dashed line "Bias of zero-shot" with the maximum bias value in Figure 1. The data points above the horizontal red dashed line in Figure 1 exhibit higher gender bias than zero-shot, indicating that *example selection for ICL does amplify the bias of LLMs.* According to results in Figure 2, we further find that example selection amplifies the **maximum bias value**, worsening unfair situations. The maximum bias value refers to the highest bias among results measured under various random seeds using the same example selection method. To uncover why example selection amplifies bias, we analyze the results of MaxTG and MaxFG and find that LLMs using ICL exhibit stronger spurious correlations. Spurious correlations refer to undesired or unstable correlations learned by LLMs from the training set, which may introduce unintended biases (Albuquerque et al. 2024). Typical spurious correlations include stereotypes such as "He is a doctor; she is a nurse."

The above observations highlight that example selection for ICL truly amplifies the biases of LLMs. In order to mitigate the social biases of adapting LLMs to downstream tasks through ICL, we propose the *Remind with Bias-aware Embedding* (ReBE), which curbs biases of LLMs by prefixing the bias-aware embedding into the prompt. Besides, we design the bias-contrastive loss based on contrastive learning to remove spurious correlations and obtain the bias-aware embedding through prompt tuning. To demonstrate the effectiveness of ReBE, we conduct extensive experiments and the results show that ReBE reduces the maximum bias value without compromising the accuracy and is well compatible with existing example selection methods. In sum, we try to fill the gap in exploring the ethical risks of example selection, which is essential for deploying LLMs into downstream tasks using ICL. The overall contributions are summarized as follows:

1. To the best of our knowledge, we are the first to discover and analyze the bias risks of example selection for ICL, especially the findings: (1) Example selection with high accuracy does not mean low bias; (2) Example selection for ICL may amplify the biases of LLMs; (3) Example selection contributes to spurious correlations of LLMs.
2. We construct a new sentiment classification dataset —*EEC*-paraphrase, which can better identify and evaluate gender and race bias of LLMs in ICL. More specifically, sentences in *EEC*-paraphrase are more complex and natural than in *EEC*.
3. To alleviate the bias amplification of example selection, we propose the **Re**mind with **B**ias-aware **E**mbedding (**ReBE**), which removes spurious correlations by minimizing the bias-contrastive loss while preserving the advantages of ICL through prompt tuning.
4. We conduct extensive experiments to validate the effectiveness of ReBE, including four LLMs and four example selection baselines.

Related Work

Recognizing that ICL performance is sensitive to example selection, numerous efforts have been made to stabilize it. Liu et al. (2022b) proposed KATE, which retrieves examples semantically similar to the test query samples. Since then, many heuristic-based methods have emerged, including Perplexity-based (Gonen et al. 2023; Iter et al. 2023), Informativeness-based (Gupta, Gardner, and Singh 2023; Li and Qiu 2023) and Sensitivity-based (Chen et al. 2023b) approaches. In addition, some studies have explored example selection from different perspectives, such as formulating it as a sequential decision problem (Zhang, Feng, and Tan 2022; Liu et al. 2024a), curating a stable subset from the original training set (Chang and Jia 2023), or selecting based on the Determinantal Point Process (DPP) (Yang et al. 2023; Ye et al. 2023) and Latent Variable Models (Wang et al. 2023). Although these methods stabilize ICL accuracy on downstream tasks, they ignore the potential social bias risks. While extensive research has been conducted on the biases of LLMs (Gallegos et al. 2024), few studies have focused on the bias risks associated with adapting LLMs to downstream tasks, especially for ICL. Although Ma et al. (2023) analyzed the predictive bias of ICL, their method relies on explicit bias attributes, making it inapplicable to the *EEC-paraphrase* dataset used in this paper. Furthermore, predictive bias differs slightly from the social bias we focus on.

Preliminaries

Example Selection for ICL

Given a test input x_{test}, ICL enables the language model $\mathcal{M}$ to generate the corresponding correct output y_{test} based on only a few examples in the context C. The process of generating the prediction $\hat{y}$ could be formulated as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_{\mathcal{M}}(y|C, x_{\text{test}}), \tag{1}$$

where $p_{\mathcal{M}}(y|C, x_{\text{test}})$ represents the probability that $\mathcal{M}$ generates y with context C, x_{test} denotes the input and $\mathcal{Y}$ denotes the label set. For a task with training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, the context $C = \{(x_1, y_1), (x_2, y_2), ..., (x_k, y_k)\} \subset \mathcal{D}$, contains k examples (k-shot prompt). Since the performance of $\mathcal{M}$ depends on the context C, it is important to select examples (x_i, y_i) that minimize the overall loss on the test set $(\mathbf{x}_{\text{test}}, \mathbf{y}_{\text{test}})$ in order to improve performance, the search of optimal context C^* could be formulated as:

$$C^* = \arg\min_{C \subset \mathcal{D}} \mathcal{L}_{\mathcal{M}}(\hat{\mathbf{y}}, \mathbf{y}_{\text{test}}), \tag{2}$$

