Fine-tuning LLMs with Cross-Attention-based Weight Decay for Bias Mitigation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in Natural Language Processing (NLP) tasks but often propagate societal biases from their training data, leading to discriminatory outputs. These biases are amplified by the models' self-attention mechanisms, which disproportionately emphasize biased correlations with sensitive tokens, like "he" or "she", reflecting the sensitive attributes such as gender and race. To address this issue, we propose a novel fine-tuning method, called **Cr**oss-**A**ttention-based **W**eight **D**ecay (**CrAWD**), which modifies the LLM architecture to mitigate bias. CrAWD introduces a cross-attention mechanism between an input sequence and a sensitive token sequence, enabling the model to identify and selectively decay the attention weights of tokens associated with sensitive tokens. This reduces the influence of biased association on the model's generation while maintaining task performance. Evaluations on real-world datasets demonstrate the effectiveness of our proposed CrAWD method. Notably, our method can handle multiple sensitive attributes by adjusting the sensitive token sequence, and it does not require full knowledge of sensitive tokens presented in the dataset, underscoring CrAWD's versatility in promoting fair LLMs across various applications.

1 Introduction

Large language models (LLMs) trained on vast datasets have demonstrated remarkable capabilities on different natural language processing (NLP) tasks, including text generation (Brown et al., 2020) and, text classification (Zhang et al., 2024). However, their widespread usage in real-world settings has also raised concerns about the propagation of societal biases embedded in their training data (Kiritchenko and Mohammad, 2018). These biases can result in discriminatory outputs, affecting downstream applications in significant ways (Bender et al., 2021; Zhao et al., 2017a). Addressing bias in LLMs is essential for building more equitable and responsible AI systems.

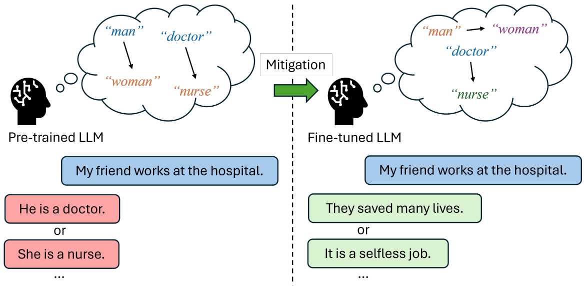


Figure 1: Text generation bias due to biased associations encoded in pre-trained LLMs.

Bias in the NLP tasks is reported as biased or stereotypical associations with some sensitive attributes, such as gender, race, language, religion, etc. (Bolukbasi et al., 2016; Barocas et al., 2019). For LLMs, the bias arises from the training data and is then encoded in the models' internal architectures. Pre-training on large corpora drawn from the web often means the data contains inherent societal biases related to race, gender, and other sensitive attributes (Caliskan et al., 2017). Pre-trained LLMs also exhibit significant bias amplification, meaning that even subtle biases in the training data become more pronounced in the model's output (Bender et al., 2021). The transformer-based models allocate attention weights to different tokens during training using the self-attention mechanisms (Vaswani et al., 2017). When exposed to biased training content, the model can disproportionately focus on biased associations with sensitive tokens, such as gendered pronouns or racially coded words. This leads to biased predictions, especially in tasks like text classification and generation (He et al., 2022; Haque et al., 2024). For example (shown in Figure 1), LLM learns a biased association between gender and occupations from the training data via self-attentions. For the next token generation, "My friend works at the hospital. He is a ___", it generates “doctor” with a high probability due to the high attention weights from “hospital” and “he”, where the association between “he” and “doctor” is biased. A fair LLM should generate tokens based mainly on the relevant context of “hospital”.

Various approaches have been proposed to mitigate bias in LLMs, ranging from data-level interventions to model-level adjustments. Data-level methods include balancing the training data or removing biased examples (Zhao et al., 2017b), while model-level techniques focus on architectural modifications or post-processing steps that reduce the impact of biased outputs (Dong et al., 2024; Dige et al., 2023). For example, Dong et al. (2024) measures bias as explicit mentions of gender pronouns. Their fine-tuning method directly adds a gender probability loss, derived from their bias metric, to the total loss to mitigate bias. Bias mitigation fine-tuning to minimize a specific evaluation metric requires prior knowledge of the evaluation task and sensitive tokens, which lack generalizability across different types of biases in LLMs. There is no single metric to summarize all types of biases in LLMs. These studies focus on bias defined by specific model output, which limits their mitigation methods to generalize to other tasks. Instead of a specific metric on the model output, we believe that the model’s internal self-attention mechanism plays an important role in encoding and perpetuating bias.

To address these challenges, we propose a bias mitigation technique, called **Cr**oss-**A**ttention-based **W**eight **D**ecay (**CrAWD**), which modifies the LLM architecture to reduce the influence of biased association during LLM fine-tuning. Specifically, our approach introduces a cross-attention mechanism between an input sequence and a reference sequence consisting of sensitive tokens, such as "he" and "she". This cross-attention enables the model to identify the input tokens that have strong correlations with sensitive tokens. By selectively decaying the attention weights of these tokens during fine-tuning, we can scale down their biased influence on the model’s predictions without significantly degrading the overall performance of correctness. The contribution of this work is as follows:

- This work addresses bias in LLMs by proposing a novel cross-attention-based bias mitigation technique. Our method effectively reduces the influence of biased associations in the input sequences, enabling the model to learn the contextual information without unintended biases.
- Our method achieves a good balance between bias mitigation and performance preservation. It maintains the self-attention component to capture contextual relationships and employs a decayed cross-attention component to reduce the influence of biased associations.
- This method does not require full knowledge of the potential sensitive tokens in the fine-tuning task. It can work with multiple sensitive attributes simultaneously by adjusting the reference sensitive sequence.
- The evaluation on real-world datasets demonstrates the versatility and effectiveness of our proposed method in a variety of settings.

