# Debiasing the Fine-Grained Classification Task in LLMs with Bias-Aware PEFT

**Anonymous ACL submission**


## Abstract

Fine-grained classification via LLMs is susceptible to more complex label biases compared to traditional classification tasks. Existing bias mitigation strategies, such as retraining, *post-hoc* adjustment, and parameter-efficient fine-tuning (PEFT) are primarily effective for simple classification biases, such as stereotypes, but fail to adequately address prediction propensity and discriminative ability biases. In this paper, we analyze these two bias phenomena and observe their progressive accumulation from intermediate to deeper layers within LLMs. To mitigate this issue, we propose a bias-aware optimization framework that incorporates two distinct label balance constraints with a PEFT strategy targeting an intermediate layer. Our approach adjusts less than 1% of the model's parameters while effectively curbing bias amplification in deeper layers. Extensive experiments conducted across 12 datasets and 5 LLMs demonstrate that our method consistently outperforms or matches the performance of full-parameter fine-tuning and LoRA, achieving superior results with lower perplexity.


## 1 Introduction

Large language models (LLMs) have demonstrated exceptional capabilities across a wide range of natural language processing (NLP) tasks (Qin et al., 2023; Li et al., 2024; Wei et al., 2022, 2023; Huo et al., 2023). Among these, fine-grained classification via LLMs (figcLLM) has gained significant attention in practical applications such as mental health assessment, recommendation systems, and conversational AI, owing to its ability to capture subtle distinctions between labels (Zhang and Guo, 2024; Luna-Jimenéz et al., 2024; Lin et al., 2025; Zhao et al., 2024; Xie and Pu, 2021; Welivita et al., 2021).

However, figcLLM introduces complex label biases that are not typically observed in traditional

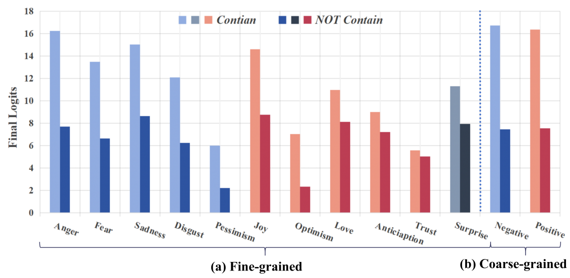



Figure 1: Average predicted logits of Gemma2-9b-it (Team, 2024) for each emotion label in TweetEmotion dataset (Mohammad et al., 2018). Figure (a) shows results for fine-grained categories, while Figure (b) displays results for coarse-grained categories.

classification tasks. Specifically, we have identified two distinct types of bias: (1) **prediction propensity bias**, where the model assigns disproportionately high probabilities to labels associated with high-frequency words from its pretraining corpus, and (2) **discriminative ability bias**, where the model struggle to differentiate between positive and negative samples for certain low-frequency labels.

Taking emotion detection as a case study, as illustrated in Figure 1, among the 11 emotion categories, the model assigns significantly higher probabilities to "anger" and "joy" compared to "pessimism" and "anticipation" due to the higher frequency of "anger" and "joy" in pretrain corpus. This demonstrates LLM outputs a clear preference for high-frequency emotion categories. Moreover, the model exhibits weak discriminability on "anticipation" and "trust", often producing nearly identical outputs regardless of whether these labels are present in the samples. Interestingly, when the same dataset was evaluated using coarse-grained labels ("positive", "negative"), these two phenomena were largely mitigated. It suggests that these two biases are closely linked to the complex and fine-grained task with low-frequency words as labels, which are typically absent in traditional classification tasks, thereby rendering these biases less noticeable.

Existing approaches to mitigating traditional label bias, such as stereotypes bias (Gira et al., 2022; Guo et al., 2022) and emotion bias (Fei et al., 2023; Hassan and Alikhani, 2023), can be broadly categorized into three groups: *post-hoc* correction techniques (Zhao et al., 2021; Fei et al., 2023; Yang et al., 2024; Mamta et al., 2024), full model retraining or fine-tuning (Thakur et al., 2023; Hassan and Alikhani, 2023; Zhou et al., 2023; He et al., 2022; Guo et al., 2022), and parameter-efficient fine-tuning (PEFT) (Hu et al., 2021; Gira et al., 2022; Xie and Lukasiewicz, 2023). *Post-hoc* methods primarily focus on correcting the model's final outputs while overlooking the underlying process of bias propagation and accumulation from intermediate layers to deep layers (Section 3). Although retraining-based approaches can be effective by adjusting the model's internal representations, they are computationally intensive and susceptible to catastrophic forgetting when applied to LLMs (Kirkpatrick et al., 2017; Gira et al., 2022). PEFT provides a trade-off between computational efficiency and adaptability, achieving performance comparable to full fine-tuning. Nevertheless, it struggles with figcLLM tasks, as it fails to explicitly address the intertwined nature of biases related to both prediction propensity and discriminative ability.

To mitigate these two biases, we propose a bias-aware optimization framework that incorporates two distinct loss functions, each targeting a specific bias type. First, to mitigate prediction propensity bias, we introduce a constraint that regulates the logits distribution across labels, ensuring a more balanced prediction tendency. Second, to enhance discriminative ability, we employ a contrastive loss that strengthens the model's capacity to distinguish between positive and negative samples for each specific label.

Furthermore, to reduce the amount of parameters for fine-tuning, we use interchange ablation to identify early layers where bias starts to propagate and key parameters which cause most effects on outputs. This enables targeted intervention at a certain layer to suppress bias accumulation as the model depth increases.

Through extensive experiments across 5 LLMs and 12 datasets, we demonstrate that our proposed approach effectively mitigates label bias, leading to improved classification performance and more balanced label predictions. Our method not only outperforms *post-hoc* correction techniques but also achieves results comparable to or exceeding those of full fine-tuning and PEFT-based methods, while maintaining lower perplexity.

Our main contributions are as follows.

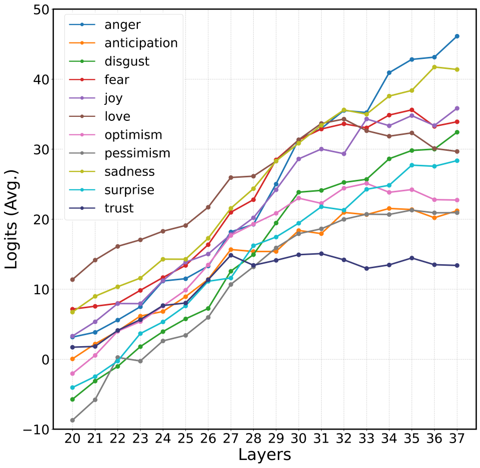
(1) We identify and analyze two specific phenomena of fine-grained label biases in LLMs and reveal that these biases originate from the progressive accumulation of erroneous predictions in intermediate layers, which become amplified in the deeper layers.

