Prompts De-Biasing Augmentation to Mitigate Gender Stereotypes in Large Language Models

Jinyuan Chen, Sebastian Binnewies and Bela Stantic

February 21, 2025

Prompts De-Biasing Augmentation to Mitigate Gender Stereotypes in Large Language Models

Bob Jinyuan Chen[1], Sebastian Binnewies[1], and Bela Stantic[1]

[1]School of Information and Communication Technology
jinyuan.chen@griffithuni.edu.au, s.binnewies@griffith.edu.au,
b.stantic@griffith.edu.au

Abstract. Large Language Models (LLMs) have manifested impressive ability in the natural language processing (NLP) area, especially in the power of generating and understanding human languages. However, the training of LLMs is a double-edged sword; on the one hand, LLMs gain the ability to understand and generate text training from the human context, but on the other, they also inevitably inherit the negative, stereotyped, and biased semantics in the context. Therefore, how to mitigate bias and stereotypes in generative LLMs is important to build a healthy, ethical, and fair environment for use in real-world scenarios. Previous studies have proposed strategies for fine-tuning models to mitigate gender stereotypes. Unfortunately, labelling and generating high-quality, debiased data for fine-tuning is a costly process. Although Counterfactual Data Augmentation (CDA) and sentence templates provide low-cost possibilities, they may also introduce new biases. In this work, we introduce a new method to augment neutral sentences for fine-tuning LLMs to mitigate gender stereotypes, named Prompts De-Biasing Augmentation (PDA). Compared with the reversal attributed words in sentences augmented by CDA, the data augmented by PDA proposed in this work to fine-tune LLMs can more effectively reduce gender stereotypes while maintaining the generative ability of the pre-trained model. In addition, this work also proposed three metrics to quantify gender inclusiveness in an unlabelled gender stereotype benchmark. The experimental results show that the neutral sentences augmented by PDA have a better debiasing performance for the parameter-efficient fine-tuning (PEFT) in six evaluation metrics under three test benchmarks rather than the gender reversal sentence augmented by CDA.

Keywords: Gender Bias & Stereotype · Large Language Models · De-Biasing Augmentation · LoRA & QLoRA fine-tuning · Fairness of LLMs

1 Introduction

Motivation: Pre-trained large language models (LLMs) have fundamentally reshaped the field of natural language processing (NLP) [10]. With rapidly developing hardware technology and various innovative designs such as transformer architecture and self-attention mechanisms, LLMs are getting more intelligent in understanding the rules and characteristics of language. However, the powerful language learning capabilities of LLMs gained from the pre-trained stage, trained from the enormous scale of uncurated human-historical data, may also potentially internalize harmful content in models. Negative language features such as prejudice, discrimination, misunderstanding, and stereotypes in human sentences may be unconsciously inherited [5,18].

Before the development of pre-trained LLMs, there were already some studies discussing gender bias and stereotypes in the field of natural language processing (NLP) [1]. For example, some research found that bias does exist in word embedding [2]. Then, the bias issue was propagated into the encoder and decoder structures as transformer-based pre-trained models were developed. This has led to several studies dedicated to uncovering and measuring biases and stereotypes in LLMs, such as benchmark datasets, StereoSet [16], CrowS-Pairs [17] and BOLD [4] and so on. In addition, some de-biasing techniques such as iterative nullspace projection (INLP) [17], Counterfactual Data Augmentation (CDA) [30] and Auto-debiasing [6], were proven effective.

However, traditional methods struggle to leverage benchmarks without explicit gender bias labels fully. In addition, Blodgett et al. [1] also emphasized the need to clarify how and to whom bias is harmful, as reversing gender roles in stereotyped sentences may shift bias rather than remove it. Specifically, CDA-enhanced fine-tuning can inadvertently introduce new biases. For instance, replacing “male-engineer” with “female-engineer” might reduce the stereotype of the males but overcorrect, making “female” overly associated with “engineer” anti-stereotype becomes another form of bias, creating a bias for females from the generation of models [24,25].

Contributions: Therefore, this research aims to propose a method for enhancing the inclusiveness of de-biased data specifically for fine-tuning LLMs. This approach is designed to make biased data more inclusive while avoiding introducing new biases while fine-tuning. Additionally, this work also proposes three metrics to quantify the model’s inclusiveness for gender on an unlabelled benchmark.