2 Related Work

Bias in NLP systems has emerged as a significant concern, particularly as language models become increasingly integral to a wide range of applications. Bolukbasi et al. (2016) and Caliskan et al. (2017) demonstrated that societal biases like gender, racial, and cultural stereotypes, which may lead to discriminatory outcomes are embedded in model representations. Large-scale datasets, often sourced from unfiltered content on the internet, are a primary contributor to bias in language models. These datasets frequently reflect societal prejudices, which models inadvertently learn and sometimes amplify during text generation (Sheng et al., 2019; Bender et al., 2021). As Wan et al. (2023) highlighted, such biases do not just persist but may intensify in model outputs, leading to harmful stereotypes being reproduced at scale.

To address this issue, researchers have developed several tools to measure bias in large language models (LLMs). Notable among these are the StereoSet dataset (Nadeem et al., 2021) and the Crows-Pairs dataset (Nangia et al., 2020), both designed to assess the presence of societal bias in generated text. In addition to measuring bias, various mitigation techniques have been explored. One popular approach involves generation-based strategies, such as zero-shot (Gallegos et al., 2024; Liu et al., 2024) and few-shot prompting (Wang et al., 2023; Ko et al., 2023; Ma et al., 2023), where carefully crafted prompts are employed to guide models toward more equitable outputs. Another strategy is the use of Chain of Thought (CoT) reasoning, which has been shown to improve the fairness of model generations by making the model's decision-making process more transparent and deliberate (Tian et al., 2023).

While these approaches have made significant strides in mitigating bias during text generation, there has been limited research on fine-tuning techniques specifically aimed at addressing bias during model training. For instance, Dong et al. (2024) proposed metrics to assess both explicit and implicit gender bias in generated text, incorporating these metrics into the loss function as a regularization technique to reduce bias. Similarly, He et al. (2022) introduced an auxiliary model that predicts protected attributes, using the negative log-likelihood from these predictions as an energy-based constraint to minimize the influence of biased tokens on the output. Contrastive learning has also been employed as a bias mitigation technique, where stereotypical data points are pushed away while non-stereotypical data points are pulled closer in the model's latent space (Zhou et al., 2024).

3 Bias in LLMs from the Attention Perspective

Large Language Models (LLMs) are pre-trained on vast corpora of data, which often contain social biases associated with sensitive attributes such as gender, race, or ethnicity. These biases are embedded in the model during pre-training and can propagate into downstream tasks. We aim to mitigate such bias embedded in the pre-trained LLMs using a fair fine-tuning method.

We use the next token generation task for LLM fine-tuning. Given an input sequence $\boldsymbol{x} = \{t_i\}_{i=1}^{n}$, where t_i denotes the i-th token in the sequence and n is the sequence length, the LLM predicts the next token t_{n+1} by adjusting its parameters θ to maximize the conditional probability $P_\theta(t_{n+1}|t_1, t_2, \ldots, t_n)$. LLMs utilize the self-attention (SA) mechanism to learn a robust and contextual representation of each token (Vaswani et al., 2017). SA mechanism assigns attention weights to all input tokens $\{t_1, t_2, \ldots, t_n\}$ based on their contextual relevance to the target token t_{n+1}. Specifically, the self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (1)$$

where Q, K, and V represent the query, key, and value matrices, and d is the embedding dimension. The attention mechanism computes attention weight α_{ij}, which quantifies how much token t_i contributes to the updated representation of token t_j. In other words, the attention weight α_{ij} determines the degree of influence that token t_i has on token t_j during the representation learning process. A higher α_{ij} means that more weight is given to the information from the token t_i when updating the representation of the token t_j. The hidden state of each token representation is updated as a weighted sum of all tokens' representations in the sequence, including itself.

Unintended biases are embedded into LLMs when a sensitive token t_s (such as "he", "Alice", etc.) related to a sensitive attribute receives disproportionately higher attention weights compared to a non-sensitive token in the task context. The disproportionate association of t_s in the attention mechanism is the root of the bias. It skews the probability distribution on the next token prediction, causing it to prioritize the sensitive token in its output. It leads to biased or stereotypical associations in LLM representation, prediction, and generation tasks, reflected in different evaluation metrics.

Our goal is to mitigate the effect of the sensitive token t_s during fine-tuning, ensuring that the representation learned from the self-attention mechanism does not disproportionately emphasize the biased associations, thus reducing bias in the model's output while maintaining overall performance in downstream tasks.

4 Cross-Attention-based Weight Decay

In this section, we propose a novel fair fine-tuning method, called **Cr**oss-**A**ttention-based **W**eight **D**ecaying (**CrAWD**), to mitigate bias in pre-trained LLMs. The core mechanism in CrAWD is a cross-attention (CA) mechanism between input sequences and reference sensitive tokens. The cross-attention mechanism helps to identify potentially biased associations. Through weight decay, the fine-tuned LLM tries to deemphasize these biased associations. In the end, it learns a debiased contextual representation for downstream tasks.

4.1 Overview

Figure 2 shows the overall architecture of CrAWD. Other than the regular input sequence $\boldsymbol{x}$, CrAWD also takes a reference sensitive sequence $\boldsymbol{b}$ as input. The reference sensitive sequence consists of a series of sensitive tokens that represent various