(2) We propose a simple yet PEFT strategy, incorporating two bias balance losses. This approach requires adjusting less than 1% of the total parameters.

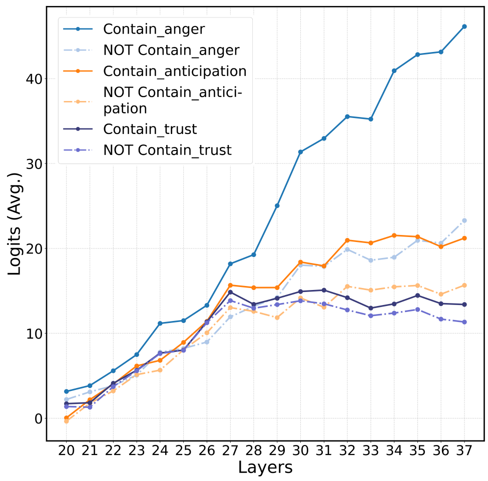
(3) We conduct extensive experiments, demonstrating the effectiveness of our method in figcLLM tasks while showcasing its adaptability to other domains.

## 2 Related Works

**Label bias.** Existing works used to mitigate label bias can be roughly divided into three categories. (1) Retraining-based approaches. Depending on whether they involve data manipulation or not, these methods are further divided into two strategies: data-based and algorithm-based (Thakur et al., 2023). The former balances the training dataset through techniques such as counterfactual data generation or resampling (Xie and Lukasiewicz, 2023; He et al., 2022; Thakur et al., 2023), while algorithm-based approaches modify the architecture or training constraints (Zhou et al., 2023; Hassan and Alikhani, 2023). However, they are difficult to apply to fine-grained tasks or are computationally expensive. (2) PEFT-based methods. Gira et al. (2022) proposed a new fine-tuning strategy by adding a linear to the input and output of the model and unfreezing some parameters. (3) *Post-hoc* approaches: These methods attempt to correct label biases after the model has made its predictions. For example, CC (Zhao et al., 2021) and DC (Fei et al., 2023) recalibrate predictions based on the unbalanced probability distributions generated by the model for free-text inputs (e.g.,"N/A" or random tokens). Additionally, Yang et al. (2024) pruned the top-K neurons contributing most to biased labels. Although these *post-hoc* approaches mitigate bias to some extent, they predominantly focus on adjusting the label probabilities in the final




(a) *Contain* of each label




(b) Distinguish ability of specific labels

Figure 2: The changing trend of *Contain*, *NOT Contain* of labels in the Gemma2 (9B) model from 20th layer to 37th layer.

output or target only a limited, discrete subset of neurons. As a result, they overlook the ongoing accumulation of bias within the intermediate layers of the model, making it challenging to fundamentally address the root causes of bias.

**Intermediate layers.** Recent studies have investigated the effectiveness of intermediate layers in large language models (Skean et al., 2024; Chen et al., 2024b; Sawtell et al., 2024; Valeriani et al., 2023). Valeriani et al.'s (2023) work demonstrated that the semantic information is better expressed at the intermediate layers. In a similar vein, Skean et al. (2024) and Sawtell et al. (2024) observed that the intermediate layers of a transformer-based model yield superior performance on various downstream tasks, including classification of embeddings. Our approach further reveals that the influence of bias is markedly diminished in the intermediate layers compared to the deeper layers, and we also show how the hidden state of the intermediate layer can be used to efficiently train a fairer LLMs for a wide range of tasks.

## 3 Bias Accumulation Analysis

To investigate the dynamics and effects of bias within the model, we performed a visual analysis on the TweetEmotion dataset (Mohammad et al., 2018) using an early exit strategy (Teerapittayanon et al., 2016; Elbayad et al., 2020; Schuster et al., 2022). This method applies language heads ($lm\_head$), which is a unembedding matrix, directly to the hidden states of intermediate layers.

We first randomly sampled a class-balanced subset from training data and conducted evaluation under a zero-shot setting, without explicit instructions. For each target label, we divided the samples into two types: those whose true label contained the target label ($Contain$) and those whose true label did not ($NOT\ Contain$). Using the Gemma2-9b-it model, we predicted the target labels at each layer and calculated the mean logits for each of the two sample sets. For instance, for the label "anger", we recorded the logits as $Contain_{anger}$ and $NOT\ Contain_{anger}$, respectively.

The experimental results are presented in Figure 2 (a-b). Figure 2a illustrates the variation of $Contain$ across all labels with respect to model depth, while Figure 2b compares the depth-dependent changes in $Contain$ and $NOT\ Contain$ of both the high-frequency word ("anger") and low-frequency words ("anticipation" and "trust"), providing a clear contrast. From these figures, we observe that fine-grained label biases exist even in the intermediate layers:

(1) Preference for high-frequency labels. In Figure 2a, the $Contain$ values for high-frequency labels (e.g., "anger", "sadness", "joy") are consistently higher than those for low-frequency labels (e.g., "anticipation", "trust", "pessimism"). Furthermore, the gap between high- and low-frequency labels grows with increasing model depth beginning with intermediate layers.

(2) Difficulty distinguishing low-frequency labels. In Figure 2b, the distance between $Contain$ and $NOT\ Contain$ is significantly wider for high-frequency labels such as "anger" than for low-frequency labels such as "anticipation" and "trust". The gap is also progressively amplified as the