Overall, the second section provides an overview of recent works focused on mitigating stereotypes and biases. Following this, the methodology section presents three experiments designed to validate our two primary contributions. The experimental results and related discussions are provided in the fourth section, followed by future directions and limitations for this research.

2 Related Work

2.1 Bias in LLMs

The issue of fairness in large language models (LLMs) poses significant challenges to their broader application. Recent research has demonstrated that biases and stereotypes are prevalent in many downstream tasks involving transformer-based pre-trained models, such as feature extraction, binary classification tasks [28], sentiment analysis [14], and generative tasks [12]. Navigli [18], Gallegos [5], had noted that large-scale pre-trained LLMs are typically trained on tens of billions of text sequences sourced from the internet, books, and media, which inherently contain biases and stereotypes related to gender, race, culture, and other societal attributes. For example, gender stereotypes present in historical data, such as associating certain roles or attributes with specific genders, are likely to be embedded in the training data. Consequently, LLMs may inherit these biases, unconsciously reproducing these characteristics during generation and perpetuating biased outputs.

2.2 Mitigating bias in LLMs

Therefore, mitigating bias in LLMs has consistently been at the core of fairness research. Several studies have investigated various strategies for mitigating bias and have reviewed approaches for detecting stereotypical biases [12,5]. In this section, we categorize de-biasing strategies into two major approaches: model-centric and data-centric.

Model-centric methods focus on adjusting the structure or training process of the model to minimize its reliance on bias when pre-trained with biased data. One prominent model-centric approach is the fine-tuning strategy, which involves adjusting model parameters to alter the distribution of each token in the logits. This adjustment improves next-token prediction, contributing to more equitable and less biased generated content. Several parameter-efficient fine-tuning (PEFT) methods have also been developed to lower fine-tuning costs. Among these, approaches based on low-rank adapters have demonstrated effectiveness in de-biasing tasks [20].

Furthermore, prefix-tuning and prompt-tuning have been utilized for de-biasing by incorporating pre-defined prefixes during training [27]. In addition to these methods, other model-centric de-biasing strategies include regularization-based de-biasing [19], adversarial training to mitigate bias [11], and de-biasing by additional auxiliary classifiers [13].

With one prevailing **data-centric** method being the augmentation of training datasets with counterfactual attributes, counterfactual data augmentation (CDA) weakens the impact of biased attributes by inverting the attribute words of biased sentences. For instance, Ranaldi et al. [22] implemented Counterfactual Data Augmentation (CDA) using the adapter [7] training on the PANDA dataset [21]. Similarly, Prakash et al. [20] also mentioned using the PANDA and counterfactual data augmentation (CDA) method to explore the de-biasing of each layer within the model in the proposed Layered Bias method. In addition, CDA is also widely used in various methods to solve the problem of data scarcity and enhance de-biased data [30,17].

However, despite its effectiveness, CDA is limited in some situations. Traditional dictionary-based CDA can result in unnatural sentence structures and lack generalization capabilities, which may even introduce new biases [24] [25] [1]. Muli et al. [15] found that CDA may not achieve the expected counterfactual effects, thereby affecting the model's performance. Topo et al. [26] and Huang et al. [8] also discovered similar problems in their respective experiments and proposed solutions. Therefore, the work elaborates on three research questions below:

- How to create inclusive data for fine-tuning without introducing new biases?
- What's the difference in efficiency for LoRA and QLoRA in de-biasing tasks?
- How to evaluate gender inclusivity on an unlabelled benchmark?

3 Methodology

The work harnesses gender-stereotypical sentences from Winobias [29] as one of the datasets. We applied two strategies to counteract the stereotypical sentences. We first enhanced training data by the Counterfactual Data Augmentation (CDA) [30], which creates anti-stereotypical versions of the biased sentences in Winobias. Then, the training data is augmented using the proposed 'Prompts De-Biasing Augmentation (PDA)' method to create the inclusive text neutrally. There is an example that compares how the biased sentence is de-biased enhanced by CDA and PDA in an individual.

Bias: "[The nurse] argued with the doctor because [she] is angry with solutions."

CDA: "[The nurse] argued with the doctor because [he] is angry with solutions."

PDA: "[The nurse] argued with the doctor because [we] are angry with solutions,
while [nurse] is the [career] for [anyone], [regardless of] gender."

