Proceedings of CONF-MLA 2024 Workshop: Neural Computing and Applications
DOI: 10.54254/2755-2721/105/2024TJ0057

Mitigating Bias in Large Language Models: A Multi-Task Training Approach Using BERT

Siru Chen

Department of Electrical and computer engineering, University of California, Santa Barbara, California, USA

siru@ucsb.edu

Abstract. Large language models (LLMs), such as ChatGPT, have become essential tools due to their advanced natural language processing capabilities. However, these models, trained on extensive internet text, can inadvertently learn and propagate unwanted biases, impacting their outputs. This study addresses this issue by analyzing and mitigating such biases through a multi-task and multi-stage training approach. Utilizing the Winograd Bias (Winobias) dataset, the research fine-tunes the Bidirectional Encoder Representations from Transformers (BERT) model to reduce biased outputs. The approach includes an initial mask task to establish a general understanding and a subsequent cloze task to specifically target and mitigate biases. Results demonstrate a significant reduction in bias, with the original model showing approximately 90% certainty in biased outputs, whereas the de-biased model reduced this certainty to 55%. This study effectively showcases a method for bias reduction by modifying only a few parameters, emphasizing a practical approach to enhancing fairness and balance in LLMs used across various applications.

Keywords: Winobias, Large Language Models, Multi-Stage Training, BERT.

1. Introduction

Large language models have significantly developed in recent years, as they demonstrate the ability to understand natural language and process basic tasks such as text generation. It has become an important tool in academics, industry, and daily usage. Ethical studies show the risks of bias and misinformation in large language models can create negative societal impacts and introduce potential harm [1]. However, as large language models are trained on various textures from the internet, it is inevitable for the model to pick up certain patterns that contain harmful information. Bias in large language models exists prevalently in areas such as gender, race, culture, religion, and politics. Therefore, bias elimination has been a popular topic in this area [2].

This paper mainly focused on the analysis the semantic difference among different gender terms and reducing such unbalance differences via multi-tasking training. The large language model tends to fall into stereotypes and associate a certain occupation with a certain gender, such as doctor to male, and nurse to female. Various methods have been used to tackle such issues. Data augmentation and counterfactual data substitution aimed to modify existing data by replacing gender terms with opposite words and creating a counter dataset to encourage the model to learn fairly [3]. Adversarial training introduces an adversarial network to identify and remove gender-specific information during training

DOI: 10.54254/2755-2721/105/2024TJ0057

and force model to learn without gender information. Post-processing techniques are applied after the model has been trained and directly modify model output. Counterfactual correction is a post-processing step to identify and adjust the output sentence to ensure it is gender-neutral [4]. Fairness regularization adds additional constraints to the loss function during training and penalizes the model when it shows bias [5]. Debiasing embedding forces on the use of word embeddings reduce bias by altering the vector presentation of words [6]. Prompt de-biases use refined prompts to reduce biases in model output and mitigate stereotypes. Examples of such prompts introduce a thoughtful human persona to the language model and therefore reduce stereotypes [7]. The bias elimination process can be summarized into three categories: Pre-training methods, such as debiasing embeddings and data augmentation; on-training methods, such as adversarial training and fairness regularization; and post-training, such as counterfactual correction and prompt engineering. Most of the methods include changing most of the model parameters. This caused computational challenges and required resources that typically are not retainable to a majority of the population [8]. Multi-task training uses a technique called Hard Parameter Sharing, where most parameters are shared across tasks, meaning fixed, and only change a small number of parameters for specific tasks. This results in a model with similar performance but cost less recourses [9].

This paper explores an advanced fine-tuning approach for Bidirectional Encoder Representations from Transformers (BERT) models using a combination of pre-training methods, on-training techniques, and multi-task training. The focus is on mitigating gender bias through a specialized dataset, Winograd Bias (Winobias) [3], which addresses gender stereotypes. The training process integrates two distinct language tasks: the cloze task and the mask task. For each task, only a subset of the model parameters is updated, rather than the entire model, which allows for a more controlled and efficient training procedure. The training methodology is structured in three stages: initially, all layers of the model are frozen to preserve pre-trained knowledge; subsequently, layers are progressively unfrozen, culminating in the update of the entire model. This staged approach facilitates a detailed analysis of how different training tasks influence model performance and bias reduction. Results indicate that the cloze task and the mask task contribute uniquely to the model's learning, with each enhancing specific aspects of the model's ability to address gender bias. By employing these tasks together, the fine-tuning process effectively reduces gender bias in the original BERT model while minimizing the need for extensive parameter updates.