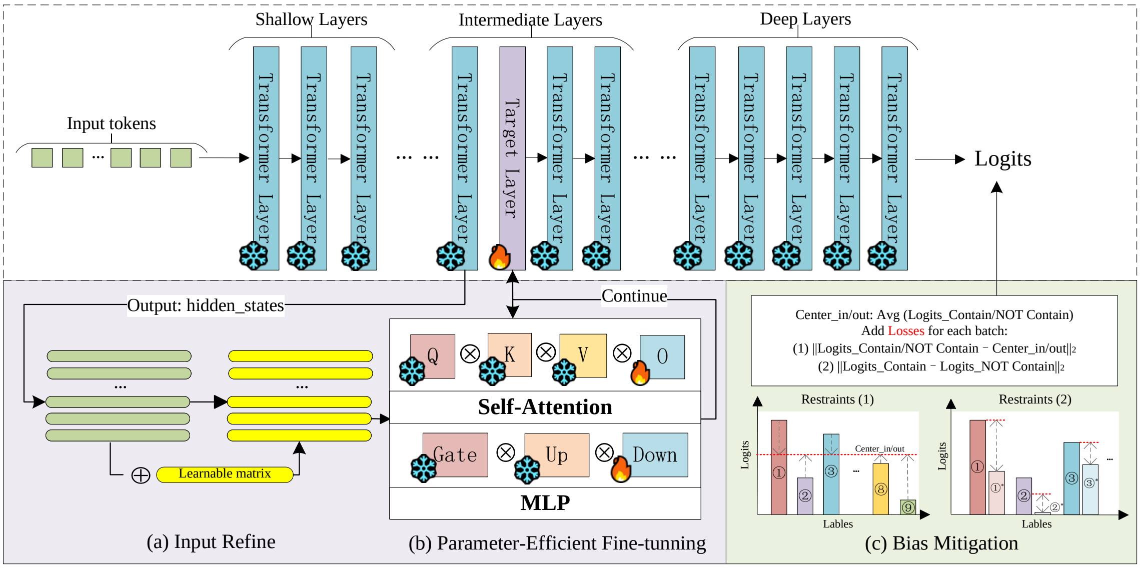



Figure 3: The overview of our method.

model depth increases at the beginning of intermediate layers.

Upon analyzing the common causes of these two biases, we conclude that they primarily stem from incorrect predictions made in the intermediate layers. These errors accumulate and propagate through deeper layers, ultimately influencing the final predictions. Thus, suppressing the accumulation of biases at the intermediate layers emerges as a feasible and effective strategy for bias mitigation.

# 4 Methodology

This section provides a comprehensive overview of the proposed methodology, as depicted in Figure 3. The task definition is first introduced, followed by a detailed discussion of the proposed approach, which comprises two key components: the determination of fine-tuning parameters, and the incorporation of bias balance constraints.

## 4.1 Task Definition

Given a supervised natural language processing (NLP) dataset $(X, Y)$, where $X$ denotes the input texts and $Y$ represents the corresponding category labels, along with a prompt template $P$, such as "Review: $[X]$. Emotion:", model is parameter-efficiently fine-tuned to learn the mapping: $\mathcal{M}(P, X) \rightarrow Y$. This process enhances the model's ability to mitigate undesirable associations between biases and labels.

## 4.2 Overview

### 4.2.1 Determine Fine-tuning Parameters

The computational cost of fine-tuning all layers is substantial. A key focus is to determine whether similar results can be achieved by fine-tuning only a small number of parameters in specific layers. Based on the analysis in Section 3, we found that biases largely arise from the accumulation of error predictions as the model deepens. Consequently, we aim to correct the early manifestations of bias by intervening in the internal states of one selected intermediate layer.

**Intermediate target layer.** We identify the target layer for fine-tuning by analyzing the extent to which the model's internal mechanisms contribute to biased predictions, using the interchange ablation method. Specifically, we replace the activation values of the golden samples in selected components with the corresponding hidden representations of biased samples, and observe the resulting changes in the final output. We then select the decoder layer where the largest change occurs as the target for subsequent interventions. A more detailed implementation can be found in Appendix A.

**Unfreeze parameters.** For the selection of parameters to fine-tune within the target layer, we draw on theoretical insights from "memory component" (Chen et al., 2024a) and validate our choices through extensive experiments. Taking the Gemma2 model as an example, each decoder consists of a self-attention module ($q\_proj$, $k\_proj$,

$v\_proj$, $o\_proj$) and a feedforward network module ($gate\_proj$, $up\_proj$, $down\_proj$). Chen et al.'s (2024a) research indicates that the attention output matrix ($o\_proj$) and the final projection layer of the MLP ($down\_proj$) exhibit stronger memory characteristics, retaining rich knowledge acquired during pre-training. Motivated by this finding, we selectively fine-tune only the $o\_proj$ and $down\_proj$ parameters within the target layer, while keeping all other weights frozen to ensure parameter efficiency. Additional experimental results on alternative parameter combinations are provided in the Appendix B. Furthermore, to enhance the effectiveness of the target layer, we refine its input by integrating a learnable parameter into the hidden representations it receives.

#### 4.2.2 Bias Balance Loss

To address two specific types of fine-grained label biases, we design two corresponding bias-balancing constraints to complement the original language modeling loss during fine-tuning. For each batch, we separately compute the logits for samples that $Contain$ and $NOT\ Contain$ each label $c$, denoted as $H_c^C$ and $H_c^N$, respectively, based on the final predicted logits. Then, these are aggregated to form $H^C$ and $H^N$ across all labels.

$$\begin{aligned} H^C &= [H_{c1}^C, ..., H_{cn}^C]^T \\ H^N &= [H_{c1}^N, ..., H_{cn}^N]^T \end{aligned} \tag{1}$$

where $n$ is the number of label types appearing in a batch.

(1) Prediction propensity bias. To reduce the gap between the model's predicted logits for high- and low-frequency labels, we aimed to minimize the internal differences within $H^C$ and $H^N$. To achieve this, we apply an L2 norm constrain to regulate the distance between $H_c^C$ and $H_c^N$ relative to their respective centroids, $ct_{in}$ and $ct_{out}$.

$$\begin{aligned} \mathcal{L}_{bal1} &= \|H^C - ct_{in}\|_2 + \|H^N - ct_{out}\|_2, \\ \text{where } ct_{in} &= \frac{1}{|Y|}\sum_{c \in Y} H_c^C, \\ ct_{out} &= \frac{1}{|Y|}\sum_{c \in Y} H_c^N \end{aligned} \tag{2}$$

(2) Discriminative ability bias. To enhance the model's sensitivity to all labels, we constrained the distance between $H_c^C$ to $H_c^N$ for each label $c$, also utilizing the L2 norm.

$$\mathcal{L}_{bal2} = -\|H^C - H^N\|_2 \tag{3}$$

Finally, we define the final loss in the fine-tuning phase as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{LM} + \beta\mathcal{L}_{bal1} + \gamma\mathcal{L}_{bal2} \tag{4}$$

where $\mathcal{L}_{LM}$ is the language modeling loss, $\alpha$. $\beta$ and $\gamma$ are hyperparameters.