Then, data augmented by CDA and PDA are used for fine-tuning the Mistral 7B model using LoRA [7] and QLoRA [3]. The resulting fine-tuned models are evaluated using stereotype score (SS), language modelling score (LMS), and idealized context association test (ICAT) [16], as well as our proposed metrics: neutral-log-likelihood comparison score (NLCS), neutral dominance frequency (NDF), and composite fairness score (CFS). These evaluations are conducted on benchmark datasets: Crows-Pairs [17], StereoSet [16], and Winogender [23]. The results are analyzed and discussed in Section 4. The whole process is shown in Figure 1. The colors and dashed lines inherit the corresponding tasks and continuous workflows, respectively. The two data augmentation methods (PDA, CDA) are trained by LoRA and QLoRA to obtain five fine-tuned models (including the original model), which are tested by six indicators under three benchmarks (the proposed inclusive metrics are highlighted in red). Three steps demonstrate detailed process about the methodology flow:

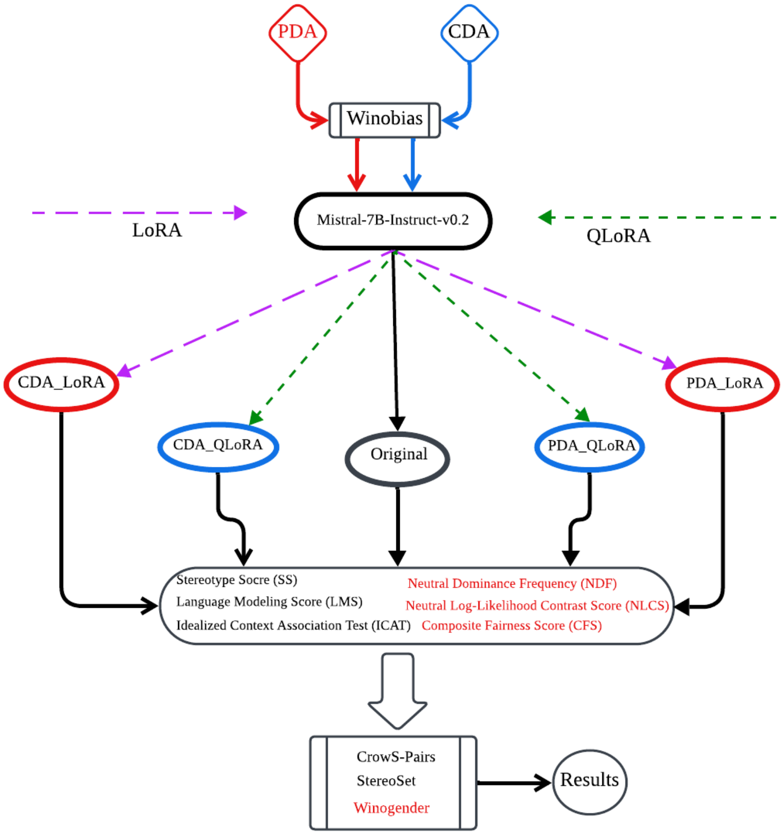
PDA
CDA
Winobias
LoRA
Mistral-7B-Instruct-v0.2
QLoRA
CDA_LoRA
CDA_QLoRA
Original
PDA_QLoRA
PDA_LoRA
Stereotype Socre (SS)
Language Modeling Score (LMS)
Idealized Context Association Test (ICAT)
Neutral Dominance Frequency (NDF)
Neutral Log-Likelihood Contrast Score (NLCS)
Composite Fairness Score (CFS)
CrowS-Pairs
StereoSet
Winogender
Results

Fig. 1: The Experiment Flow of Proposed Methodology

- This work utilizes 3,168 sentences (including 1,584 gender-stereotyped pairs) from the Winobias benchmark [29] to explore the use of predefined neutral prompts for transforming gender-biased sentences into neutral versions. It aims to develop a cost-effective method for obtaining fine-tuning de-biased data while minimizing the risk of introducing new biases.

- Using Mistral-7B-Instruct-v0.2 as the base model, we applied low-rank adaptation (LoRA) [7] and Quantized LoRA [3] fine-tuning on the data enhanced by Counterfactual Data Augmentation (CDA) and Prompts De-Biasing Augmentation (PDA). We aim to determine whether the neutral sentences augmented by PDA outperform the counterfactual sentences enhanced by CDA in terms of de-biasing effectiveness while also comparing the efficiency of de-biasing through LoRA and QLoRA fine-tuning.