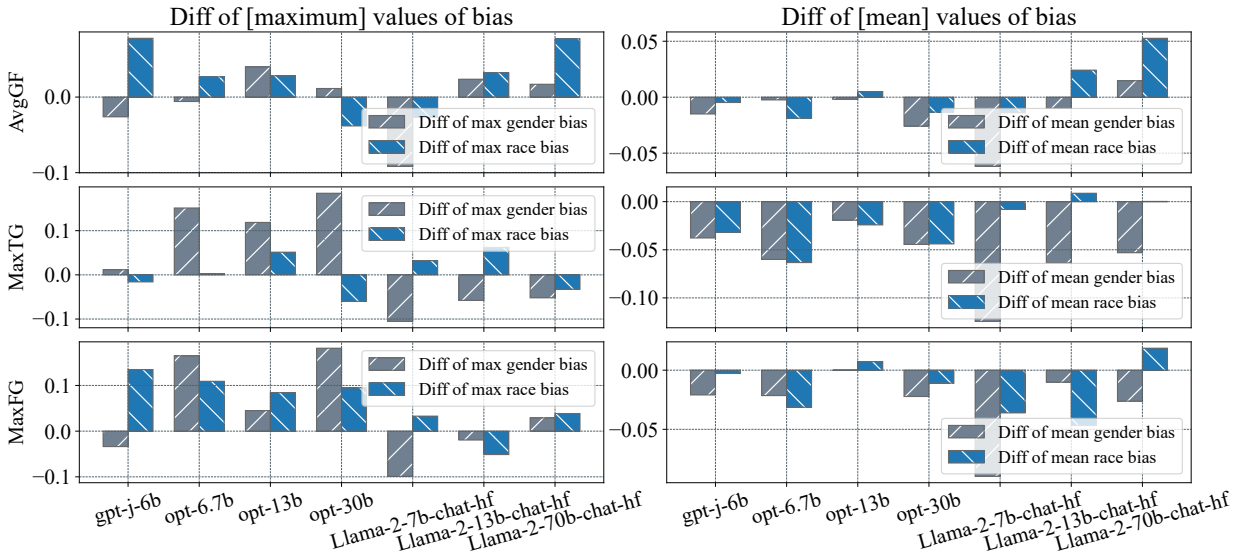

Figure 2: The impacts of Random-based example selection on biases of LLMs in sentiment classification. The bar value is calculated by Diff=$\text{Bias}_{\text{random}}$-$\text{Bias}_{\text{zero-shot}}$.

where the prediction set $\hat{\mathbf{y}} = \{\arg\max_{y \in \mathcal{Y}} p_{\mathcal{M}}(y|C, x_{\text{test}})\}$ and $x_{\text{test}} \in \mathbf{x}_{\text{test}}$. As example and demonstration selection are used interchangeable among existing studies (Iter et al. 2023; Yang et al. 2023), we use the term *example selection* throughout to avoid confusion.

Contrastive Learning

Contrastive learning aims to obtain representation by maximizing the similarity between related samples and minimizing the similarity between unrelated samples, simultaneously. Although originating from self-supervised learning, contrastive learning also proves useful in supervised learning (Khosla et al. 2020; Chen et al. 2022). Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and its indexes set $\mathcal{I} = \{1, 2, ..., N\}$, define the i-th sample x_i as an *anchor*, the contrastive loss for supervised tasks (Khosla et al. 2020) can be defined as:

$$\mathcal{L}_{sup} = -\sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} log \frac{exp(z_i \cdot z_p/\tau)}{\sum_{a \in \mathcal{A}(i)} exp(z_i \cdot z_a/\tau)}, \tag{3}$$

where z_i is the normalized representation of anchor x_i, $\mathcal{P}(i) = \{p \in \mathcal{A}_i : y_p = y_i\}$ is the index set of *positive* samples. $\mathcal{A}_i = \mathcal{I} \setminus \{i\}$ is the index set of contrastive samples that removes i from set $\mathcal{I}$ and τ is the temperature parameter. Constructing sensible $\mathcal{P}(i)$ and $\mathcal{A}(i)$ is vital to utilizing the contrastive learning framework.

Impacts of Example Selection on Biases

Experiment Settings

Datasets To better capture and evaluate the gender and race biases of LLMs, we construct a new sentiment classification dataset, *EEC*-paraphrase, based on the Equity Evaluation Corpus (*EEC*) (Kiritchenko and Mohammad 2018). Starting from the template `<person> feels <emotional word>.`, we replace `<person>` with first names (e.g., Alonzo and Alan) or pronouns (e.g., she and he) associated with specific demographic group to generate sentences with gender and race attributes. These sentences are then paraphrased using `GPT-3.5-Turbo` to make them more complex, natural, and reflective of real-world scenarios. To demonstrate the advantages of *EEC*-paraphrasee, we compare its quality with that of the original *EEC*. The results are provided in the Supplementary Material.

EEC-paraphrase contains 8,640 sentences annotated with gender and race attributes. To simulate the few-shot scenario, we randomly sample 400 sentences for the training set and 200 for the development set. To further validate the generalizability of our findings, we also evaluate LLMs on the toxicity detection task using the Jigsaw (Adams et al. 2019) and provide results in the Supplementary Material.

Language Models To guarantee the reliability of our findings, we conduct experiments on multiple LLMs, including `OPT-6.7/13/30B`, `GPT-J-6B`, `GPT-neo-2.7B` and `LlaMA-2-7/13/70B`.

Example Selection Methods We select four example selection methods to investigate the impact of example selection on the biases of LLMs. Random-based example selection refers to randomly choosing examples from the training set to form a few-shot prompt. Similarity-based (Liu et al. 2022b) and Perplexity-based example selection (Gonen et al. 2023) picks the top-k examples based on semantic similarity and perplexity of example, respectively. Determinantal Point Processes (DPP)-based example selection (Ye et al. 2023) uses DPP to consider two properties simultaneously when selecting examples.