## 5 Experiment

### 5.1 Experimental setup

**Datasets.** We conducted extensive experimental evaluations on five fine-grained tasks and seven coarse-grained task datasets. The fine-grained tasks include emotion detection (SuperTweetEval (Antypas et al., 2023): TweetEmotion (Mohammad et al., 2018), TweetHate (Sachdeva et al., 2022), GoEmotions (Demszky et al., 2020), EmpatheticDialogues (Rashkin et al., 2019)) and fine-grained sentiment analysis (SST-5 (Socher et al., 2013)). The coarse-grained tasks encompass social bias question answering (SBQA (Parrish et al., 2022): BBQ-Age, BBQ-SES, BBQ-Disability, BBQ-Gender), topic classification (AGNews (Zhang et al., 2015)), natural language inference (RTE (Dagan et al., 2006)), and sentiment analysis (SST-2 (Socher et al., 2013)). Notably, the SBQA dataset differs from other datasets in that it contains a number of inconsistent candidate labels. For instance, in the socioeconomic status bias dataset BBQ-SES, the labels include terms such as *poor people*, *low-income people* and *the truck driver*. Further details about the datasets and the division of the training set can be found in Appendix C.

**Baseline.** For the fine-tuning approach, we compared parameter-efficient fine-tuning (LoRA (Hu et al., 2021)) and full-parameter fine-tuning. Additionally, we compared the *post-hoc* methods CC (Zhao et al., 2021), DC (Fei et al., 2023) and CRISPR (Yang et al., 2024). A detailed description of the baselines is provided in Appendix D.

**Models and Implementation Details.** In our work, we utilized five LLMs, all sourced from HuggingFace[1]: Gemma2-2b-it, Gemma2-9b-it (Team, 2024), Mistral-7b-Instruct (Jiang et al., 2023), Llama3.2-1b, Llama3.2-3b (Grattafiori et al., 2024). The primary experiments were conducted on fine-grained tasks, both with and without instructions in a zero-shot setting. Other experiments were carried

[1] https://huggingface.co

| Model | All Layers | Fine-grained | | | | | Coarse-grained | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TE | GE | ED | TH | S5 | S2 | BA | BS | BD | BG | AG | RTE |
| Gemma2 (2B) | 26 | 15 | 15 | 12 | 6 | 22 | 15 | 22 | 21 | 22 | 21 | 12 | 19 |
| Gemma2 (9B) | 42 | 22 | 22 | 22 | 28 | 28 | 25 | 25 | 25 | 25 | 25 | 22 | 25 |
| Mistral (7B) | 32 | 12 | 12 | 12 | 12 | 12 | - | - | - | - | - | - | - |
| Llama3 (1B) | 16 | 9 | 10 | 8 | 5 | 10 | - | - | - | - | - | - | - |
| Llama3 (3B) | 28 | 14 | 14 | 8 | 14 | 14 | - | - | - | - | - | - | - |

Table 1: Target layer of each dataset in our experience. Because of the better performance overall the fine- and coarse-grained tasks on Gemma2 models (2B and 9B), we conduct the coarse-grained tasks (adaptability) only on these two models.

| Model | Method | TweetEmotion | GoEmotions | EmpathicDialogues | TweetHate | SST-5 |
|---|---|---|---|---|---|---|
| Gemma2 (2B) | Original | 59.33 | 27.96 | 30.05 | 53.87 | 38.38 |
| | CC | 42.42 (-16.91) | 12.14 (-15.82) | 13.98 (-16.07) | 8.35 (-45.52) | 38.32 (-0.06) |
| | DC | 65.41 (+6.08) | 20.04 (-7.92) | 35.50 (+5.45) | 24.55 (-29.32) | 40.96 (+2.58) |
| | CRISPR | 61.13 (+1.80) | 30.43 (+2.47) | 27.13 (-2.92) | 56.52 (+2.65) | 35.44 (-2.94) |
| | LoRA | 73.95 (+14.62) | 50.11 (+22.15) | 51.39 (+21.34) | 15.83 (-38.04) | 54.83 (+16.45) |
| | Full FT | 71.41 (+6.08) | 49.76 (+21.80) | **58.95 (+28.90)** | 14.28 (-39.59) | 48.87 (+10.49) |
| | Ours | **75.87 (+16.54)** | **56.17 (+28.21)** | 57.31 (+27.26) | **65.67 (+11.80)** | **55.70 (+17.32)** |
| Gemma2 (9B) | Original | 66.34 | 24.03 | 45.04 | 56.28 | 54.86 |
| | CC | 65.18 (-1.16) | 25.19 (+1.16) | 44.73 (-0.31) | 22.49 (-33.79) | 45.34 (-9.52) |
| | DC | 67.94 (+1.60) | 24.97 (+0.94) | 47.83 (+2.79) | 32.33 (-23.95) | 51.69 (-3.17) |
| | CRISPR | 69.72 (+3.38) | 26.63 (+2.60) | 45.78 (+0.74) | 57.26 (+0.98) | 50.13 (-4.73) |
| | LoRA | 74.34 (+8.00) | 49.24 (+25.21) | 56.66 (+11.62) | 14.90 (-41.38) | 56.63 (+1.77) |
| | Full FT | 74.27 (+7.93) | 51.13 (+27.10) | 57.36 (+12.32) | 14.62 (-41.66) | 55.81 (+0.95) |
| | Ours | **75.49 (+9.15)** | **54.29 (+30.26)** | **59.13 (+14.09)** | **70.42 (+14.14)** | **61.08 (+6.22)** |
| Mistral (7B) | Original | 67.31 | 32.45 | 47.07 | 63.06 | 37.99 |
| | CC | 64.95 (-2.36) | 22.30 (-10.15) | 48.25 (+1.18) | 49.17 (-13.89) | 34.17 (-3.82) |
| | DC | 62.72 (-4.59) | 24.51 (+7.94) | 41.89 (-5.18) | 29.89 (-33.17) | 40.51 (+2.52) |
| | CRISPR | 67.05 (-0.26) | 27.53 (-4.92) | 50.76 (+3.69) | **70.51 (+7.45)** | 44.82 (+6.83) |
| | LoRA | 71.81 (+4.50) | 48.15 (+15.70) | 59.34 (+12.27) | 13.47 (-49.59) | 53.42 (+15.43) |
| | Full FT | 72.69 (+5.38) | 48.94 (+16.47) | 58.22 (+11.15) | 14.84 (-48.22) | 53.70 (+15.71) |
| | Ours | **73.91 (+6.60)** | **50.34 (+17.89)** | **60.19 (+13.12)** | 55.67 (-7.39) | **57.74 (+19.75)** |
| Llama3 (1B) | Original | 42.71 | 8.95 | 19.97 | 18.25 | 24.90 |
| | CC | 49.94 (+7.23) | 12.30 (+3.35) | 30.31 (+10.34) | 4.87 (-13.38) | 20.84 (-4.06) |
| | DC | 50.95 (+8.24) | 16.42 (+7.47) | 37.32 (+17.35) | 4.82 (-13.43) | 29.20 (+4.30) |
| | CRISPR | 43.09 (+0.38) | 9.15 (+0.20) | 21.18 (+1.21) | 23.82 (+5.57) | 24.31 (-0.59) |
| | LoRA | 71.65 (+28.94) | 48.90 (+39.95) | 49.92 (+29.95) | 15.04 (-3.21) | 55.74 (+30.84) |
| | Full FT | 71.71 (+29.00) | 48.66 (+39.71) | 51.34 (+31.37) | 14.90 (-3.35) | 55.57 (+30.67) |
| | Ours | **72.55 (+29.84)** | **50.81 (+41.86)** | **51.36 (+31.39)** | **46.29 (+28.04)** | **57.67 (+32.77)** |
| Llama3 (3B) | Original | 46.35 | 12.70 | 19.27 | 3.48 | 27.27 |
| | CC | 51.06 (+4.71) | 5.85 (-6.85) | 28.85 (+9.58) | 14.31 (+10.83) | 20.79 (-6.48) |
| | DC | 57.92 (+11.57) | 15.97 (+3.27) | 35.53 (+16.26) | 12.65 (+9.17) | 31.12 (+3.85) |
| | CRISPR | 50.80 (+4.45) | 17.51 (+4.81) | 22.38 (+3.11) | 18.63 (+15.15) | 30.94 (+3.67) |
| | LoRA | 70.92 (+24.57) | 45.17 (+32.47) | 59.16 (+39.89) | 15.57 (+12.09) | **57.23 (+29.96)** |
| | Full FT | 73.82 (+27.47) | 50.21 (+37.51) | **61.03 (+41.76)** | 14.97 (+11.49) | 55.42 (+28.15) |
| | Ours | **74.22 (+27.87)** | **50.55 (+37.85)** | 58.98 (+39.71) | **65.99 (+62.51)** | 55.50 (+28.23) |