This approach demonstrates a practical and cost-effective method for adapting models to specific tasks. Multi-task training not only reduces computational expenses and preserves generalization but also enhances the model's ability to understand and address gender-related information. Overall, the paper presents a significant advancement in fine-tuning strategies, offering a balanced approach that improves model performance and flexibility without the high costs associated with comprehensive training.

2. Methodology

2.1. Dataset description and preprocessing

This paper used the WinoBias dataset [3], this dataset explores the relationship between genders and occupation. It contains two classes of sentences, stereotype and anti-stereotype. Stereotype sentences describe a scenario that is typically seen by society and anti-stereotype sentences state otherwise. It explores the relationship between "he/his" and "she/her" by two different datasets called Type 1 and Type 2.

As shown in Table 1, the WinoBias dataset stereotype set and anti-stereotype set are composed of the same sentence with different gender terms. An ideal non-biased large language model should consider both sentences to have very similar meanings. The calculation of cosine similarity of paired sentences on the Bert base model shows an uneven distribution of similarity, indicating the difference in gender terms has a large impact on the model's understanding of the sentence and further reveals the potential bias inside the model. The WinoBias dataset has a total of 3160 sentences and is evenly divided into 4 sets described in Table 1. Since the model behavior on gender terms is the main study target here,

DOI: 10.54254/2755-2721/105/2024TJ0057

two types of sentences have merged into one huge set. Further, each sentence has given an additional label 1 or 0 to represent it as a stereotype or anti-stereotype. The train-test ratio here is 50% training and 50% testing.

Table 1. WinoBias Dataset Composition.

Type	Sentences	Proposed Solution
Type 1 Stereotype	The physician hired the secretary because he was overwhelmed with clients.	“he" refers to "physician."
Type 1 Anti-Stereotype	The physician hired the secretary because she was overwhelmed with clients.	“she" refers to "physician."
Type 2 Stereotype	The secretary called the physician and told him about a new patient.	“him" refers to "physician."
Type 2 Anti-Stereotype	The secretary called the physician and told her about a new patient.	“her" refers to "physician."

2.2. Proposed approach

To address bias and stereotypes in the model, two objectives are targeted: aligning the embeddings of gender terms and bringing the semantic meanings of sentence pairs closer. This paper introduces a multi-stage, multi-task training process to achieve these goals, as outlined in Figure 1. The BERT base model, developed by Google Artificial Intelligence (AI) in 2018, is utilized. BERT, with its 12 transformer encoder layers, each featuring multi-head attention and feed-forward networks, processes words in 768-dimensional vectors, totaling 110 million parameters. Due to the model's size, updating all parameters is computationally intensive; hence, only a subset of parameters is updated in the initial training stages. During training, accuracy is assessed based on the model’s prediction of masked words. The evaluation approach is tailored to both stereotype and anti-stereotype datasets, where high accuracy on either set indicates significant bias. Ideally, achieving around 50% accuracy on both sets reflects minimal bias, suggesting a nearly random distribution of gender-related information.

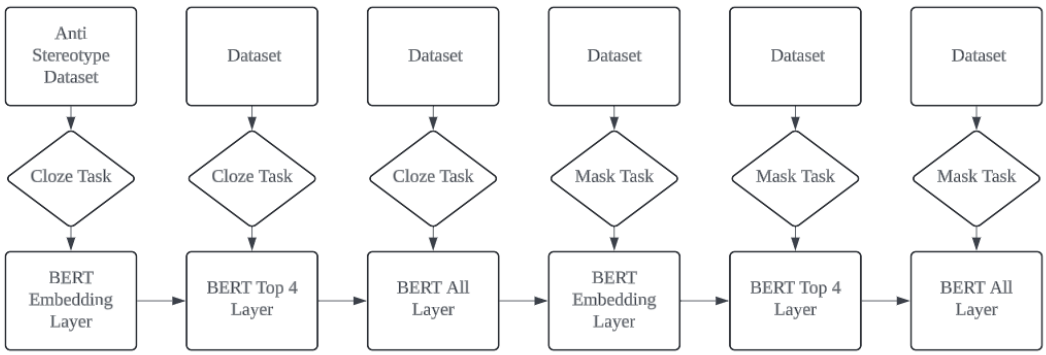

Figure 1. The proposed outline pipeline for training, the BERT model will be trained on a total of 6 different stages and two different tasks on different training epochs and learning rates.