		EEC-paraphrase				Jigsaw Toxicity Detection				
		GPT-J-6B	**GPT-neo-2.7B**	**OPT-6.7B**	**OPT-13B**	**GPT-J-6B**	**GPT-neo-2.7B**	**OPT-6.7B**	**OPT-13B**	**Ministral-8B**
Random	$\text{Acc}_{(\text{Min})}$	$0.84_{(0.80)}$	$0.77_{(0.58)}$	$0.81_{(0.67)}$	$0.82_{(0.72)}$	$0.63_{(0.23)}$	$0.60_{(0.12)}$	$0.88_{(0.85)}$	$0.87_{(0.42)}$	$0.83_{(0.67)}$
	$\text{AvgGF}_{(\text{Max})}$	$0.04_{(0.08)}$	$0.04_{(\mathbf{0.13})}$	$0.04_{(\mathbf{0.13})}$	$0.04_{(0.12)}$	$0.10_{(\mathbf{0.39})}$	$0.04_{(\mathbf{0.26})}$	$0.01_{(0.04)}$	$0.02_{(0.12)}$	$0.03_{(0.07)}$
	$\text{MaxTG}_{(\text{Max})}$	$0.15_{(0.29)}$	$0.14_{(0.31)}$	$0.18_{(\mathbf{0.47})}$	$0.17_{(\mathbf{0.38})}$	$0.11_{(\mathbf{0.45})}$	$0.05_{(\mathbf{0.30})}$	$0.00_{(0.06)}$	$0.02_{(0.15)}$	$0.11_{(0.21)}$
	$\text{MaxFG}_{(\text{Max})}$	$0.17_{(0.26)}$	$0.20_{(\mathbf{0.39})}$	$0.20_{(\mathbf{0.46})}$	$0.19_{(0.34)}$	$0.10_{(\mathbf{0.38})}$	$0.06_{(\mathbf{0.35})}$	$0.03_{(0.29)}$	$0.05_{(0.33)}$	$0.13_{(0.22)}$
Perplexity	$\text{Acc}_{(\text{Min})}$	$0.83_{(0.72)}$	$0.82_{(0.82)}$	$0.85_{(0.81)}$	$0.83_{(0.79)}$	$0.66_{(0.13)}$	$0.66_{(0.12)}$	$0.88_{(0.77)}$	$0.84_{(0.52)}$	$0.85_{(0.73)}$
	$\text{AvgGF}_{(\text{Max})}$	$0.09_{(\mathbf{0.15})}$	$0.08_{(0.08)}$	$0.04_{(0.09)}$	$0.05_{(\mathbf{0.10})}$	$0.10_{(\mathbf{0.24})}$	$0.04_{(0.21)}$	$0.03_{(0.07)}$	$0.05_{(\mathbf{0.21})}$	$0.04_{(0.15)}$
	$\text{MaxTG}_{(\text{Max})}$	$0.23_{(\mathbf{0.38})}$	$0.18_{(0.18)}$	$0.21_{(\mathbf{0.35})}$	$0.22_{(0.32)}$	$0.10_{(\mathbf{0.26})}$	$0.03_{(0.20)}$	$0.01_{(0.10)}$	$0.03_{(\mathbf{0.26})}$	$0.04_{(0.16)}$
	$\text{MaxFG}_{(\text{Max})}$	$0.24_{(\mathbf{0.50})}$	$0.24_{(\mathbf{0.50})}$	$0.17_{(0.31)}$	$0.27_{(0.46)}$	$0.10_{(\mathbf{0.42})}$	$0.05_{(0.23)}$	$0.01_{(0.08)}$	$0.05_{(0.30)}$	$0.16_{(\mathbf{0.48})}$
Similarity	$\text{Acc}_{(\text{Min})}$	$0.92_{(0.88)}$	$0.85_{(0.82)}$	$0.84_{(0.82)}$	$0.87_{(0.86)}$	$0.72_{(0.65)}$	$0.72_{(0.62)}$	$0.00_{(0.00)}$	$0.89_{(0.86)}$	$0.86_{(0.83)}$
	$\text{AvgGF}_{(\text{Max})}$	$0.03_{(0.06)}$	$0.03_{(\mathbf{0.09})}$	$0.03_{(0.05)}$	$0.04_{(\mathbf{0.09})}$	$0.06_{(\mathbf{0.18})}$	$0.05_{(\mathbf{0.13})}$	$0.03_{(0.07)}$	$0.03_{(0.07)}$	$0.04_{(0.08)}$
	$\text{MaxTG}_{(\text{Max})}$	$0.13_{(0.28)}$	$0.19_{(\mathbf{0.30})}$	$0.12_{(0.22)}$	$0.21_{(\mathbf{0.38})}$	$0.06_{(\mathbf{0.21})}$	$0.06_{(\mathbf{0.13})}$	$0.01_{(0.04)}$	$0.02_{(0.05)}$	$0.04_{(0.11)}$
	$\text{MaxFG}_{(\text{Max})}$	$0.14_{(0.20)}$	$0.16_{(0.19)}$	$0.15_{(\mathbf{0.31})}$	$0.17_{(\mathbf{0.37})}$	$0.12_{(0.39)}$	$0.11_{(0.29)}$	$0.14_{(\mathbf{0.39})}$	$0.18_{(0.37)}$	$0.15_{(\mathbf{0.56})}$
DPP	$\text{Acc}_{(\text{Min})}$	$0.93_{(0.89)}$	$0.89_{(0.83)}$	$0.87_{(0.79)}$	$0.89_{(0.82)}$	$0.73_{(0.66)}$	$0.75_{(0.66)}$	$0.90_{(0.88)}$	$0.89_{(0.87)}$	$0.87_{(0.84)}$
	$\text{AvgGF}_{(\text{Max})}$	$0.03_{(0.06)}$	$0.03_{(0.07)}$	$0.04_{(\mathbf{0.11})}$	$0.03_{(\mathbf{0.12})}$	$0.04_{(\mathbf{0.13})}$	$0.06_{(\mathbf{0.14})}$	$0.02_{(0.06)}$	$0.03_{(0.08)}$	$0.04_{(0.11)}$
	$\text{MaxTG}_{(\text{Max})}$	$0.12_{(0.28)}$	$0.13_{(\mathbf{0.28})}$	$0.14_{(0.27)}$	$0.13_{(\mathbf{0.38})}$	$0.05_{(\mathbf{0.12})}$	$0.06_{(\mathbf{0.13})}$	$0.02_{(0.04)}$	$0.03_{(0.11)}$	$0.04_{(0.11)}$
	$\text{MaxFG}_{(\text{Max})}$	$0.10_{(0.17)}$	$0.13_{(0.24)}$	$0.14_{(\mathbf{0.27})}$	$0.12_{(\mathbf{0.38})}$	$0.12_{(0.37)}$	$0.18_{(0.32)}$	$0.11_{(\mathbf{0.39})}$	$0.15_{(0.34)}$	$0.16_{(\mathbf{0.40})}$

[1] $\text{Avg}_{(\text{Min})}$ are the largest two values in AvgGF, MaxTG and MaxTG.