Table 2: The main results in the instruction setting. The **bold**/underlined font means the best/the second best result.

out exclusively with instructions. The templates and task instructions employed can be found in Appendix H. For the target layer selection step, we randomly selected 20 samples from the training set for evaluation, with the results presented in Table 1. Regarding hyperparameters, the learning rate was set to 5e-5, the batch size to 16, $\alpha = 1, \beta = 1, \gamma = 1$. All training was performed using FP16 precision on NVIDIA GeForce RTX 3090 GPUs.

## 5.2 Main Results

### 5.2.1 Fine-grained Classification

We evaluated the bias mitigation performance of our method and several baselines for fine-grained label biases. Table 2 presents the weighted F1 scores of various methods across five fine-grained datasets under the instruction setting, with results

| Model | TweetEmotion | | GoEmotions | |
|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 |
| Ours | 53.15 | 75.87 | 53.55 | 56.17 |
| $w/o\ \mathcal{L}_{bal1}$ | 47.46 | 73.62 | 52.00 | 54.60 |
| $w/o\ \mathcal{L}_{bal2}$ | 34.29 | 68.27 | 35.09 | 37.84 |
| $w/o\ \mathcal{L}_{bal1,2}$ | 33.65 | 67.63 | 49.69 | 49.94 |
| $w/o\ refine$ | 50.32 | 75.52 | 51.13 | 54.68 |
| unfreeze ($down$) | 49.83 | 74.70 | 52.80 | 55.75 |
| unfreeze ($o$) | 49.17 | 74.30 | 48.41 | 53.34 |
| unfreeze (-) | 37.13 | 69.10 | 27.54 | 29.63 |
| unfreeze ($q$, $k$, $v$) | 45.88 | 73.31 | 46.48 | 50.85 |
| unfreeze ($gate$, $up$) | 50.06 | 75.13 | 50.88 | 54.50 |

Table 3: Ablation experiments.

for the no-instruction setting available in Appendix E. The findings indicate that existing *post-hoc* methods (CC, DC, CRISPR) are limited in effectively mitigating fine-grained label biases. Particularly when applied to the TweetHate dataset, which exhibits a severe label imbalance, both CC and DC lead to a notable decline in task performance. While CRISPR shows some improvement in the instruction setting, its performance still lags behind that of the fine-tuning methods. In contrast, training-based methods, which adjust the model's intrinsic representations, are more effective in mitigating the negative impact of bias. However, on the TweetHate dataset, both full-parameter fine-tuning and LoRA fail to improve the metric, highlighting the complexity of the figcLLM task compared to traditional classification tasks. Notably, our approach achieves performance comparable to, or even better than, LoRA and full-parameter fine-tuning methods, despite updating far fewer parameters. This underscores the effectiveness of our strategy in suppressing bias accumulation within the deeper layers by intervening at the intermediate layer.

#### 5.2.2 Ablation

We also conducted ablation experiments using the Gemma2-2b-it model on the TweetEmotion and GoEmotions datasets to assess the impact of our proposed bias balance losses, learnable refine parameter ($refine$), and the choice of training components on the final task performance. Specifically, TweetEmotion is a multi-label classification task, for which we computed accuracy using the exact match principle. In each ablation experiment, we ensured that all settings remained constant except for modifications in the conditions under investigation. The results of these experiments are presented in Table 3.

In Table 3, $w/o\ \mathcal{L}_{bal1}$, $w/o\ \mathcal{L}_{bal2}$, and $w/o\ \mathcal{L}_{bal1,2}$ represent the removal of one or both bias balance losses, respectively. The last five lines represent different parameter combinations for unfreezing. The results reveal that omitting the balance losses significantly impairs task performance, with removal of $\mathcal{L}_{bal2}$ leading to greater degradation than removal of $\mathcal{L}_{bal1}$. This suggests that enhancing the model's discriminative ability for low-frequency labels is crucial for improving task performance. Moreover, freezing all components in the target layer severely hinders bias mitigation. Fine-tuning $q$, $k$, and $v$ in the target layer proves less effective than other combinations, while fine-tuning only $o$ and $down$ yields the best results with fewer parameters. More parameter combination experiments can be found in Appendix B.

#### 5.2.3 Parameter Analysis

Furthermore, Appendix F provides an analysis of the impact of varying the location of the target layer, training multiple decoder layers, and hyper-parameters $\alpha$, $\beta$, $\gamma$ on task performance.