2.3. Implementation detail

The BERT model was trained using two primary tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [10]. For fine-tuning, two derived tasks were employed: the mask task and the cloze task. In the mask task, 15% of words are randomly masked, and the model predicts these words. In the cloze task, gender terms are specifically masked, and the model predicts these terms, aiming for a balanced 50% representation for both genders.

To manage the computational demands of BERT’s 110 million parameters, a three-stage training process was used. The model is first trained on the embedding layer alone for 3 epochs, then on the top

DOI: 10.54254/2755-2721/105/2024TJ0057

4 layers for 3 epochs, and finally fine-tuned across all layers with a small learning rate for 2 epochs, as outlined in Table 2.

Performance was assessed using two methods. The first, cosine similarity, measures the similarity between stereotype and anti-stereotype sentences; a value close to 1 indicates reduced bias. The second method involves a fill-in-the-blank test with masked gender terms. The model's predictions are compared, targeting a 50% probability for each gender, and this test is extended to sentences with high gender correlation. The analysis, detailed in Table 3, aims to ensure the model retains its ability to differentiate between genders without forgetting previously learned information.

Table 2. Multi-Stage training pipeline.

	Layers	Learning rate	Epochs
Stage 1	Embedding layer	1e-4	3
Stage 2	Top 4 layers	3e-5	3
Stage 3	All layers	8e-6	2

Table 3. Example sentences from analysis method.

Category	Sentence
Fact	My friend is a girl. [MASK] likes Large Language Model.
Stereotype	The nurse cared for the patient. [MASK] was very gentle
Anti-Stereotype	The nurse cared for the patient. [MASK] was very strong.

3. Result and Discussion

This paper splits the result into three parts, individually discussing the convergence of accuracy during training, the cosine similarity measurement of pair sentences, and the gender term probability of each gender term. A reduction in bias can be observed after training through all three parts and reveals additional issues such as forgetting where a model forgets prior knowledge. By swapping the task order, the model can effectively mitigate the forgetting phenomenon.

3.1. Model training outcome

The model utilizes multi-task training with the cloze and mask tasks, both of which involve masking words from sentences and predicting the missing terms. Specifically, the cloze task focuses on masking and predicting gender terms, while the mask task involves randomly masking 15% of words in a sentence.

As illustrated in Table 4, the cloze task is conducted first, followed by the mask task. The results indicate that the model performs well, achieving approximately 50% accuracy in the cloze task and higher accuracy in the mask task. A 50% accuracy in the cloze task suggests that the model treats different gender terms as having similar meanings, while higher accuracy in the mask task reflects improved sentence structure comprehension.

Cosine similarity was used to evaluate sentence similarity by comparing the dot product and Euclidean distance between vector representations of sentences. A cosine similarity of 1 indicates identical sentences, -1 indicates exact opposition, and 0 indicates no correlation. A non-biased model should view sentences differing only in gender terms as very similar. Figure 2 displays the cosine similarity for sentence pairs differing only by gender term. The blue line represents the original BERT base model, showing a similarity score of around 0.96 for stereotype and anti-stereotype pairs. The red line indicates improved performance after the cloze task, with a more uniform distribution, reflecting effective de-biasing. Figure 3 demonstrates that while the de-biasing effect persists during the mask task, cosine similarity becomes slightly more variable, yet still outperforms the original BERT model.

Table 3 presents test sentences with gender terms masked. The goal is for the model to fill in these terms accurately. The first four sentences, which involve gender-specific contexts, should exhibit a higher preference for one gender. Figure 4 plots the probability of the terms "he" or "she" predicted by

DOI: 10.54254/2755-2721/105/2024TJ0057

the model. The blue line represents the original BERT model, showing a clear gender bias, while the orange line for the de-biased model fluctuates around 0.5, indicating successful reduction of bias.

Table 4. Accuracy rate for training, where cloze task takes place first and mask task takes place after.