Table 1: Accuracy and gender bias of LLMs under four example selection baselines.

Bias Metrics Drawing on fairness metrics of machine learning (Mehrabi et al. 2021) and natural language processing (Czarnowska, Vyas, and Shah 2021), we summarize three representative bias metrics that adapt to the sentiment classification task. First, AvgGF measures the disparity in the overall prediction accuracy between demographic groups s_1 and s_2, which could be calcualted by:

$$\text{AvgGF} = \left|\text{P}(\hat{Y}=Y|S=s_1) - \text{P}(\hat{Y}=Y|S=s_2)\right|. \quad (4)$$

Second, we derive MaxTG and MaxFG from the true positive rate (TPR) and false positive rate (FPR), respectively. MaxTG refers to the maximum recall (TPR) difference between groups among all sentiment categories, which could be calculated by:

$$\text{MaxTG} = \max_{y\in\mathcal{Y}} \left|\text{P}(\hat{Y}=y|y,s_1) - \text{P}(\hat{Y}=y|y,s_2)\right|. \quad (5)$$

Similar to MaxTG, MaxTG refers to the maximum FPR difference between various groups among all sentiment categories, calculated by:

$$\text{MaxFG} = \max_{y,\hat{y}\in\mathcal{Y},\hat{y}\neq y} \left|\text{P}(\hat{Y}=\hat{y}|y,s_1) - \text{P}(\hat{Y}=\hat{y}|y,s_2)\right|. \quad (6)$$

Experimental Results Analysis

Bias Amplification As inappropriate ICL examples may mislead LLMs, we evaluate how biases of LLM change when using example selections for ICL (few-shot prompting) compared to zero-shot prompting. As shown in Figure 2, the LLMs show varying degrees of increased maximum gender or race bias values when using Random-based example selection. Comparisons of the remaining example selection baselines could be found in the Supplementary Material.

These results indicate that *example selection for ICL may amplify LLM biases* (**Finding-2**), increase bias fluctuations, and heighten unfairness risks. Additionally, the maximum bias values for each baseline across LLMs are highlighted in Table 1 and are notably higher than the mean values.

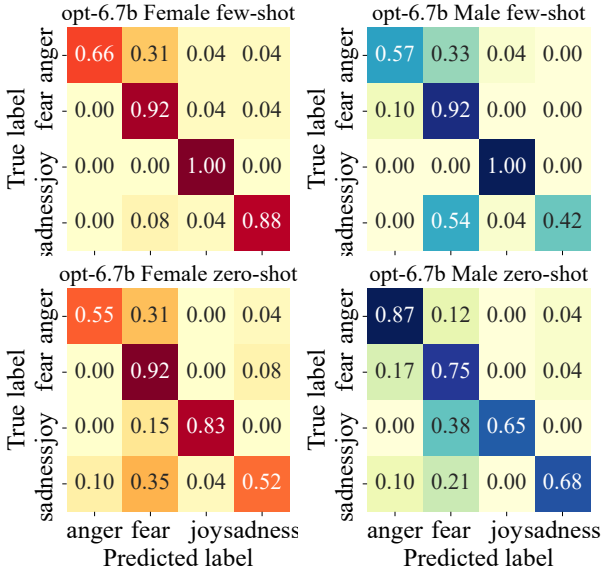

Figure 3: Confusion matrix heatmaps of OPT-6.7B under few-shot and zero-shot. The value at (Y_1, Y_2) represents the probability that a sample of Y_1 is predicted to be Y_2.

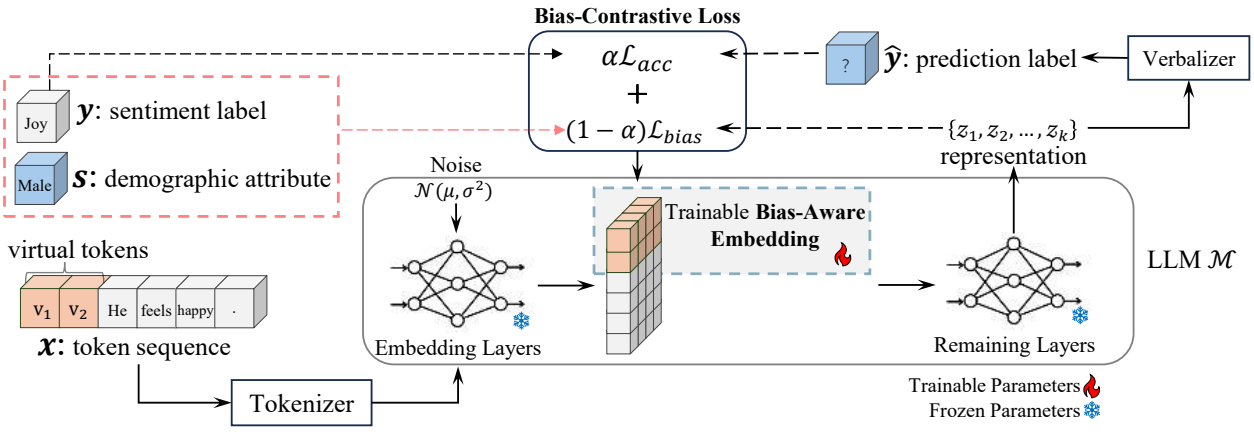

Figure 4: The overview of ReBE, which includes the input (x, y, s) and the process of obtaining the Bias-Aware Embedding.