### 5.3 Adaptability

We also test the adaptability of our method on coarse-grained tasks. Given the social bias question answering tasks and the more balanced label in the classification datasets, we chose accuracy as the evaluation metric for this experiment. Additionally, based on the observations in Section 5.2, where the Gemma2 models (2B and 9B) consistently outperformed others, we limited this section to the Gemma2 family of models.

Table 4 presents the performance of our approach in comparison with other baselines across four types of coarse-grained tasks. Consistent with the results from the fine-grained tasks, our method achieves superior performance on most of the datasets, particularly excelling on the topic classification dataset (AGNews) and the age bias dataset (BBQ-Age). These results strongly highlight the adaptation capability of our approach.

### 5.4 Perplexity

The fine-tuning approach is susceptible to the issue of "catastrophic forgetting", where the fine-tuned model may lose some of its original language modeling capability. To assess the impact of different fine-tuning methods on this aspect, we calculated the perplexity of the model before and after train-

| Model | Method | Type of Datasets | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | SBQA (BBQ) | | | | SA | TC | NLI |
| | | Age | SES | Disability | Gender | SST-2 | AGNews | RTE |
| Gemma2 (2B) | Original | 69.14 | 77.75 | 71.95 | 65.63 | 90.17 | 77.73 | 74.65 |
| | CC | 52.34 (-16.80) | 52.89 (-24.86) | 50.20 (-21.75) | 52.88 (-12.75) | 90.65 (+0.48) | 54.86 (-22.87) | 77.52 (+2.87) |
| | DC | 56.03 (-13.11) | 55.34 (-22.41) | 56.47 (-15.48) | 54.46 (-11.17) | 93.26 (+3.09) | 62.30 (-15.43) | 79.23 (+4.58) |
| | CRISPR | 70.10 (+0.96) | 79.06 (+1.31) | 69.18 (-2.77) | 68.33 (+2.70) | 92.08 (+1.91) | 76.64 (-1.09) | 77.25 (+2.60) |
| | LoRA | 82.60 (+13.46) | **98.77 (+21.02)** | 91.88 (+19.93) | 99.50 (+33.87) | **96.61 (+6.44)** | 91.47 (+13.74) | 81.85 (+7.20) |
| | Full FT | 86.60 (+17.46) | 96.75 (+19.00) | **92.06 (+20.11)** | 98.85 (+33.22) | 94.94 (+4.77) | 90.97 (+13.24) | 84.64 (+9.99) |
| | Ours | **96.98 (+27.84)** | 97.54 (+19.79) | 91.95 (+20.00) | **99.61 (+33.98)** | 95.63 (5.46) | **97.02 (+19.29)** | **84.82 (+10.17)** |
| Gemma2 (9B) | Original | 85.45 | 85.73 | 86.22 | 88.40 | 95.61 | 86.61 | 75.62 |
| | CC | 65.43 (-20.02) | 65.74 (-19.99) | 71.35 (-14.87) | 69.79 (-18.61) | 95.56 (-0.05) | 85.86 (-0.75) | 75.53 (-0.09) |
| | DC | 80.82 (-4.63) | 76.86 (-8.87) | 81.66 (-4.56) | 88.69 (+0.29) | 95.12 (-0.49) | 86.11 (-0.50) | 79.71 (+4.09) |
| | CRISPR | 86.45 (+1.00) | 84.33 (-1.40) | 85.51 (-0.71) | 89.62 (+1.22) | 95.53 (-0.08) | 86.47 (-0.14) | 77.99 (+2.37) |
| | LoRA | 94.04 (+8.59) | 99.42 (+13.69) | 97.32 (+11.10) | **99.95 (+11.55)** | 95.73 (+0.12) | 92.01 (+5.40) | 82.44 (+6.82) |
| | Full FT | 95.23 (+9.78) | 99.55 (+13.82) | **97.86 (+11.64)** | 99.26 (+10.86) | 95.90 (+0.29) | 94.06 (+7.45) | 86.61 (+10.99) |
| | Ours | **98.19 (+12.74)** | **99.77 (+14.04)** | 97.57 (+11.35) | 99.67 (+11.27) | **96.10 (+0.49)** | **97.87 (+11.26)** | **93.09 (+17.47)** |

Table 4: The results of generalization. The **bold**/underlined font means the best/the second best result.

| Method | WikiText-2: Perplexity (↓) | | | | |
|---|---|---|---|---|---|
| | Gemma2 (2B) | Gemma2 (9B) | Mistral (7B) | Llama3 (1B) | Llama3 (3B) |
| Original | 18.80 | 13.60 | 6.37 | 11.37 | 9.04 |
| LoRA | 35.68 | 34.29 | 8.04 | 22.05 | 15.59 |
| Full FT | 23.48 | 37.08 | 10.57 | 22.97 | 9.20 |
| Ours | 21.94 | 13.52 | 6.48 | 11.47 | 9.06 |

Table 5: The results of perplexity on fine-tuned methods.

ing, using the WikiText-2 datasets (Merity et al., 2016). As an example, we used the model saved after fine-tuning on the TweetEmotion, and the results are presented in Table 5.

It is evident that for the model fine-tuned using our method, the perplexity remains nearly identical to that of the initial model, indicating that our fine-tuning approach has minimal impact on the language modeling capability. In contrast, models fine-tuned with LoRA and full-parameter fine-tuning exhibit a significant increase in perplexity to varying degrees.

### 5.5 Visualisation

To demonstrate the mitigation effect of our fine-tuned model on fine-grained label biases, we visualized the $Contain$ and $NOT\ Contain$ of labels on TweetEmotion, as detailed in Section 3. The corresponding results are provided in Appendix G.

## 6 Conclusion

This work addresses the mitigation of label biases in Large Language Models (LLMs) for fine-grained classification tasks. We identify two distinct forms of fine-grained label biases within LLMs, named prediction propensity bias and discriminative ability bias, and explore the underlying causes of these biases, i.e., erroneous predictions in the intermediate layers are accumulated and amplified as the model depth increases. To counteract this issue, we propose two bias balance losses to parameter-efficiently fine-tune an intermediate layer. Notably, our method requires training less than 1% of the model's total parameters. Extensive experiments across a range of tasks and datasets demonstrate that our approach not only exceeds existing *post-hoc* methods in mitigating label biases, but also achieves performance comparable to, or even exceeding, that of full-parameter fine-tuning and LoRA. Our findings underscore the potential of intervening in the middle layer to enhance the fairness and accuracy of LLMs in fine-grained classification tasks.