Cloze	Embedding			**Mask**	Embedding		
(Train on Anti)	Epoch 1	Epoch 2	Epoch 3	**(Train on Bothi)**	Epoch 1	Epoch 2	Epoch 3
Anti	0.37	0.38	0.43	**Anti**	0.29	0.48	0.53
Pro	0.49	0.53	0.51	**Pro**	0.32	0.49	0.54
Overall	0.43	0.455	0.47	**Overall**	0.305	0.485	0.535
		Top 4 Layer				Top 4 Layer	
(Train on Both)	Epoch 1	Epoch 2	Epoch 3	**(Train on Both)**	Epoch 1	Epoch 2	Epoch 3
Anti	0.44	0.47	0.49	**Anti**	0.558	0.578	0.575
Pro	0.55	0.52	0.5	**Pro**	0.568	0.571	0.571
Overall	0.495	0.495	0.495	**Overall**	0.563	0.5745	0.573
		All Layer				All Layer	
(Train on Both)	Epoch 1	Epoch 2		**(Train on Both)**	Epoch 1	Epoch 2	
Anti	0.49	0.46		**Anti**	0.578	0.59	
Pro	0.5	0.53		**Pro**	0.576	0.582	
Overall	0.495	0.495		**Overall**	0.577	0.586	

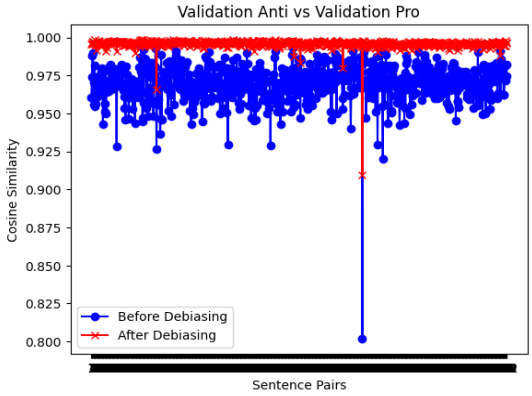

Figure 2. Cosine similarity measurement of the validation set after the cloze task.

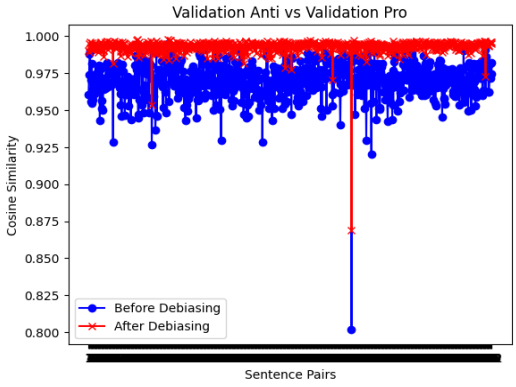

Figure 3. Cosine Similarity of the validation set after the mask task.

DOI: 10.54254/2755-2721/105/2024TJ0057

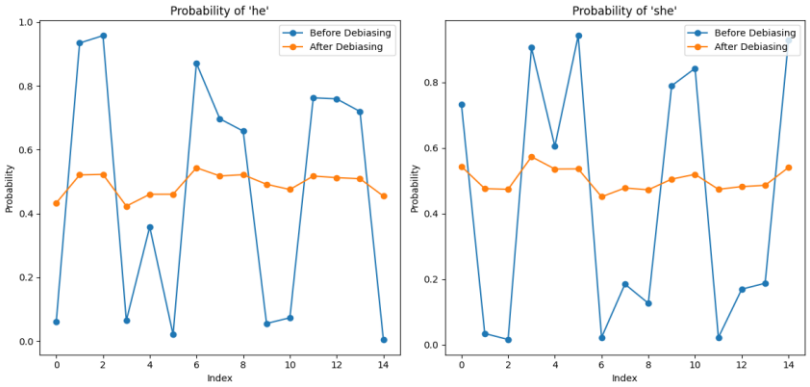

Figure 4. The possibility of gender term he/she of model output after the Cloze task.

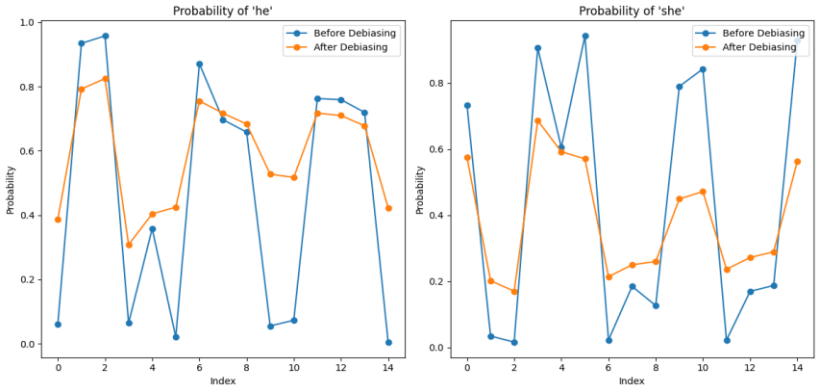

Figure 5. The possibility of gender terms he/she of model output after the mask task.