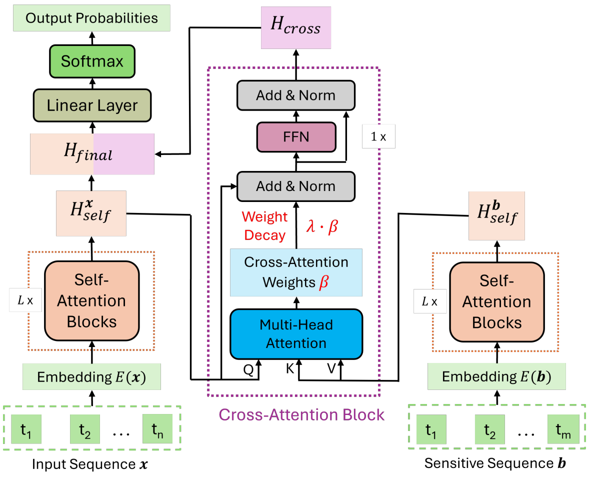


Figure 2: The CrAWD Architecture

sensitive attributes, such as gender, race, religion, ethnicity, or language. The self-attention blocks (SABs) learns the hidden representations $H^{\boldsymbol{x}}_{\text{self}}$ and $H^{\boldsymbol{b}}_{\text{self}}$ for $\boldsymbol{x}$ and $\boldsymbol{b}$, respectively. Then CrAWD introduces a cross-attention block (CAB), which takes in $H^{\boldsymbol{x}}_{\text{self}}$ and $H^{\boldsymbol{b}}_{\text{self}}$ to identify biased associations. Then, CrAWD instructs the model to deemphasize these biased associations by decaying the cross-attention weights. The final representation H_{final} of LLM consists of two parts: the regular SA representation $H^{\boldsymbol{x}}_{\text{self}}$ for the input $\boldsymbol{x}$ to preserve the model performance, and the CA representation after weight decay H_{cross} to reduce the influence of biased associations. In this way, the LLM can generate fair output without over-penalizing the overall performance.

4.2 Reference Sensitive Sequence

To identify biased associations in the pre-trained LLMs, CrAWD takes a pre-defined list of sensitive tokens as references, denoted as reference sensitive sequence $\boldsymbol{b} = \{t_s\}_{s=1}^{m}$, where m is the length of the sequence. The reference sensitive tokens are representative sensitive tokens related to one or multiple sensitive attributes of interest. We do not require $\boldsymbol{b}$ to be a complete list with the full pre-defined knowledge of potential biases in the fine-tuning dataset and downstream tasks. It can be just one reference token for each sensitive attribute group. The cross-attention mechanism only uses $\boldsymbol{b}$ as a reference. In the embedding space, it examines similar tokens related to the sensitive attributes and discovers biased associations with those tokens.

The input embeddings of the input sequence $\boldsymbol{x}$ and the reference sensitive sequence $\boldsymbol{b}$ are $E(\boldsymbol{x})$ and $E(\boldsymbol{b})$, respectively. Each of them is passed through L self-attention (SA) layers, feed-forward networks (FFN), and normalization, to produce self-attended hidden representations $H^{\boldsymbol{x}}_{\text{self}} \in \mathbb{R}^{n\times d}$ and $H^{\boldsymbol{b}}_{\text{self}} \in \mathbb{R}^{m\times d}$. Here, $H^{\boldsymbol{x}}_{\text{self}}$ and $H^{\boldsymbol{b}}_{\text{self}}$ represent the updated last hidden state of the SA block that captures the contextual relationships within the input sequence.

4.3 Cross-Attention Mechanism

After getting the hidden representations $H^{\boldsymbol{x}}_{\text{self}}$ and $H^{\boldsymbol{b}}_{\text{self}}$ from the input sequence and the reference sensitive sequence. CrAWD calculates cross-attention weights from the cross-attention (CA) mechanism, which can be used to identify biased associations in the pre-trained LLMs.

The CA block contains a single cross-attention layer with both $H^{\boldsymbol{x}}_{\text{self}}$ and $H^{\boldsymbol{b}}_{\text{self}}$ as its input. The cross-attention weights are calculated by following Equation 1, where the query, key, and value matrices are computed through linear projections:

$$\begin{aligned} Q &= H^{\boldsymbol{x}}_{\text{self}} W_Q \quad \text{with} \quad W_Q \in \mathbb{R}^{d\times d}, \\ K &= H^{\boldsymbol{b}}_{\text{self}} W_K \quad \text{with} \quad W_K \in \mathbb{R}^{d\times d}, \\ V &= H^{\boldsymbol{b}}_{\text{self}} W_V \quad \text{with} \quad W_V \in \mathbb{R}^{d\times d}. \end{aligned}$$

The cross-attention mechanism enables the model to compute the influence of sensitive attributes on the input sequence $\boldsymbol{x}$ by comparing the query Q (from $\boldsymbol{x}$) with the key K (from $\boldsymbol{b}$), and using the value V (from $\boldsymbol{b}$) to update the hidden state. The hidden representation $H^{\boldsymbol{x}}_{\text{self}}$ of the input sequence $\boldsymbol{x}$ represents the contextual information within the input sequence, which may include biased information with any sensitive token. The hidden representation $H^{\boldsymbol{b}}_{\text{self}}$ of the reference sensitive sequence $\boldsymbol{b}$ represents the sensitive attributes and all potential sensitive tokens related to them (not just the t_s listed in $\boldsymbol{b}$). The cross-attention weights capture the strength of biased associations between the input sequence and the sensitive tokens. Specifically, a higher cross-attention weight β_{is} from the cross-attention matrix represents a higher possibility that there are biased associations between the input token t_i and the sensitive attribute represented by t_s. It can be a direct bias between t_i and another input token in $\boldsymbol{x}$ that is similar to t_s, or an indirect bias between t_i and a sensitive context (without the presence of a direct sensitive token).

4.4 Weight Decay

The cross-attention mechanism captures the potential biased association between the input token t_i and the sensitive attributes. To mitigate bias, we use weight decay to deemphasize (but not completely ignore) these tokens during the fine-tuning.

From the cross-attention matrix, we select the top-$K\%$ of attention weights as these weights reflect the most likely biased associations between the input token t_i and the sensitive attributes. CrAWD decays the attention weights on these tokens by a small factor λ during the fine-tuning, The attention weights for these tokens are adjusted as follows:

$$\beta'_{is} = \begin{cases} \lambda \cdot \beta_{is}, & \text{if } \beta_{is} \geq \tau, \\ \beta_{is}, & \text{otherwise,} \end{cases} \tag{2}$$

where τ is the cutoff threshold at the top-$K\%$ attention weights.