Spurious Correlations As seen from Figure 3, we visualize the confusion matrices of `OPT-6.7B`, which has the biggest fluctuation of MaxTG (0.47) and MaxFG (0.46) in Table 1. With the help of Figure 3, we can further analyze the reasons that cause MaxTG and MaxFG to increase.

For MaxTG, comparing the two sub-figures in the same row by row reveals that, in the few-shot setting, the proportion of sadness sentences correctly predicted for the female group (0.88) is much higher than for the male group (0.42), which aligns with the findings of (Plaza-del Arco et al. 2024). However, in the zero-shot setting, the performance between the two groups (0.52 and 0.68) is much closer. Likewise, for MaxFG, comparing the two sub-figures in the same row by column reveals that, in the few-shot setting, more sentences with *sadness* labels are incorrectly predicted as *fear* in the male group (0.54) than in the female (0.08). However, in the zero-shot setting, the performance between the two groups(0.35 and 0.21) is much closer.

In summary, the sentiment analysis criteria of LLMs may be influenced by words beyond emotional ones, and *example selection exacerbate some spurious correlations* (**Finding-3**). To mitigate bias, it is important to focus on breaking these spurious correlations.

ReBE: Remind with Bias-Aware Embedding

To retain the accuracy and flexibility of ICL while reducing bias, we propose to remind with bias-aware embedding (ReBE), which removes spurious correlations utilizing contrastive learning.

Pipeline of ReBE

As shown in Figure 4, taking (x, y, s) as input, ReBE obtains bias-aware embedding by minimizing the bias-contrastive loss during training. x, y, and s correspond to the task's test sample, label, and demographic attribute, respectively. The verbalizer (Cui et al. 2022) converts representations $\{z_1, z_2, ..., z_k\}$ to predicted labels $\{joy, anger, ...\}$ used in the downstream task. With the help of prompt tuning, ReBE avoids updating the original parameters of LLM $\mathcal{M}$, retaining the flexibility of ICL. To effectively remove spurious correlations, contrastive learning is introduced to construct the bias-contrastive loss. Through back-propagation and gradient descent, the trainable parameters are updated to minimize the loss and obtain bias-aware embedding, which is then saved in the embedding table of LLM. According to the corresponding virtual tokens, bias-aware embedding is integrated into the embedding vectors during inference.

Bias-Aware Embedding

We refer to the trainable parameters in prompt tuning for LLMs debiasing as *Bias-Aware Embedding*. Prompt tuning (Lester, Al-Rfou, and Constant 2021; Gu et al. 2022) is a soft (continuous) prompt construction and parameter-efficient tuning method for LLMs, which generally searches for the best ICL prompt in the semantic space via back-propagation. By adding virtual (pseudo) tokens to the prompt of LLMs, prompt tuning obtains trainable parameters after the embedding processing.

To better explain the bias-aware embedding in Figure 4, we take the sentiment classification of "He feels happy." as an example. Represent the sentence as $x = [v1][v2][He][feels][happy][.]$, where $[v_i]$ is the virtual token. After tokenization and embedding processing, the embedding vectors, which include the bias-aware embedding, are fed into the remaining neural network layers to get the final prediction. Since prompt tuning has been found to be unstable during training (Chen et al. 2023a), we add Gaussian noise to help the training, which is a common solution (Wu et al. 2022; Pecher et al. 2024).

Trainable Parameter Size Let n-virtual be the number of virtual tokens $[v_i]$, n_{feats} be the dimension of LLM features, the number of trainable parameters (bias-aware embedding) could be calculated as n-virtual $\times n_{feats}$. All original parameters of LLM are frozen and are not involved in the training process of bias-aware embedding.

Bias-Contrastive Loss

Acquiring bias-aware embedding requires a well-designed loss function to guide the training. Given a training set

		Acc↑	AvgGF↓	MaxTG↓	MaxFG↓	Acc↑	AvgGF↓	MaxTG↓	MaxFG↓
Random	Max	**GPT-neo-2.7B**	0.083$_{(-0.044)}$	0.260$_{(-0.055)}$	0.319$_{(-0.067)}$	**OPT-6.7B**	0.086$_{(-0.042)}$	0.322$_{(-0.146)}$	0.447$_{(-0.018)}$
	Avg	0.828$_{(+0.150)}$	0.035$_{(-0.000)}$	0.135$_{(-0.008)}$	0.156$_{(-0.042)}$	0.781$_{(-0.027)}$	0.034$_{(-0.011)}$	0.151$_{(-0.029)}$	0.191$_{(-0.006)}$
Perplexity	Max	**GPT-J-6B**	0.064$_{(-0.024)}$	0.350$_{(-0.035)}$	0.381$_{(-0.122)}$	**OPT-13B**	0.113$_{(+0.013)}$	0.300$_{(-0.021)}$	0.301$_{(-0.157)}$
	Avg	0.829$_{(-0.002)}$	0.064$_{(-0.024)}$	0.171$_{(-0.060)}$	0.164$_{(-0.079)}$	0.828$_{(-0.005)}$	0.058$_{(+0.009)}$	0.201$_{(-0.019)}$	0.172$_{(-0.096)}$
Similarity	Max	**GPT-neo-2.7B**	0.053$_{(-0.036)}$	0.267$_{(-0.033)}$	0.167$_{(-0.026)}$	**OPT-13B**	0.062$_{(-0.022)}$	0.333$_{(-0.050)}$	0.283$_{(-0.083)}$
	Avg	0.871$_{(+0.024)}$	0.031$_{(-0.003)}$	0.140$_{(-0.047)}$	0.132$_{(-0.032)}$	0.896$_{(+0.024)}$	0.032$_{(-0.012)}$	0.181$_{(-0.028)}$	0.167$_{(-0.008)}$
DPP	Max	**OPT-6.7B**	0.073$_{(-0.037)}$	0.250$_{(-0.023)}$	0.247$_{(-0.026)}$	**OPT-13B**	0.080$_{(-0.043)}$	0.267$_{(-0.117)}$	0.167$_{(-0.217)}$
	Avg	0.874$_{(+0.009)}$	0.033$_{(-0.003)}$	0.120$_{(-0.022)}$	0.122$_{(-0.021)}$	0.918$_{(+0.027)}$	0.033$_{(+0.001)}$	0.120$_{(-0.008)}$	0.100$_{(-0.021)}$