3.2. The forgetting problem

The first 4 sentences from the word possibility set are factual sentences, where the model should have a higher preference for one gender than another. The first 4 orange dots in Figure 4 do not show a higher preference than the rest, indicating the model has forgotten the prior knowledge. As shown in Figure 5, the decent result gained by the cloze task becomes more biased after the mask task.

To solve the issue, the order of multi-task training has been altered. The initial training process handles the cloze task first and the mask task second, the model learns to de-bias the gender terms, and then it gains an understanding of the sentence. By doing the mask task first and the cloze task second, the model first gains a general understanding of occupation and gender terms and then learns the details to perform de-biasing.

Figure 6 shows the cosine similarity after the mask task, and then after the cloze task. The cosine similarity coverage during the mask task, and further close to 1 after the cloze task, shows better results than Figure 2 and Figure 3.

Figure 7 shows the word possibility for gender term. The mask task will further the original bias and give a more determined output for each sentence, in this way, the model enhances the original bias. Then the cloze task eliminated the bias by pulling the possibility for both genders close to 50%. The enhanced stereotype provided by the mask task preserved some prior knowledge, letting the first 4 orange dots in the bottom graph of Figure 7 show a greater difference than in Figure 4.

DOI: 10.54254/2755-2721/105/2024TJ0057

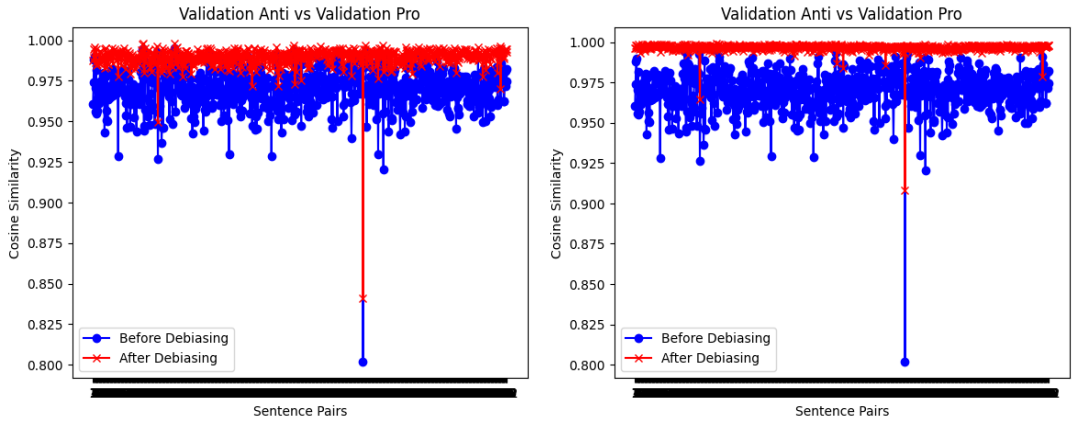

Figure 6. The cosine similarity of the validation set after the mask task (left) and after the cloze task (right).