Weight decay on β reduces the impact of biased associations on the final prediction. The fine-tuned model mainly focuses on the other associations actually related to the task context. Thus, the model can make unbiased predictions in a context similar to the fine-tuning task.

After weight decay, the decayed cross-attention weights join the residual connection, followed by a normalization layer. Then the output goes through a position-wise fully connected feed-forward network (FFN) layer and a normalization layer, to compute the cross-attention hidden representation as the output of the CA block:

$$H_{\text{cross}} = \text{FFN}(\text{Cross-Attention}^{\text{decay}}(H^{\boldsymbol{x}}_{\text{self}}, H^{\boldsymbol{b}}_{\text{self}})),$$

where Cross-Attention weights β is modified by following Equation 2, and $H_{\text{cross}} \in \mathbb{R}^{n \times d}$. The resulting cross-attention hidden representation H_{cross} encodes the identified biased associations in LLMs.

4.5 Final Hidden Representation

Once the biased associations are deemphasized through weight decay, the final hidden representation of the input sequence is obtained by combining the self-attention representation $H^{\boldsymbol{x}}_{\text{self}}$ with the decayed cross-attention representation H_{cross}. The final hidden representation H_{final} is computed as:

$$H_{\text{final}} = \text{Concat}(H^{\boldsymbol{x}}_{\text{self}}, H_{\text{cross}}).$$

Here, the self-attention component captures the essential contextual information within the input sequence, while the decayed cross-attention component selectively adjusts the influence of biased associations with sensitive attributes without over-penalization. By integrating these two components, the final hidden representation $H_{\text{final}} \in \mathbb{R}^{n \times 2d}$ incorporates both rich contextual information for the fine-tuned task and mitigation of unintended biases. At the end, the final hidden representation H_{final} is passed through linear and softmax layers to predict the next token t_{n+1} in the sequence.

Our mitigation method also applies to transformer-based text classification models.

5 Experiment Setup

5.1 Research Questions

We aim to answer the following questions with our experiments:

- **RQ1:** Does fine-tuning with CrAWD mitigate bias in different pre-trained LLMs?
- **RQ2:** How does the performance of CrAWD compare to existing fine-tuning techniques for mitigating different biases in text generation or classification?
- **RQ3:** How does identification of biased association (with different top-$K\%$) and weight decay (with different λ) affect the utility-fairness trade-off of CrAWD?

5.2 Datasets

The **Jigsaw Unintended Bias in Toxicity** dataset (cjadams et al., 2019) consists of approximately 2 million public comments annotated for toxicity and the protected attributes of the comment targets. We select 21,000 records for fine-tuning.

The **Bias in Bios** dataset (De-Arteaga et al., 2019) contains textual biographies used to investigate bias in NLP models, featuring 15 different occupations and a balanced gender distribution with binary gender as the sensitive attribute. We randomly select 100k records for fine-tuning.

The **Stanford Natural Language Inference (SNLI)** corpus (Bowman et al., 2015) comprises 570k human-written English sentence pairs, each consisting of a premise and a hypothesis labeled as *entailment*, *contradiction*, or *neutral*. It mentions sensitive attributes like gender, ethnicity, age, etc. We select 100k premises for fine-tuning.

The **Measuring Hate Speech** Corpus (Sachdeva et al., 2022) contains 50,070 social media comments, annotated by 11,143 Amazon Mechanical

Turk contributors using faceted Rasch measurement theory (RMT) to assess hate speech. A subset of 27,818 comments, focused on racial hate speech detection, includes 11,418 hate speech records and 16,400 non-hate speech records. We use it for the evaluation of text classification tasks.

5.3 Baselines

For text generation, we use different pre-trained models for evaluations, including GPT2 (Radford et al., 2019), Llama2 (7B version) (Touvron et al., 2023), Llama3 (8B version)(Dubey et al., 2024), and Falcon (5B version) (Almazrouei et al., 2023), Flan-T5 (3B version) (Chung et al., 2022), DeepSeek (6.7B version) (DeepSeek-AI et al., 2025).

We adopt several fine-tuning methods as baselines. We first compare our model with the **Vanilla Fine-tuning** on pre-trained models. Vanilla Fine-tuned models gain task-specific knowledge from the fine-tuning datasets, but there is no fairness mechanism built in to mitigate bias.

Indirect Bias Mitigation (IBM) (Haque et al., 2024) method uses attention-based explanation to calculate similarity between important tokens in the instance and sensitive information and then uses the similarity as a regularizer in the loss function to mitigate bias.

Debias Tuning (DT) (Dong et al., 2024) explore different metrics to disclose explicit and implicit gender bias in LLMs. They design a regularization term for each metric and add it to the loss function to mitigate bias in fine-tuning.

Deep Soft Debias (DSD) (Rakshit et al., 2025) is a post-hoc debiasing technique that routes token embeddings through a compact residual MLP. A dual objective retains original geometric structure while forcing outputs to lie orthogonal to a hand-crafted bias subspace, thereby suppressing linear demographic bias with negligible task-performance loss.

For text classification, we use the pre-trained BERT-base model (Devlin et al., 2019) as the base model. Other than **Vanilla** and **IBM**, we also adopt the following fine-tuning methods as baselines.

Adversarial Debias (AD) (Zhao et al., 2018) is an in-processing mitigation technique that employs adversarial learning to reduce the correlation between the predicted outcome and the protected attribute, aiming to achieve equality of opportunity.

Controlling Bias Exposure (CBE) (He et al., 2022) is an in-processing mitigation technique that employs an auxiliary model to predict a protected attribute. The negative log-likelihood from this prediction acts as an energy-based constraint, regulating the impact of biased tokens on the output.