- To evaluate the performance of models, we employed various quantitative implicit bias metrics on various fine-tuned Mistral-7B: the original model, the CDA-enhanced fine-tuned model, and the PDA-enhanced fine-tuned model. Both QLoRA and LoRA were used for fine-tuning to explore the efficiency of lower-cost parameter-efficient methods. Additionally, this work introduces three metrics to evaluate models' de-biasing on an unlabelled benchmark, which aims to measure the gender inclusivity and fairness of models.

3.1 Prompts De-biasing Augmentation

The work proposes Prompts De-Biasing Augmentation (PDA) to mitigate gender bias in LLMs. PDA mainly enhances the neutrality of gender stereotypes by replacing gender words with neutral pronouns and phrases that semantically emphasize the neglect of gender. The Winobias dataset contains 1,584 pairs of sentences, each with gender stereotypes and corresponding anti-stereotypes. CDA works by reversing gender descriptions within stereotype sentences to form counter-stereotypical versions and replacing occupation-related terms to ensure that gender-biased and balanced sentences represent each occupation.

However, considering that CDA not only introduces new biases but also has the limitation of failing to work in unlabelled gender-biased texts, the PDA method was born. The design of the PDA method involves several complex modifications to the original stereotype texts. It targets texts or sentences that contain gender references but lack explicit labels. These modifications transform the occupational descriptions and gender characteristics into neutral and inclusive formulations. This process results in a batch of balanced and unbiased sentences. The overall process of PDA's logic is given in Algorithm 1, and here are four processes as follows to describe Algorithm 1 in more detail:

Gender pronoun replacement: Firstly, we used spaCy to perform a partial part-of-speech analysis on each sentence, paying attention to the use of pronouns (such as “he”, “she”, “him”, etc.) and randomly selecting constructed gender-neutral pronouns for replacement, such as “they”, “people”, “someone”, “we”. Meanwhile, PDA considered whether the verb form immediately following the pronoun needs to be modified according to the replaced pronoun (such as changing “is” to “are”) to ensure grammar consistency. In addition, replace the object case of the word with the corresponding neutral object pronoun, such as “him”. The replacement specifications are as follows: “he”: [“they”, “people”, “someone”, “we”], “she”: [“they”, “people”, “someone”, “we”], “him”: [“them”, “people”, “someone”, “us”], “his”: [“their”, “people's”, “someone's”, “our”]. Furthermore, “her” will be judged to possessive or object by “en_core_web_sm” from spaCy and then “her” will be given different replacement words “her”: [“their”, “people's”, “someone's”, “our”] or “her”: [“them”, “people”, “someone”, “us”].

Introducing gender-neutral occupational terms: Secondly, We designed a set of gender-neutral occupational terms, including words like “career”, “job”, “role”, etc., to replace potentially gender-biased occupational names in the biased sentence. When each sentence is enhanced to be neutral, using these neutral terms makes the entire sentence semantically gender-neutral. (e.g. “profession”, “job”, “work”, “role”, “position”, “occupation”, “career”, “employment”, “task”, “responsibility”, “field”, “duty”, “assignment”, “service”).

Algorithm 1: Prompts De-Biasing Augmentation (PDA) Logic

```
Input: Input:Stereotype sentences from Winobias dataset
Output: Balenced sentences by neutral Prompts De-Biasing
        Augmentation (PDA)
Define lists of neutral professions, neutrality expressions, and pronoun
  replacements and fixed prompts at the end of sentence foreach
  sentence in dataset do
    Find bracketed terms using regular expression if fewer than two
      bracketed terms found then
        Return original sentence
    else
        Extract first and second bracketed terms if second term is
          "her" then
            Determine if "her" is possessive or objective using POS
              tagging Replace with appropriate neutral pronoun (e.g.,
              "their", "them")
        else
            foreach gendered pronoun in list do
                if second term matches pronoun then
                    Replace with neutral pronouns if next word is "is" or
                      "was" then
                        Replace with singular neutral replacement (e.g.
                          "someone", "person")
        Replace the second term in the sentence
        if replacement is plural then
            Adjust verb if it's in third-person singular form
        Choose random neutral profession, neutrality expression, and
          connecting clause Append the modified rules to the original
          biased sentences
Saved & Output modified dataset to CSV
```