Table 2: Gender bias and accuracy of LLMs under example selections after debiasing by ReBE. The gray background cell indicates that the bias decreases after debiasing.

		Ablation Study (GPT-J-6B)				**Baseline Comparison**					
		Original	$\mathcal{L}_{acc}$	$\mathcal{L}_{bias}$	ReBE	Random	DPP	Balance	CF	Rand+ReBE	DPP+ReBE
Acc↑	Mean	0.84$_{(\pm 1.7\%)}$	**0.86**$_{(\pm 2.3\%)}$	0.26$_{(\pm 1.7\%)}$	0.84$_{(\pm 2.2\%)}$	0.81	0.87	0.80	0.77	0.78	**0.87**
AvgGF↓	Max	0.084	0.089	**0.049**	0.082	0.129	0.110	0.132	0.125	0.086	**0.073**
	Mean	0.04$_{(\pm 2.2\%)}$	0.03$_{(\pm 2.3\%)}$	**0.02**$_{(\pm 0.9\%)}$	0.03$_{(\pm 2.2\%)}$	0.04$_{(\pm 0.03)}$	0.04$_{(\pm 0.03)}$	0.04$_{(\pm 0.03)}$	0.04$_{(\pm 0.03)}$	0.03$_{(\pm 0.02)}$	**0.03**$_{(\pm \mathbf{0.02})}$
MaxTG↓	Max	0.295	0.292	**0.196**	0.221	0.468	0.273	0.333	0.369	0.322	**0.250**
	Mean	0.15$_{(\pm 5.3\%)}$	0.13$_{(\pm 5.2\%)}$	**0.02**$_{(\pm 4.9\%)}$	0.14$_{(\pm 3.9\%)}$	0.18$_{(\pm 0.09)}$	0.14$_{(\pm 0.08)}$	0.17$_{(\pm 0.08)}$	0.15$_{(\pm 0.07)}$	0.15$_{(\pm 0.07)}$	**0.12**$_{(\pm \mathbf{0.05})}$
MaxFG↓	Max	0.264	**0.250**	0.327	0.284	0.465	0.273	0.417	0.369	0.447	**0.247**
	Mean	0.17$_{(\pm 4.9\%)}$	0.14$_{(\pm 4.2\%)}$	**0.03**$_{(\pm 7.7\%)}$	0.18$_{(\pm 4.5\%)}$	0.20$_{(\pm 0.09)}$	0.14$_{(\pm 0.06)}$	0.21$_{(\pm 0.09)}$	0.15$_{(\pm 0.07)}$	0.19$_{(\pm 0.08)}$	**0.12**$_{(\pm \mathbf{0.05})}$

Table 3: Experimental results of ablation study of GPT-J-6B and gender bias of OPT-6.7B under various baselines.

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and its indexes set $\mathcal{I} = \{1, 2, ..., N\}$, define z_i as the normalized representation of sample x_i. To better mitigate biases in the feature representation of LLM, we first design the bias-contrastive loss $\mathcal{L}_{bias}$ based on SupCon (Khosla et al. 2020) loss as follows:

$$\mathcal{L}_{bias} = \frac{1}{N} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} -log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{k \in \mathcal{A}(i)} exp(z_i \cdot z_k / \tau)}, \tag{7}$$

where $\mathcal{P}(i) = \{j \in \mathcal{I} : y_j = y_i, s_j \neq s_i\}$, represents the set of indexes of examples with the same label and different demographic attribute s_j as z_i. Conversely, $\mathcal{A}(i) = \{k \in \mathcal{I} : y_k \neq y_i, s_k = s_i\}$, represents the set of indexes of examples with the different label and same demographic attribute as z_i. τ is the temperature parameter of contrastive learning.

On the other hand, to retain the accuracy of ICL, we introduce the loss $\mathcal{L}_{acc}$ based on cross-entropy loss. Following the convention, we define the $\mathcal{L}_{acc}$ as:

$$\mathcal{L}_{acc} = \frac{1}{N} \sum_{i \in \mathcal{I}} -log \frac{exp(p_i)}{\sum_{y \in \mathcal{Y}} exp(p_i^y)}, \tag{8}$$

where p_i is the probability that z_i is predicted to be the ground-truth label, p_i^y is the probability that z_i is predicted to be the label y, and label set $\mathcal{Y} = \{joy, anger, sadness, fear\}$. Finally, we obtain bias-aware embedding by minimizing the weighted sum of the above two objectives: $\mathcal{L}_{total} = \alpha \mathcal{L}_{acc} + (1 - \alpha)\mathcal{L}_{bias}$, where α is the parameter that balances the accuracy and fairness. As shown in Figure 4, the total loss $\mathcal{L}_{total}$ is used to optimize the bias-aware embedding via back-propagation.

Experimental Results with ReBE

To validate the few-shot performance of ReBE, we conduct debiasing experiments on a training set of 400 samples, split from the *EEC*-paraphrase. We select two LLMs with the largest AvgGF in each baseline in Table 1 as the objects for eliminating gender bias. Due to hardware limitations, we exclude `OPT-30b` and `Llama-2-70b` from the choices. ReBE is implemented based on the Huggingface PEFT library and previous work (Nguyen and Wong 2023). Experimental results of race bias are available in the Supplementary Material.