5.4 Metrics

Perplexity (PPL) is a metric used to evaluate the quality of text generated by a language model. It measures how well the model predicts a sequence of words by assessing the probability of the model's predictions across a given text. (lower is better).

For text classification, we use **Accuracy** as a utility metric to evaluate the correctness of the classification model

We use the **Idealized Context Association Test (ICAT)** from StereoSet (Nadeem et al., 2021) to evaluate the model's tendency to produce stereotypical associations. ICAT combines a Language Modeling Score (LMS) and a Stereotype Score (SS). A higher ICAT indicates strong modeling with reduced stereotyping.

The **Word Embedding Association Test (WEAT)** (Caliskan et al., 2017) measures bias via cosine similarity of target-attribute word sets (lower scores imply less bias).

The **Gender Attribute Score (GAS)** is a straightforward metric for evaluating gender bias in generated sentences by checking for the presence of gender-specific words in a list (see Appendix A.2). Lower values indicate more neutral text.

The **True Positive Rate (TPR) gap** (Zhao et al., 2018) evaluates Equality of Opportunity by measuring differences in TPRs between subgroups (lower gaps are fairer).

Finally, the **Area Under Similarity Curve (AUSC)** (Haque et al., 2024) evaluates indirect bias via attention-based explanations; higher AUSC suggests greater reliance on sensitive context tokens.

6 Result Analysis

6.1 Fine-tuning Performance (RQ1)

We fine-tune different pre-trained LLMs using CrAWD on the Jigsaw dataset. Table 1 shows the bias in the pre-trained models, the vanilla fine-tuned models, and the CrAWD fine-tuned models. We run all the models 5-repeat rounds and report the average value of each metric in Table 1. For ICAT, which combines LM performance (LMS) and stereotype bias (SS), CrAWD has the highest ICAT for each LLM. Notably, CrAWD sig-

Table 1: Average fine-tuning performance of CrAWD in different pre-trained LLMs over 5-repeat runs. $*$ denotes that the model fine-tuned by our proposed CrAWD method has statistically significant differences with both pre-trained and vanilla fine-tuned models under a one-tailed t-test with $p < 0.05$.

LLM	**Pre-Trained**		**Vanilla**		**CrAWD**	
	ICAT↑	**WEAT↓**	**ICAT↑**	**WEAT↓**	**ICAT↑**	**WEAT↓**
GPT2	63.817	0.980	67.078	0.693	67.997	0.374*
Llama2-7B	65.527	0.863	67.756	0.722	69.560*	0.406*
Llama3-8B	64.924	0.583	64.706	0.557	66.136*	0.345*
Falcon-7B	66.476	0.498	67.711	0.340	69.471*	0.247*
FlanT5-3B	66.000	0.496	67.123	0.508	68.241*	0.431
Deepseek-6.7B	66.667	0.221	65.432	0.897	68.452*	0.200

Table 2: Gender bias in text generation in fine-tuned models on the Bias in Bios dataset

LLM	**Model**	**GAS↓**	**WEAT↓**	**PPL↓**
Llama2-7b	Vanilla	0.860	0.582	15.360
	IBM	0.800	0.613	17.108
	DT	0.790	0.589	21.896
	DSD	0.750	0.526	15.553
	CrAWD	0.340	0.194	22.646
Llama3-8B	Vanilla	0.895	0.640	19.906
	IBM	0.835	0.632	28.527
	DT	0.870	0.697	41.533
	DSD	0.900	0.704	41.677
	CrAWD	0.550	0.639	46.993

nificantly outperforms both pre-trained and standard fine-tuned models across all LLMs, except for GPT-2. This suggests that CrAWD effectively reduces stereotype bias while maintaining high utility. For WEAT, which measures the association bias in the embedding space, CrAWD has the lowest WEAT for each LLM. The results indicate that fine-tuning with CrAWD can effectively mitigate association bias. Specifically, CrAWD consistently achieves significantly lower WEAT scores than both pre-trained and fine-tuned methods across all pre-trained LLMs, except for FLanT5-3B and Deepseek-6.7B. A possible reason for the relatively low WEAT scores of pre-trained FlanT5 and DeepSeek is that these models have already incorporated bias-mitigation mechanisms in their word embeddings during pre-training (e.g., instruction tuning for FlanT5 (Chung et al., 2022) and domain-diverse training corpora for Deepseek (DeepSeek-AI et al., 2025)). In addition, the vanilla fine-tuning Deepseek model resulted in a dramatic increase in bias, suggesting that the vanilla fine-tuning process without bias-mitigation mechanisms can amplify biases even in an initially unbiased LLM. In conclusion, it is advantageous to use CrAWD during LLM fine-tuning for all LLMs.

For qualitative assessment, we analyse outputs from Llama-2-7B fine-tuned on Bias in Bios with the Vanilla and CrAWD (Figure. 3, Appendix A.4). Vanilla fine-tuning introduces explicit gender tokens and reproduces occupational gender stereotypes, whereas CrAWD produces gender-neutral generations devoid of such bias. In Appendix A.4, Figure. 4, attention-map visualization for the prompt “The doctor said that. . . ” illustrates the effect: cross-attention can successfully identify the biased association between “doctor” in the prompt and “she” in the reference bias token with the highest attention score in the cross attention matrix.

6.2 Comparison with Baselines (RQ2)

We use three datasets to evaluate CrAWD against other fine-tuning methods to mitigate different biases in text generation or classification. We fine-tune Llama2-7B and Llama3-8B for text generation and BERT-base for classification.

We evaluate gender bias in text generation in Llama2-7B and Llama3-8B models fine-tuned on the Bias in Bios dataset. As shown in Table 2, CrAWD has the lowest scores on GAS and WEAT. It mitigates bias in the model’s attention mechanism, which is reflected in multiple bias metrics. For text generation utility evaluated by PPL, CrAWD has a slightly higher utility cost in this setting as the result of utility-fairness trade-off. Debias Tuning can only mitigate for a specific metric target, GAS, so it does not generalize to other evaluations.