Use conjunctions and neutral expressions: PDA also includes a range of neutral expressions at the end of enhanced sentences to highlight that these roles and tasks are not gender-restricted, such as “regardless of”, “irrespective of”, “without consideration of”, “without regard to”, “independent of”, “without distinction of", "unconcerned with", "not influenced by", and "apart from". Moreover, PDA uses a series of predefined template sentences to combine the enhanced sentences with neutral descriptions to form templates. For example, "While [ ] is the [ ] for anyone, [ ] gender", "It doesn't matter who [ ] is; the [ ] is for anyone, [ ] gender", "Regardless of who [ ] is, the [ ] belongs to everyone, [ ] gender". These expressions are randomly combined through different connectives and sentence structures, which could diversify the semantics of balanced sentences and reduce the impact of solidification when fine-tuning the model.

Human Evaluation: Besides, some grammatical errors are inevitable in the generation process, such as plural nouns needing to be used for plural persons. Therefore, we manually checked using Grammarly to ensure they were correct in terms of grammar and semantics. Using the PDA method, 1584 defined biased sentences in the Winobias benchmark were modified into neutral de-biased texts in batches.

3.2 Mitigating Gender Bias By PEFT

Since various de-biasing methods have been proposed, full parameter fine-tuning has always been one of the most direct and effective methods. However, due to the limitations of computing resources and training costs, recent studies have proposed parameter-efficient fine-tuning (PEFT) strategies, such as prompts, adapters, LoRA, etc. They are proposed with the hope of reducing computing and storage costs. They only need to update a small number of parameters of the pre-trained model instead of fine-tuning the entire model parameters, which could significantly reduce the computing resources while maintaining the model's performance.

Low-rank Adaptation (LoRA): This part discusses how to fine-tune the datasets augmented with CDA (Counterfactual Data Augmentation) and PDA (Prompts De-Biasing Augmentation) using two parameter-efficient fine-tuning methods: LoRA and QLoRA. The experiment aims to explore the effectiveness of different data augmentation strategies in mitigating gender bias and the impact of different PEFT methods in terms of efficiency. We used Mistral-7B-Instruct-v0.2 [1] as the base model, which has sound generation and language understanding capabilities. It surpasses the Llama 2 13B – Chat model on both manual and automated benchmarks [9], making it a good base model for exploring gender bias issues. To verify the efficiency of the enhanced dataset under different strategies, we applied two fine-tuning methods: LoRA and QLoRA.

In our implementation of **LoRA** shown in Figure 2, to fine-tune the Mistral-7B model, the pre-trained model weights are denoted as $W_0 \in \mathbb{R}^{d \times d}$. LoRA introduces a low-rank decomposition to effectively adapt the weights with minimal parameter updates. The new weights W_{LoRA} are represented as :

$$W_{\text{LoRA}} = A \times B$$

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

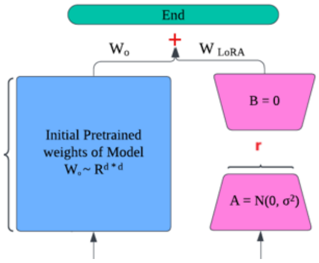

Fig. 2: The structure of LoRA Fine-tune

- $A \sim \mathcal{N}(0, \sigma^2)$ represents a randomly initialized matrix sampled from a normal distribution with variance σ^2. This initialization is crucial for learning effective transformations during adaptation.
- $B = 0$ indicates that the impact of LoRA is initially nullified, meaning the pre-trained weights W_0 are unaffected at the beginning.

The final adapted weights used during training are represented as:

$$W_{\text{adapted}} = W_0 + W_{\text{LoRA}}$$

This approach allows the model to utilize the pre-trained knowledge stored in W_0 while applying efficient and lightweight adjustments via W_{LoRA}. The adjustment process is driven by data enhanced through either PDA or CDA, which aim to reduce gender bias within the model outputs.

Although the principles of QLoRA and LoRA are similar, in QLoRA fine-tuning, we further quantified the model weights based on LoRA and used 4-bit precision for calculation, which significantly reduces memory and computing costs compared with LoRA.