Effectivity of ReBE

As shown by the blue subscripts in Table 2, the average gender bias of most LLMs decreases after debiasing by ReBE, which works for all example selection baselines. Concerning the issue that example selection may amplify the maximum bias value, the "Max" row in Table 2 shows a significant reduction in maximum bias. In addition, Figure 5 more intuitively shows the changes in accuracy, AvgGF, MaxTG and MaxFG of `GPT-neo-2.7B` before and after debiasing. The variances of the three biases all decrease, resulting in a more concentrated distribution, indicating improved stability of the bias. In addtion, according to Table 2, the sentiment classification accuracy of LLMs is not significantly affected

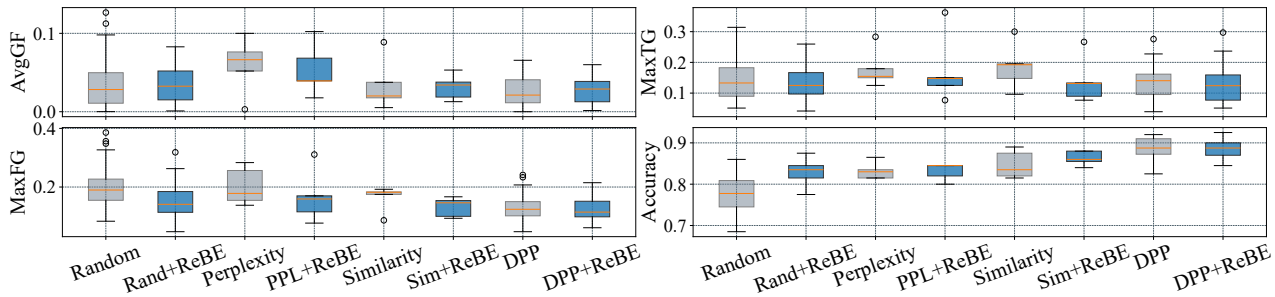

Figure 5: The accuracy and gender bias comparison of GPT-neo-2.7B under example selection methods with and w/o ReBE.

after using ReBE. The above experimental results demonstrate that **ReBE meets the requirement of reducing bias without significantly compromising the accuracy**.

More importantly, the results in Table 2 and Figure 5 demonstrate that **ReBE is compatible with existing example selection methods**. By combining example selection with ReBE, it is possible to achieve high accuracy and low bias of LLMs.

Baseline Comparison

Regarding baseline selection, although (Hu, Liu, and Du 2024) proposed Fairness via Clustering Genetic (FCG) algorithm, it cannot be applied to sentiment analysis or toxicity detection because it requires explicit feature vectors for clustering. Since there are no other debiasing methods specifically for ICL, we compare ReBE with two context augmentation methods: **counterfactual context** and **gender-balanced context**. Compared with the counterfactual context (CF) and gender-balance (Balance) context method, ReBE is compatible with existing example selection methods and can achieve lower bias and higher accuracy.

Ablation Study

To further demonstrate that the reduction in bias results from the $\mathcal{L}_{bias}$ rather than improved accuracy, we conduct ablation studies using the $\mathcal{L}_{acc}$ and $\mathcal{L}_{bias}$ to replace the $\mathcal{L}_{total}$ to train `GPT-J-6B`, respectively. As shown in Table 3, the maximum values of AvgGF and MaxTG of $\mathcal{L}_{acc}$ are much higher than those of ReBE, even though the accuracy is slightly improved. In contrast, $\mathcal{L}_{bias}$ achieves lower bias but sacrifices accuracy. Therefore, $\mathcal{L}_{bias}$ is actually responsible for bias reduction, and $\mathcal{L}_{acc}$ guarantees accuracy.

Parameter Analysis

To illustrate the influence of parameters on ReBE, we conduct the following parameter analysis.

k-shot Since the coverage of examples affects the accuracy of ICL (Gupta, Gardner, and Singh 2023), the number of examples in prompt of ICL k should be large enough. However, redundant information caused by excessive examples may decline the performance of ICL. As shown in Figure 6, the accuracy of LLMs after debiasing increases with the rise in k, while the biases tend to decrease initially and then increase. Therefore, considering accuracy and biases, we chose $k = 18$ as our experiment setting. We provide the k-shot experimental results of other LLMs in the Supplementary Material.

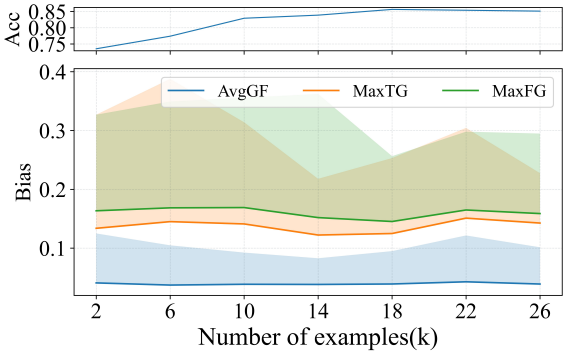

Figure 6: The accuracy and gender bias of GPT-J-6B using ReBE under different k-shot.

n-virtual We ues n-virtual to represent the number of virtual prompt tokens which decides the size of trainable parameters. ReBE needs enough parameters to correct LLMs' biases, but large n-virtual takes up more prompt space. We collect the accuracy and bias results of `GPT-J-6B` using ReBE under different n-virtual, and find there is no apparent relationship between n-virtual and bias. Due to space limitations, we provide the corresponding experimental results in the Supplementary Material.