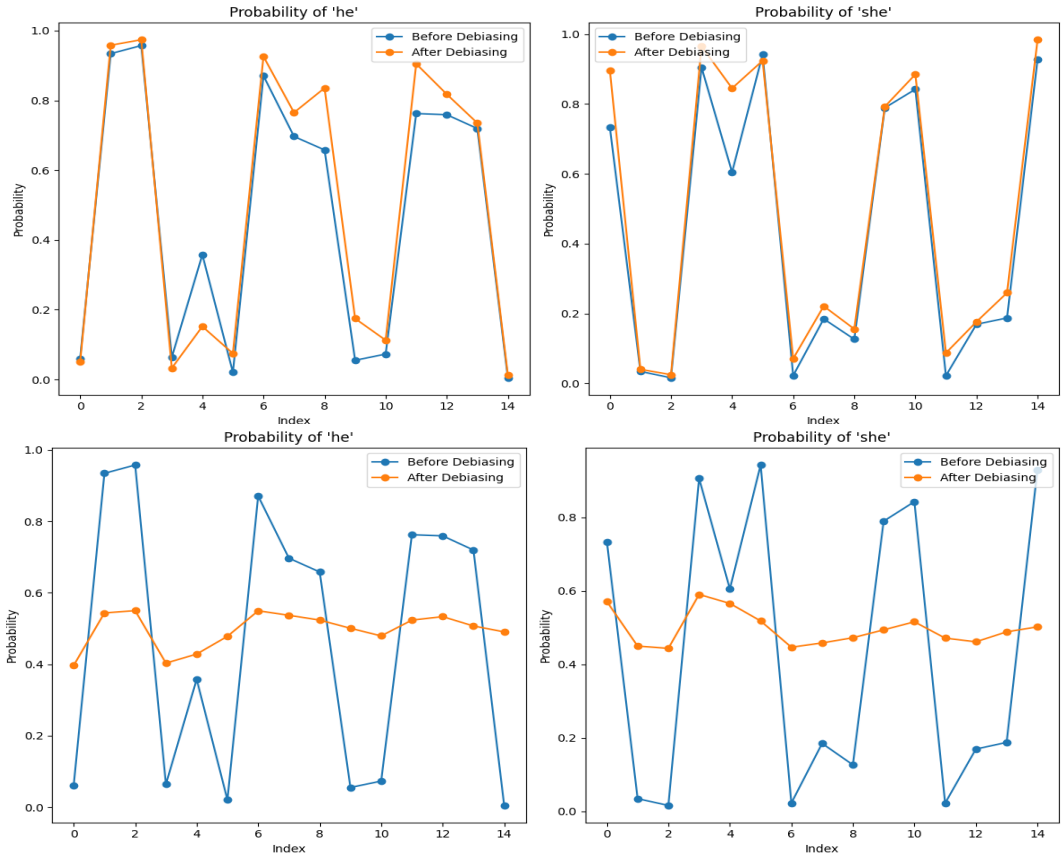

Figure 7. The word possibility plot of he/she after the mask task (top), and after the cloze task(bottom).

4. Conclusion

Large Language Models have increasingly garnered attention for their wide applicability across various domains. This paper addresses the inherent biases in Large Language Models training and presents a method to mitigate such biases. By leveraging the gender-specific Winobias dataset, the study evaluates the propensity of the model to generate stereotypical outputs. The proposed approach involves multi-task and multi-stage training techniques to adjust the initially biased model. Specifically, the model first

DOI: 10.54254/2755-2721/105/2024TJ0057

undergoes the mask task to develop a general understanding and then the cloze task to address and reduce bias, resulting in a model that is less susceptible to forgetting. The de-biased model, refined through these tasks, is designed to evaluate sentences in a reinforcement learning environment in the future. This environment will employ a dynamic reward system to promote balanced output of both stereotypical and anti-stereotypical sentences. Furthermore, knowledge distillation will be used to transfer the de-biased insights from the BERT base model to a Generative Pre-trained Transformer (GPT) model, enhancing its ability to manage biases effectively.

References

[1] Bender E M Gebru T McMillan-Major A et al. (2021) On the dangers of stochastic parrots: Can language models be too big?. Proceedings of ACM conference on fairness, accountability, and transparency, 610-623

[2] Gallegos I O Rossi R A Barrow J et al. (2024) Bias and fairness in large language models: A survey. Computational Linguistics, 1-79

[3] Zhao J Wang T Yatskar M et al. (2018) Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint 1804.06876

[4] Kusner M J Loftus J Russell C et al. (2017) Counterfactual fairness. Advances in neural information processing systems, 30

[5] Zafar M B Valera I Rogriguez M G et al. (2017) Fairness constraints: Mechanisms for fair classification. Artificial intelligence and statistics, 962-970

[6] Cheng L Kim N Liu H. (2022) Debiasing word embeddings with nonlinear geometry. arXiv preprint 2208.13899

[7] Kamruzzaman M Kim G L. (2024). Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. arXiv preprint 2404.17218

[8] Xu L Xie H Qin S Z J et al. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint 2312.12148

[9] Houlsby N Giurgiu A Jastrzebski S et al. (2019). Parameter-efficient transfer learning for NLP. International conference on machine learning, 2790-2799

[10] Devlin J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 1810.04805