Hyper-parameters and Training Settings: For experimental accuracy, all hyper-parameter configurations and low-rank adaptation settings are the same when using LoRA and QLoRA to fine-tune the texts of CDA and PDA, respectively, as shown below. The process involves loading the pre-trained Mistral-7B-Instruct-v0.2 model and applying the LoRA configuration to fine-tune the model. LoRA/QLoRA enables adaptive fine-tuning by adding low-rank matrices to existing model weights. This technique introduces a low-rank decomposition, allowing efficient parameter updates and minimizing memory overhead.

```
lora_config = LoraConfig(r=8, lora_alpha=32, lora_dropout=0.1,
bias="none", task_type=TaskType.CAUSAL_LM, inference_mode=False)
```

In **QLoRA** fine-tuning, the only difference from LoRA is to use 4-bit NF4 quantization and 8-bit AdamW optimizer for training. It specially loads the model in 4-bit precision and computing in 16-bit, reserving in 4-bit to reduce the loss from quantization. Meanwhile, NF4 is a better choice for the quantify accuracy, which can effectively reduce the losses caused by quantification.

```
bnb_config = BitsAndBytesConfig(
load_in_4bit=True, bnb_4bit_use_double_quant=True,
bnb_4bit_quant_type="nf4",  bnb_4bit_compute_dtype=torch.float16)
```

3.3 Quantify Bias in Fine-tuned Models

To evaluate the effectiveness of fine-tuned models in mitigating gender bias, we used six quantitative metrics [5] on three benchmark datasets. These benchmarks comprise binary or multi-tuple sentences with gender and occupation words. Two of the benchmarks (Stereoset [16], CrowS-Pairs [17]) are labelled biased with certain genders (male/female). We used the Stereotype Score (SS), Language Model Score (LMS), and Idealized Context Association Test (ICAT) proposed by Nadeem et al. [16] to evaluate models in de-biasing.

The other benchmark is Winogender [23], which consists of sentence triplets, such as: "The physician advised the patient after [he/she/they] finished the procedure". Winogender is also not labelled as biased or anti-biased by a certain gender. It is necessary to use unlabelled benchmarks to measure gender bias, which can more realistically reflect the gender inclusiveness of the model. Therefore, based on the Winogender benchmark, we proposed three quantified metrics that reflect the gender inclusiveness of the model:

The Neutral Log-Likelihood Comparison Score **(NLCS)**, Neutral Dominance Frequency **(NDF)**, and Composite Fairness Score **(CFS)** are given by Formulas 1, 2, and 3, respectively. Their overall logic is shown in Algorithm 2.

$$\text{NLCS} = \frac{1}{N_{\text{groups}}} \sum_{i=1}^{N_{\text{groups}}} \left(L_{\text{neutral}}^{(i)} - (L_{\text{male}}^{(i)} + L_{\text{female}}^{(i)}) \right) \tag{1}$$

- N_{groups}: Total number of groups of sentences.
- $L_{\text{neutral}}^{(i)}$: Log-likelihood of the neutral sentence in the i-th group.
- $L_{\text{male}}^{(i)}$ and $L_{\text{female}}^{(i)}$: Log-likelihoods of the gender sentences in the i-th group.

$$\text{NDF} = \frac{N_{\text{neutral\_dominant}}}{N_{\text{groups}}} \tag{2}$$

- $N_{\text{neutral\_dominant}}$: Number of groups where the neutral sentence has the highest log-likelihood compared to male or female sentences.
- N_{groups}: Total number of groups $\left(\frac{\text{count\_sentences}}{3}\right)$.

$$\text{CFS} = \text{NLCS} + \text{NDF} \tag{3}$$

Algorithm 2: Neutrality and Gender Bias Assessment

```
Input: Pre-trained model $f$, Tokenizer $T$, Data $D$
Output: Task Results: $task1$, $task2$, $task3$
Initialize model $f$ and tokenizer $T$;
Load fine-tuned weights into model $f$;
Load dataset $D$ from 'all_sentences.tsv';
Calculate log-likelihood for each set of male, female, and neutral
  sentences using model $f$;
Task 1: Calculate NLCS:
  $task1\_ratio \leftarrow neutral\_score - (male\_score + female\_score)$;
Task 2: Calculate NDF: Increment $task2\_counts$ if
  $neutral\_score > \max(male\_score, female\_score)$;
Task 3: Calculate CFS: $task1\_ratio + (task2\_counts/count\_all)$;
```

4 Results and Discussion

This section presents the results of our experiments and provides a comprehensive comparison of Mistral7B's de-biasing performance on three benchmarks before and after fine-tuning using LoRA and QLoRA with datasets enhanced by PDA and CDA. Table 1 shows the results of quantifying metrics for the original and fine-tuned Mistral-7B Models. SS_set, LMS and ICAT are tested on StereoSet; SS-crows are tested on CrowS-Pairs; NLCS, NDF, and CFS are tested on Winogender, highlighting the best-performing outcomes in bold for reference.