Conclusion

In this study, we have investigated the impact of example selection on the biases of LLMs. By comparing biases under four example selection baselines with biases under zero-shot, we have found that example selection for ICL amplifies the biases of LLMs. To mitigate the bias of example selection, we have proposed the *Remind with Bias-aware Embedding* (ReBE), which removes the spurious correlations by contrastive learning and retains the feasibility of ICL by prompt tuning. After extensive experiments, we have demonstrated that ReBE can mitigate the bias without significantly compromising accuracy and is compatible with existing example selection methods.

Acknowledgments

This work was supported in part by Major 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

Adams, C.; Daniel, B.; Inversion; Jeffrey, S.; Lucas, D.; Lucy, V.; and Nithum. 2019. Jigsaw Unintended Bias in Toxicity Classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Accessed: 2025-11-16.

Albuquerque, I.; Schrouff, J.; Warde-Farley, D.; Cemgil, A. T.; Gowal, S.; and Wiles, O. 2024. Evaluating Model Bias Requires Characterizing its Mistakes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 938–954. PMLR.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Chang, T.-Y.; and Jia, R. 2023. Data Curation Alone Can Stabilize In-context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8123–8144.

Chen, L.; Chen, J.; Huang, H.; and Cheng, M. 2023a. PTP: Boosting Stability and Performance of Prompt Tuning with Perturbation-Based Regularizer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13512–13525. Singapore: Association for Computational Linguistics.

Chen, Q.; Zhang, R.; Zheng, Y.; and Mao, Y. 2022. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv:2201.08702*.

Chen, Y.; Zhao, C.; Yu, Z.; McKeown, K.; and He, H. 2023b. On the Relation between Sensitivity and Accuracy in In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 155–167. Singapore: Association for Computational Linguistics.

Cui, G.; Hu, S.; Ding, N.; Huang, L.; and Liu, Z. 2022. Prototypical Verbalizer for Prompt-based Few-shot Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7014–7024. Dublin, Ireland: Association for Computational Linguistics.

Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9: 1249–1267.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 1–83.

Gonen, H.; Iyer, S.; Blevins, T.; Smith, N.; and Zettlemoyer, L. 2023. Demystifying Prompts in Language Models via Perplexity Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10136–10148.

Gu, Y.; Han, X.; Liu, Z.; and Huang, M. 2022. PPT: Pre-trained Prompt Tuning for Few-shot Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8410–8423. Dublin, Ireland: Association for Computational Linguistics.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1012–1023. Dublin, Ireland: Association for Computational Linguistics.

Gupta, S.; Gardner, M.; and Singh, S. 2023. Coverage-based Example Selection for In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13924–13950.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Hu, J.; Liu, W.; and Du, M. 2024. Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7460–7475. Miami, Florida, USA: Association for Computational Linguistics.

Iter, D.; Pryzant, R.; Xu, R.; Wang, S.; Liu, Y.; Xu, Y.; and Zhu, C. 2023. In-Context Demonstration Selection with Cross Entropy Difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1150–1162.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, 18661–18673. Curran Associates, Inc.

Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53. Association for Computational Linguistics.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059.

Li, X.; and Qiu, X. 2023. Finding Support Examples for In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6219–6235.

Lin, R.; and Ng, H. T. 2023. Mind the Biases: Quantifying Cognitive Biases in Language Model Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5269–5281. Toronto, Canada: Association for Computational Linguistics.

Liu, H.; Liu, J.; Huang, S.; Zhan, Y.; Sun, H.; Deng, W.; Wei, F.; and Zhang, Q. 2024a. Se^2: Sequential Example Selection for In-Context Learning. In *Findings of the Association for Computational Linguistics ACL 2024*, 5262–5284. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. A. 2022a. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. In *Advances in Neural Information Processing Systems*, volume 35, 1950–1965. Curran Associates, Inc.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2022b. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 100–114.

Liu, Y.; Liu, Y.; Chen, X.; Chen, P.-Y.; Zan, D.; Kan, M.-Y.; and Ho, T.-Y. 2024b. The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Language Models. In *The Twelfth International Conference on Learning Representations*.

Ma, H.; Zhang, C.; Bian, Y.; Liu, L.; Zhang, Z.; Zhao, P.; Zhang, S.; Fu, H.; Hu, Q.; and Wu, B. 2023. Fairness-guided Few-shot Prompting for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 43136–43155. Curran Associates, Inc.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Mosbach, M.; Pimentel, T.; Ravfogel, S.; Klakow, D.; and Elazar, Y. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 12284–12314.

Nguyen, T.; and Wong, E. 2023. In-context example selection with influences. *arXiv:2302.11042*.

Pecher, B.; Cegin, J.; Belanec, R.; Simko, J.; Srba, I.; and Bielikova, M. 2024. Fighting Randomness with Randomness: Mitigating Optimisation Instability of Fine-Tuning using Delayed Ensemble and Noisy Interpolation. *arXiv:2406.12471*.

Plaza-del Arco, F.; Curry, A.; Cercas Curry, A.; Abercrombie, G.; and Hovy, D. 2024. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7682–7696. Bangkok, Thailand: Association for Computational Linguistics.

Wang, X.; Zhu, W.; Saxon, M.; Steyvers, M.; and Wang, W. Y. 2023. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. In *Advances in Neural Information Processing Systems*, volume 36, 15614–15638.

Wu, C.; Wu, F.; Qi, T.; and Huang, Y. 2022. NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 680–685. Dublin, Ireland: Association for Computational Linguistics.

Yang, Z.; Zhang, Y.; Sui, D.; Liu, C.; Zhao, J.; and Liu, K. 2023. Representative Demonstration Selection for In-Context Learning with Two-Stage Determinantal Point Process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5443–5456.

Ye, J.; Wu, Z.; Feng, J.; Yu, T.; and Kong, L. 2023. Compositional Exemplars for In-context Learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 39818–39833.

Zhang, Y.; Feng, S.; and Tan, C. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9134–9148.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 12697–12706. PMLR.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv:1909.08593*.