## 7 Limitation

In this work, we have focused exclusively on LLMs with a decoder-only architecture and have not explored models with other architectural types, such as encoder-only or encoder-decoder structures. These alternative architectures warrant further investigation, particularly with respect to the variation of bias in the encoder modules, which may differ significantly from that observed in the decoders. Consequently, we plan to extend our study to include LLMs with diverse architectural configurations in future research.

## References


Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. 2024a. Unveiling language skills via path-level circuit discovery. *Preprint*, arXiv:2410.01334.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2024b. Is bigger and deeper always better? probing llama across scales and layers. *Preprint*, arXiv:2312.04333.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. Depth-adaptive Transformer. In *ICLR 2020 - Eighth International Conference on Learning Representations*, pages 1–14.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14014–14031.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69.

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

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: 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)*, pages 1012–1023.

Sabit Hassan and Malihe Alikhani. 2023. D-CALM: A dynamic clustering-based active learning approach for mitigating bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5540–5553.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. MABEL: Attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, page 11–20. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, pages 3521–3526.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18608–18616.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2025. How can recommender systems benefit from large language models: A survey. *ACM Trans. Inf. Syst.*, 43(2).

Cristina Luna-Jimenéz, Zoraida Callejas, and David Griol. 2024. Mental-health topic classification employing d-vectors of large language models. In *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 199–204.

Mamta Mamta, Rishikant Chigrupaatii, and Asif Ekbal. 2024. BiasWipe: Mitigating unintended bias in text classifiers through model interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21059–21070.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Phu Mon Thompson, Jana and Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics*, pages 2086–2105.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

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.

Mason Sawtell, Tula Masterman, Sandi Besen, and Jim Brown. 2024. Lightweight safety classification using pruned language models. *Preprint*, arXiv:2412.13435.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pages 17456–17472.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. 2024. Does representation matter? exploring intermediate layers in large language models. *Preprint*, arXiv:2412.09563.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 340–351.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, pages 51234–51252.

Fusheng Wei, Robert Keeling, Nathaniel Huber-Fliflet, Jianping Zhang, Adam Dabrowski, Jingchao Yang, Qiang Mao, and Han Qin. 2023. Empirical study of llm fine-tuning for text classification in legal document review. In *2023 IEEE International Conference on Big Data*, pages 2786–2792.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15730–15745.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9061–9073.

Jun Zhang and Yanrong Guo. 2024. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognition Letters*, 178:167–173.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

| Combination | | | | | | | Metric |
|---|---|---|---|---|---|---|---|
| $q$ | $k$ | $v$ | $o$ | $gate$ | $up$ | $down$ | F1 |
| ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | 70.66 |
| ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | 71.51 |
| ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | 73.46 |
| ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | 73.83 |
| ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | 75.03 |
| ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | 74.59 |
| ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 73.88 |
| ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | 73.03 |
| ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | 73.34 |
| ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | 73.93 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | 73.47 |
| ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 75.23 |
| ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | 74.50 |
| ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | 72.59 |
| ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | 74.09 |
| ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | 74.96 |
| ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | 74.91 |
| ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | 72.45 |
| ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | 74.30 |
| ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | 74.93 |
| ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | 74.10 |
| ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | 75.87 |
| ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | 74.83 |
| ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | 75.13 |
| ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | 74.73 |
| ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | 74.17 |
| ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | 74.63 |
| ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 74.70 |

Table 6: Results of selecting different combinations.


Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, pages 6889–6907.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4227–4241.


| Datasets | Class | Balanced | Train | Test |
|---|---|---|---|---|
| Fine-grained | | | | |
| TweetEmotion | 11 | ✗ | 886 | 3259 |
| GoEmotions | 28 | ✗ | 1000 | 5227 |
| Empathetic Dialogues | 32 | ✗ | 960 | 2538 |
| TweetHate | 7 | ✗ | 895 | 1433 |
| SST-5 | 5 | ✗ | 1000 | 2210 |
| Coarse-grained | | | | |
| BBQ-Age | - | ✗ | 368 | 3312 |
| BBQ-SES (socio-economic status bias) | - | ✗ | 686 | 6175 |
| BBQ-Disability (disability status bias) | - | ✗ | 155 | 1401 |
| BBQ-Gender (gender bias) | - | ✗ | 567 | 5105 |
| AGNews | 4 | ✓ | 760 | 6840 |
| RTE | 2 | ✗ | 248 | 2242 |
| SST-2 | 2 | ✓ | 182 | 1639 |

Table 7: Full datasets information.

## A Target Layer Selection

The specific operation of the target layer selection is as follows. Given a sample $(x_i, y_i) \in (X, Y)$ and a prompt template $P$, we prompt LLM to make predictions by connecting $x_i$ and $P$ as inputs. We identify the bias label $\widehat{y_i}$ corresponding to $x_i$ based on the logits by the last layer of the model.

$$\begin{gathered}\widehat{y_i} = argmax \mathcal{M}(c|P(x_i)) \\ \text{where } c \in Y \cap c \neq y_i\end{gathered} \tag{5}$$

Then, we connect $P$ with $y_i$ and $\widehat{y_i}$ in text form, so that we get the gold sample $s_i$ and the biased sample $\widehat{s}_i$. According to this method, we sampled a total of $S$ pairs of samples for analysis, where $i \in S$.

For each pair of samples, we re-entered $s_i$ and $\widehat{s}_i$ into LLM to capture the activation values of the studied component at each layer, recorded as $h_i$ and $\widehat{h}_i$ respectively. Next, we replace the layer by layer while ensuring that the input is still $s_i$, replacing the $h_i$ of a specified layer with the corresponding $\widehat{h}_i$ each time, and using KL divergence to calculate the distribution change of the final output before and after the replacement. Finally, we average the KL divergence of the pair of samples, and the layer $\ell$ where the maximum value appears is the target layer of the operation we are looking for.

$$\ell = argmax_{l \in L} \frac{1}{S} \sum_{i \in S} \mathcal{M}_{h_l}(s_i) log \frac{\mathcal{M}_{h_l}(s_i)}{\mathcal{M}_{\widehat{h}_l}(s_i)} \tag{6}$$

In this implementation, the studied component is focused on the output matrix of the self-attention module, i.e., $o\_proj$.