Table 1: Quantifying gender bias in fine-tuned/original models

Model_Mistral7B	SS_set	LMS	ICAT	SS_crows	NLCS	NDF	CFS
Original	58.69	74.31	43.65	67.24	3.76	$\frac{73}{240}$	4.07
CDA_LoRA_ft	59.86	**75.62**	**45.20**	62.59	4.45	$\frac{103}{240}$	4.87
PDA_LoRA_ft	**55.77**	75.35	42.01	**61.20**	**4.93**	$\frac{121}{240}$	**5.44**
CDA_QLoRA_ft	59.54	74.96	44.63	66.04	4.18	$\frac{142}{240}$	4.77
PDA_QLoRA_ft	57.61	74.35	42.83	64.12	4.73	$\mathbf{\frac{171}{240}}$	**5.44**

4.1 PDA vs CDA

The experimental results indicate that data enhanced by neutral Prompts De-Biasing Augmentation (PDA) shows superior performance across both labelled benchmarks (StereoSet and CrowS-Pairs) when using LoRA fine-tuning. Within the StereoSet evaluations, it was observed that models fine-tuned with the data enhanced by Counterfactual Data Augmentation (CDA) introduced additional biases, leading to an increase in SS_set compared to the original model.

Although CDA slightly outperformed PDA in balancing de-biasing and maintaining language coherence, both CDA and PDA achieved comparable top scores in the coherence assessment (LMS). Besides, PDA consistently demonstrated better performance across the three fairness metrics when evaluated on the Winogender benchmark, regardless of the fine-tuning method applied.

Furthermore, when considering the inclusiveness metrics proposed in the work—NLCS, NDF, and CFS. PDA performs better in mitigating bias. These metrics help quantify the inclusiveness of the model's outputs on gender fairness, and PDA exhibited a clear advantage over CDA across the benchmarks. So, PDA is a better choice for enhancing de-biasing in fine-tuning, especially for tasks that prioritize the inclusiveness of models. In summary, regardless of the quality of gender-bias labelling in text data, PDA is a practical choice for enhancing de-biasing data in fine-tuning tasks that do not prioritize balanced semantic coherence. It is especially relevant in scenarios where CDA may introduces additional biases. Using PDA-enhanced data for fine-tuning models reduces biases related to fairness and prevents the introduction of additional unintended biases.

4.2 LoRA vs QLoRA

Besides, We analyzed parameter-efficient fine-tuning (PEFT) methods—specifically, LoRA and QLoRA—and compared them with full fine-tuning. Notably, PEFT update only 13 MB of parameters, while full fine-tuning updates 27,625 MB, significantly reducing computational costs. In our experiments, LoRA achieved higher accuracy, making it more suitable for accuracy-sensitive tasks under resource constraints. Although QLoRA does not outperform LoRA in accuracy—at best, it matches LoRA on the comprehensive inclusiveness metric (CFS)—its training speed is markedly faster, completing four epochs in 11 minutes compared to 80 minutes for LoRA. Therefore, while LoRA remains the preferred choice for accuracy-critical applications, QLoRA demonstrates a clear advantage in speed. In conclusion, for bias mitigation tasks using parameter-efficient fine-tuning, LoRA is more suitable for overall, effective debiasing, whereas QLoRA is better suited for time-sensitive training scenarios.

5 Conclusion and Limitation

Overall, the results and discussion revealed the answers to our research questions **(1):** The results demonstrate that the Prompts De-Biasing Augmentation (PDA) method effectively mitigates gender-stereotypical biases without introducing new biases during fine-tuning, as evidenced by the SS and fairness metrics. **(2):** We also observed a clear distinction in the efficiency of de-biasing methods within the PEFT methods. LoRA generally outperformed QLoRA regarding task accuracy, while QLoRA demonstrated superior training efficiency with significantly reduced resource requirements. **(3):** To assess inclusiveness on unlabelled benchmarks Winogender, we introduced three new metrics: NLCS, NDF, and CFS to measure the gender inclusiveness of the model on the unlabelled benchmark Winogender and obtained consistent trends with other metrics.