To consider a more realistic scenario, in Table 3 we limit the model’s prior knowledge of the evaluation task. Both Debias Tuning and CrAWD only consider token “he” and “she” for gender during fine-tuning. The evaluation uses Appendix A.2 to calculate GAS and even more for WEAT in LLama models. CrAWD only needs a simple reference {“he”, “she”} to mitigate bias effectively. Whereas, Debias Tuning’s mitigation performance is very limited to the prior knowledge during training. It needs more information to achieve good perfor-

Table 3: Model training with different amounts of prior knowledge (in terms of sensitive tokens)

Sensitive sequence	Debias Tuning (DT)			CrAWD		
	GAS↓	WEAT↓	PPL↓	GAS↓	WEAT↓	PPL↓
he, she	0.790	0.589	21.896	0.340	0.194	22.646
he, she, man, woman	0.750	0.546	21.797	0.485	0.339	22.421
he, she, man, woman, male, female	0.710	0.595	24.339	0.405	0.295	23.103

Table 4: Racial bias in text classification in BERT-base models fine-tuned on the Hate Speech dataset

Model	Accuracy↑	TPR Gap↓	AUSC↓
Vanilla	0.944	0.181	0.717
AD	0.916	0.035	0.702
CBE	0.924	0.053	0.604
IBM	0.874	0.050	0.636
CrAWD	0.972	0.042	0.413

Table 5: The performance of CrAWD using different Top-$K\%$

Top-$K\%$	GAS↓	WEAT↓	PPL↓
10	0.485	0.660	16.609
20	0.430	0.624	15.642
30	0.400	0.606	19.687
40	0.340	0.194	22.646

Table 6: The performance of CrAWD using different weight decay values λ

Weight Decay (λ)	GAS↓	WEAT↓	PPL↓
0.1	0.176	0.665	15.180
0.2	0.340	0.194	22.646
0.3	0.125	0.341	19.298
0.4	0.410	0.421	22.197

mance on GAS and WEAT and even then the performance is not as good as CrAWD. In addition, it can be also observed that the relative performance improvement between the CrAWD method fine-tuned on "he, she" and "he, she, man, woman, male, female" is 17.14% for GAS and 52.05% for WEAT, while the performance drop for PPL is only 2.02%. The results clearly indicate that using simple references in CrAWD fine-tuned method dramatically improves LLM performance on GAS and WEAT while maintaining comparable PPL.

We also evaluate racial bias using their related sensitive tokens (see Appendix A.1) in text classification in BERT-base models fine-tuned on the Hate Speech dataset. As shown in Table 4, CrAWD has the second lowest TPR gap and lowest AUSC, indicating that it is effective in removing both direct bias in model prediction and indirect bias in model explanation. CrAWD has the highest accuracy (higher than Vanilla), which shows that CrAWD has a good balance of model utility and fairness. We also evaluate multiple types of biases in text generation in fine-tuned models on the SNLI dataset (see Appendix A.3).

6.3 Trade-off Analysis (RQ3)

We evaluate the performance of CrAWD using different values of hyperparameters when fine-tuning Llama2-7B on the Bias in Bios dataset. The values with the optimal trade-off between fairness (measured by WEAT) and utility (measured by PPL) are used for all other CrAWD experiments.

Top-$K\%$ defines the percentage of cross-attention weights selected to decay. A higher $K\%$ identifies more tokens that have biased associations with sensitive attributes. As shown in Table 5, with K increases, the model discovers more biased associations and better mitigates bias as indicated by the lower WEAT score. However, it also mis-identifies meaningful associations as bias and worsen the model utility as indicated by the higher PPL score. At Top-40%, the bias is low and the model utility is still satisfactory.

Weight Decay value λ controls the strength of weight decay applied to the identified biased tokens. A lower value of λ leads to a stronger bias mitigation, which trade-off more utility for fairness. As shown in Table 6, with λ decreases, the model deemphasizes more on tokens with biased associations. It lowers the WEAT acore and better mitigates bias. It also degrades the model utility as indicated by the higher PPL score. At $\lambda = 0.2$, the bias is low and the model utility is still satisfactory.

7 Conclusion

In this work, we introduced Cross-Attention-based Weight Decay (CrAWD), a method for mitigating multiple types of bias in LLMs. CrAWD uses cross-attention mechanism to identify and reduce biased associations with sensitive attributes during fine-tuning, without requiring prior knowledge of bias. Our approach successfully reduces bias in model outputs while preserving performance, offering a versatile and effective solution for promoting fairness in diverse NLP tasks across various domains.

8 Limitations

The CrAWD method introduces a cross-attention layer, adding extra parameters to the LLM architecture and increasing runtime complexity during training and inference. This additional computational overhead may limit its applicability in resource-constrained environments or scenarios requiring low latency. In future work, we will explore parameter-light alternative solutions to reduce computational costs, which effectively mitigate bias without increasing model complexity.

9 Ethical Considerations

We aim to mitigate biases in LLMs using the CrAWD method. While we utilize widely used corpora for training, we acknowledge that they may contain harmful or biased content, we are not responsible for any offensive material they include. Our focus is on reducing biased associations within models to generate less biased outputs, assessed through bias-related metrics, without compromising overall performance. We do not explore the potential extrinsic harms that might arise from employing the debiasing methods studied.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4349–4357.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai

Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Omkar Dige, Jacob-Junqi Tian, David Emerson, and Faiza Khan Khattak. 2023. Can instruction fine-tuned language models identify social bias through prompting? *arXiv preprint arXiv:2307.10472*.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Farsheed Haque, Depeng Xu, and Shuhan Yuan. 2024. Discovering and mitigating indirect bias in attention-based model explanations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1599–1614, Mexico City, Mexico. Association for Computational Linguistics.