However, We found that PDA is less effective than CDA in enhancing semantic coherence in the ICAT metric, which similar neutral templates may cause. Additionally, fine-tuning a 7B model with only 3,000 samples may limit generalizability. To address this, we plan to incorporate more de-biasing techniques and evaluate PDA on different model scales (e.g., 3B, 14B).

We will also explore PDA's adaptability to other stereotypes, such as race, health, and occupation. Technically, PDA will be compared with data-driven de-biasing (e.g., data filtering, contrastive learning) and model-driven methods (e.g., projection-based, RLHF) across various models. In the future, we aim to develop a benchmark only for mitigating implicit bias using the PDA method.

References

1. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050 (2020)
2. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Quantifying and reducing stereotypes in word embeddings. arXiv preprint arXiv:1606.06121 (2016)
3. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms (2023). arXiv preprint arXiv:2305.14314 (2023)
4. Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.W., Gupta, R.: Bold: Dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 862–872 (2021)
5. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational Linguistics (2024)
6. Guo, Y., Yang, Y., Abbasi, A.: 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). pp. 1012–1023 (2022)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
8. Huang, W., Liu, H., Bowman, S.R.: Counterfactually-augmented snli training data does not yield better generalization than unaugmented data. arXiv preprint arXiv:2010.04762 (2020)
9. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
10. Khan, W., Daud, A., Khan, K., Muhammad, S., Haq, R.: Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. Natural Language Processing Journal (2023)

11. Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.: Fairness without demographics through adversarially reweighted learning. Advances in neural information processing systems (2020)
12. Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models. arXiv preprint arXiv:2308.10149 (2023)
13. Liu, H., Jin, W., Karimi, H., Liu, Z., Tang, J.: The authors matter: Understanding and mitigating implicit bias in deep text classification. arXiv preprint arXiv:2105.02778 (2021)
14. Luo, H., Glass, J.: Logic against bias: Textual entailment mitigates stereotypical sentence reasoning. arXiv preprint arXiv:2303.05670 (2023)
15. Mouli, S.C., Zhou, Y., Ribeiro, B.: Bias challenges in counterfactual data augmentation. arXiv preprint arXiv:2209.05104 (2022)
16. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456 (2020)
17. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: Crows-pairs: A challenge dataset for measuring social biases in masked language models. arXiv preprint arXiv:2010.00133 (2020)
18. Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality (2023)
19. Park, S., Choi, K., Yu, H., Ko, Y.: Never too late to learn: Regularizing gender bias in coreference resolution. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. pp. 15–23 (2023)
20. Prakash, N., Lee, R.K.W.: Layered bias: Interpreting bias in pretrained large language models. In: Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP (2023)
21. Qian, R., Ross, C., Fernandes, J., Smith, E., Kiela, D., Williams, A.: Perturbation augmentation for fairer nlp. arXiv preprint arXiv:2205.12586 (2022)
22. Ranaldi, L., Ruzzetti, E.S., Venditti, D., Onorati, D., Zanzotto, F.M.: A trip towards fairness: Bias and de-biasing in large language models. arXiv preprint arXiv:2305.13862 (2023)
23. Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B.: Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301 (2018)
24. Sen, I., Samory, M., Wagner, C., Augenstein, I.: Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. arXiv preprint arXiv:2205.04238 (2022)
25. Tokpo, E.K., Calders, T.: Model-based counterfactual generator for gender bias mitigation. arXiv preprint arXiv:2311.03186 (2023)
26. Tokpo, E.K., Calders, T.: Fairflow: An automated approach to model-based counterfactual data augmentation for nlp. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer (2024)
27. Wang, Y., Demberg, V.: A parameter-efficient multi-objective approach to mitigate stereotypical bias in language models. In: Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP). pp. 1–19 (2024)
28. Zhang, Y., Zhou, F.: Bias mitigation in fine-tuning pre-trained models for enhanced fairness and efficiency. arXiv preprint arXiv:2403.00625 (2024)
29. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876 (2018)
30. Zmigrod, R., Mielke, S.J., Wallach, H., Cotterell, R.: Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. arXiv preprint arXiv:1906.04571 (2019)