Zexue He, Yu Wang, Julian McAuley, and Bodhisattwa Prasad Majumder. 2022. Controlling bias exposure for fair interpretable predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (* SEM)*, pages 43–53.

Donggeun Ko, Dongjun Lee, Namjun Park, Kyoungrae Noh, Hyeonjin Park, and Jaekwang Kim. 2023. Amplibias: Mitigating dataset bias through bias amplification in few-shot learning for generative models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4028–4032.

Zhongkun Liu, Zheng Chen, Mengqi Zhang, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2024. Zero-shot position debiasing for large language models. *arXiv preprint arXiv:2401.01218*.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.

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

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aishik Rakshit, Smriti Singh, Shuvam Keshari, Arijit Ghosh Chowdhury, Vinija Jain, and Aman Chadha. 2025. From prejudice to parity: A new approach to debiasing large language model word embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6718–6747, Abu Dhabi, UAE. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jacob-Junqi Tian, Omkar Dige, D Emerson, and Faiza Khattak. 2023. Using chain-of-thought prompting for interpretable recognition of social bias. In *Socially Responsible Language Modelling Research*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. “kelly is a warm person, joseph is a role model”: Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Song Wang, Jing Ma, Lu Cheng, and Jundong Li. 2023. Fair few-shot learning with auxiliary sets. In *ECAI 2023*, pages 2517–2524. IOS Press.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017a. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017b. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Ruixiang Zhao, Anna Zhou, and Kecheng Mao. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. 2024. Mitigating the bias of large language model evaluation. *arXiv preprint arXiv:2409.16788*.

A Appendix

A.1 Implementation Details

For LLM fine-tuning, we employ QLoRA (Dettmers et al., 2024) for parameter-efficient training with the following configuration parameters: rank = 16, alpha = 32, dropout = 0.1, and task_type = "causal_lm". The model undergoes training for 5 epochs with a learning rate of 10^{-5}. We use cross-entropy as our loss function and optimize it with the AdamW8bit optimizer. We conducted all the experiments on 2 NVIDIA Tesla A100 GPUs with 400GB memory. For experiments of RQ1, we run all the models for 5-repeat rounds and report the average value of each metric. All experiment results are reproducible using a provided set of random seeds.

For the detailed architecture of CrAWD, we use the pre-trained LLama2-7B-chat from Hugging Face as an example. LLama2-7B-chat comprises 32 decoder layers, each featuring 8 attention heads and a dropout rate of 0.1. The embedding size of the model is 4,096. In CrAWD fine-tuning, we enhance this architecture by adding a single cross-attention layer using PyTorch's nn.MultiheadAttention, is also configured with a dropout rate of 0.1. This cross-attention layer introduces additional parameters which is less than 1% of the model and it take 15% additional computing time per epoch. Both the input sequence $\boldsymbol{x}$ and the sensitive sequence $\boldsymbol{b}$ are fed into the model. We pad both sequences to a maximum length of 32, i.e., $n = m = 32$. After computing the cross-attention weights, we select the Top-40% of the highest weights and apply weight decay by multiplying them by weight decay parameter $\lambda = 0.2$, effectively reducing their values by 80%.

The sensitive sequence $\boldsymbol{b}$ consists of the following tokens {"he", "she", "gay", "straight", "young", "old", "black", "white", "Asian", "European", "American", "Mexican", "disabled", "abled", "Muslim", "Christian", "Jewish", "fat", "thin", "rich", "poor"}. It only needs one reference token for each sensitive attribute group. In the embedding space, the cross-attention mechanism examines similar tokens related to the sensitive attribute and discovers the biased associations with those tokens. For different datasets, we use a subset of $\boldsymbol{b}$. We use only gender tokens for the Bias in Bios dataset, racial tokens for the Hate Speech dataset, and the full set for the SNLI dataset.

For the WEAT calculation, we use ChatGPT o1 mini to make suggestions on the word sets for the considered sensitive attributes.

A.2 Gender Attribute Score (GAS) Evaluation Set

In the GAS metric for assessing gender bias in the model's text generation, we use the following evaluation test set {"he", "she", "him", "her", "his", "hers", "man", "woman", "male", "female", "boy", "girl", "father", "mother", "son", "daughter", "brother", "sister", "king", "queen", "actor", "actress", "husband", "wife", "uncle", "aunt", "sir", "madam"}. Any generated text containing at least one word from this set is considered as biased, whereas text that excludes all such words is considered as neutral.

A.3 Handling Multiple Types of Bias

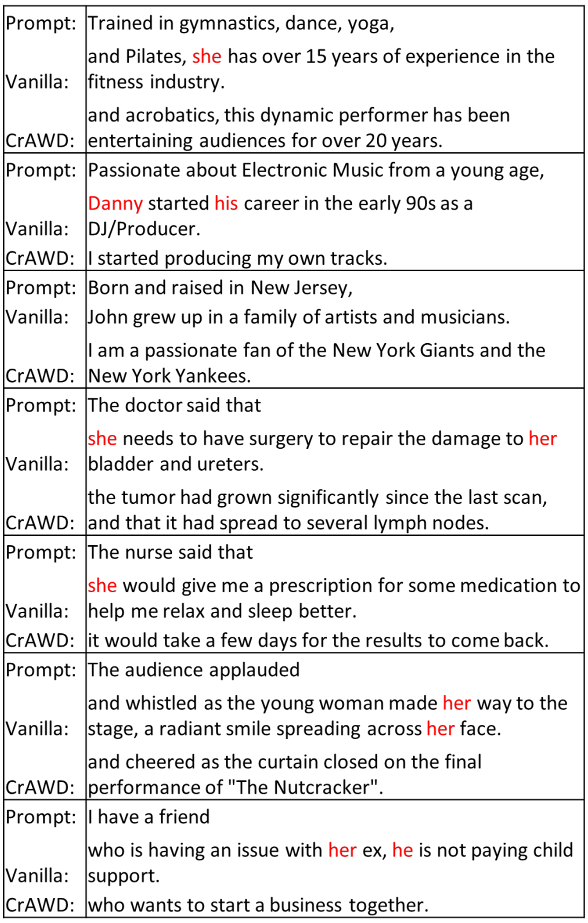
To evaluate how our model performs in mitigating multiple types of bias, we fine-tuned LLaMA2-7b and LLaMA3-8B models on the SNLI dataset, which contains sensitive tokens related to gender, race, and age (See Appendix A.1). Our reference sequence, $\boldsymbol{b}$, is designed to discover and mitigate multiple types of biases simultaneously, without relying on full knowledge of the evaluation task. As shown in Table 7, baseline models struggle to address all three sensitive attributes effectively. While they may perform well for one type of bias, they often fail to mitigate others. For instance, models like IBM might significantly reduce gender bias in Llama3-8B, but they do so at the expense of race or age bias, typically sacrificing perplexity. Similarly, Debias Tuning excels in mitigating one type of bias, but leaves other dimensions relatively unchecked. In contrast, CrAWD provides a more stable and balanced approach. It consistently ranks among the top two methods across all bias metrics in both LLaMA2-7b and LLaMA3-8B, effectively reducing gender, race, and age biases without significantly harming perplexity. This stability arises because CrAWD adapts its reference sequence $\boldsymbol{b}$ to address various sensitive tokens simultaneously, rather than focusing on a single bias at the expense of others. As a result, CrAWD offers a more holistic and reliable debiasing solution, mitigating multiple biases at once while maintaining model fluency. for 45 seconds

A.4 Case Study

For qualitative analysis, we examine some text examples generated by Llama2-7B fine-tuned on the

Table 7: Multiple bias in text generation in fine-tuned models on the SNLI dataset

LLM	Model	WEAT↓			PPL↓
		Gender	Race	Age	
LLaMA2-7b	Vanilla	0.955	0.526	0.049	18.243
	IBM	0.942	0.846	0.006	19.958
	Debias Tuning	0.929	0.667	0.072	17.512
	CrAWD	0.938	0.445	0.028	20.004
LLaMA3-8B	Vanilla	0.901	0.190	0.565	32.072
	IBM	0.546	0.302	0.984	111.783
	Debias Tuning	0.878	0.247	0.708	29.154
	CrAWD	0.885	0.349	0.576	35.952



Prompt:	Trained in gymnastics, dance, yoga,
Vanilla:	and Pilates, she has over 15 years of experience in the fitness industry.
CrAWD:	and acrobatics, this dynamic performer has been entertaining audiences for over 20 years.
Prompt:	Passionate about Electronic Music from a young age,
Vanilla:	Danny started his career in the early 90s as a DJ/Producer.
CrAWD:	I started producing my own tracks.
Prompt:	Born and raised in New Jersey,
Vanilla:	John grew up in a family of artists and musicians.
CrAWD:	I am a passionate fan of the New York Giants and the New York Yankees.
Prompt:	The doctor said that
Vanilla:	she needs to have surgery to repair the damage to her bladder and ureters.
CrAWD:	the tumor had grown significantly since the last scan, and that it had spread to several lymph nodes.
Prompt:	The nurse said that
Vanilla:	she would give me a prescription for some medication to help me relax and sleep better.
CrAWD:	it would take a few days for the results to come back.
Prompt:	The audience applauded
Vanilla:	and whistled as the young woman made her way to the stage, a radiant smile spreading across her face.
CrAWD:	and cheered as the curtain closed on the final performance of "The Nutcracker".
Prompt:	I have a friend
Vanilla:	who is having an issue with her ex, he is not paying child support.
CrAWD:	who wants to start a business together.

Figure 3: Examples generated by Llamma2-7B fine-tuned using the Vanilla and CrAWD methods

Bias in Bios dataset using the Vanilla and CrAWD methods, shown in Figure 3. The vanilla fine-tuned model generates texts with a gender narrative (with the presence of specific gender token), where it exhibits gender stereotyping bias in occupations. The CrAWD model generates gender neutral texts, which do not have a gender narrative nor show any gender bias associated with occupations.

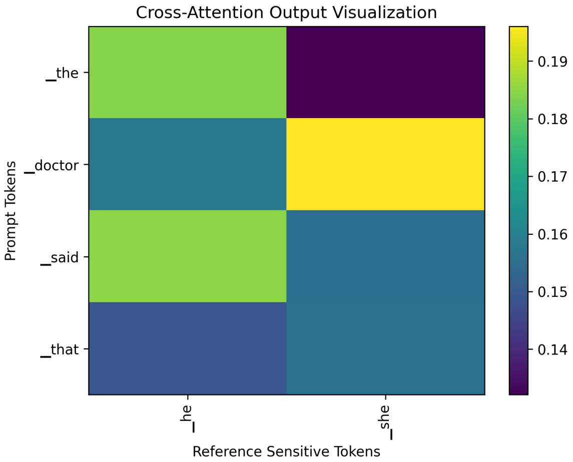


Figure 4: Cross-attention visualization for the prompt “The doctor said that. . . ”

To further inspect, we visualize the final decoder layer’s cross-attention matrix with the predefined sensitive sequence {“he”, “she”} for the prompt “The doctor said that. . . ”. Figure. 4 shows the strongest weight links to the input token “doctor” to the word “she”, meaning the model’s decision is driven mainly by the word “doctor”. This pronounced attention link exposes an implicit gender association for the profession, evidence that in the model’s training it picks up association bias with certain gender specific words which lead to biased text generation and CrAWD can successfully reduce this association with the weight decay trick.