| Model | Method | TweetEmotion | GoEmotions | EmpathicDialogues | TweetHate | SST-5 |
|---|---|---|---|---|---|---|
| Gemma2 (2B) | Original | 59.56 | 13.75 | 38.73 | 68.15 | 35.18 |
| | CC | 61.55 (+1.99) | 15.95 (+2.20) | 48.82 (+10.09) | 50.22 (-17.93) | 32.22 (-2.96) |
| | DC | 64.63 (+5.07) | 18.03 (+4.28) | 45.93 (+7.20) | 41.53 (-26.62) | 42.35 (+7.17) |
| | CRISPR | 62.47 (+2.91) | 15.63 (+1.88) | 43.27 (+4.54) | 70.63 (+2.48) | 36.19 (+1.01) |
| | LoRA | 74.78 (+15.22) | 51.53 (+37.78) | 59.17 (+20.44) | 15.45 (-52.70) | 56.81 (+21.63) |
| | Full FT | 74.67 (+15.11) | 52.90 (+39.15) | 54.79 (+16.06) | 19.90 (-48.25) | 54.52 (+19.34) |
| | Ours | **75.72 (+16.16)** | **54.55 (+40.80)** | **59.69 (+20.96)** | **72.14 (+3.99)** | **58.52 (+23.34)** |
| Gemma2 (9B) | Original | 60.86 | 21.13 | 39.05 | 64.40 | 39.09 |
| | CC | 64.27 (+3.41) | 22.17 (+1.04) | 48.74 (+9.69) | 67.07 (+2.67) | 36.48 (-2.61) |
| | DC | 67.49 (+6.63) | 22.25 (+1.12) | 46.73 (+7.68) | 44.83 (-19.57) | 47.54 (+8.45) |
| | CRISPR | 60.54 (-0.32) | 22.60 (+1.47) | 38.91 (-0.14) | 68.08 (+3.68) | 41.38 (+2.29) |
| | LoRA | 74.52 (+13.66) | **54.53 (+33.40)** | 60.88 (+21.83) | 15.63 (-48.77) | 59.12 (+20.03) |
| | Full FT | 75.74 (+14.88) | 53.64 (+32.51) | 60.21 (+21.16) | 21.35 (-43.05) | **59.62 (+20.53)** |
| | Ours | **76.21 (+15.35)** | 53.75 (+32.62) | **61.12 (+22.07)** | **71.77 (+7.37)** | 59.45 (+20.36) |
| Mistral (7B) | Original | 59.06 | 13.16 | 34.66 | 35.60 | 34.79 |
| | CC | 62.01 (+2.95) | 21.56 (+8.40) | 49.54 (+14.88) | 25.68 (-9.92) | 38.75 (+3.96) |
| | DC | 63.75 (+4.69) | 15.91 (+2.75) | 50.20 (+15.54) | 19.20 (-16.40) | 32.14 (-2.65) |
| | CRISPR | 55.89 (-3.17) | 12.53 (-0.63) | 35.56 (+0.90) | 22.17 (-13.43) | 29.78 (-5.01) |
| | LoRA | 71.80 (+12.74) | **52.66 (+39.50)** | **62.20 (+27.54)** | 15.44 (-20.16) | 55.68 (+20.89) |
| | Full FT | 72.10 (+13.04) | 52.43 (+39.27) | 61.10 (+26.44) | 18.62 (-16.98) | 54.70 (+19.91) |
| | Ours | **72.57 (+13.51)** | 52.06 (+38.90) | 61.53 (+26.87) | **36.18 (+0.58)** | **56.27 (+21.48)** |
| Llama3 (1B) | Original | 37.16 | 8.43 | 21.04 | 40.32 | 21.44 |
| | CC | 48.97 (+11.81) | 14.31 (+5.88) | 34.88 (+13.84) | 10.54 (-29.78) | 24.43 (+2.99) |
| | DC | 51.56 (+14.40) | 19.92 (+11.49) | 36.36 (+15.32) | 37.80 (-2.52) | 16.66 (-4.78) |
| | CRISPR | 36.92 (-0.24) | 7.48 (-0.95) | 20.79 (-0.25) | 53.81 (+13.49) | 13.88 (-7.56) |
| | LoRA | 73.19 (+36.03) | 48.94 (+40.51) | **58.69 (+37.65)** | 15.06 (-25.26) | **55.13 (+33.69)** |
| | Full FT | **73.77 (+36.61)** | **50.58 (+42.15)** | 56.74 (+35.70) | 53.99 (-13.67) | 53.56 (+32.12) |
| | Ours | 73.69 (+36.53) | 50.05 (+41.62) | 57.17 (+36.13) | **61.14 (+20.82)** | 54.10 (+32.66) |
| Llama3 (3B) | Original | 38.35 | 11.02 | 28.25 | 55.42 | 14.14 |
| | CC | 53.97 (+15.62) | 10.73 (-0.29) | 40.87 (+12.62) | 1.60 (-53.82) | 18.35 (+4.21) |
| | DC | 51.45 (+13.10) | 17.24 (+6.22) | 42.51 (+14.26) | 2.60 (-52.82) | 23.65 (+9.51) |
| | CRISPR | 40.71 (+2.36) | 16.58 (+5.56) | 31.83 (+3.58) | 9.10 (-46.32) | 19.17 (+5.03) |
| | LoRA | 73.71 (+35.36) | 50.71 (+39.69) | **60.86 (+32.61)** | 15.50 (-39.92) | 55.98 (+41.84) |
| | Full FT | **73.99 (+35.64)** | 51.71 (+40.69) | 58.51 (+30.26) | **58.96 (+3.54)** | 51.70 (+37.56) |
| | Ours | 73.26 (+34.91) | **52.61 (+41.59)** | 59.43 (+31.18) | 56.70 (+1.28) | **56.68 (+42.54)** |

Table 8: The main results in the no-instruction setting. The **bold**/underlined font means the best/the second best result.

## B Selection of Fine-tune Parameters

Table 6 presents the impact of unfreezing different parameter combinations on prediction performance during the fine-tuning of Gemma2 (2B). The experiments were conducted on the TweetEmotion dataset. Given the large number of possible combinations, we report results only where one or two parameters were unfrozen.

## C Datasets

The 12 datasets we used are all from the HuggingFace version. There is an extreme label imbalance problem on the fine-grained dataset, which causes that LoRA and full- parameter fine-tuning require more training data to achieve positive improvements. Therefore, in fine-grained tasks, we use a subset of the validation set or training set divided by the original version for training, but ensure that the number of training samples is within 1,000. For coarse-grained tasks, in all implementation methods, we sampled 10% of the test set for training, and the rest for testing. The details are shown in Table 7.

## D Baselines

**CC** (Zhao et al., 2021) and **DC** (Fei et al., 2023) investigated label bias in the few-shot setting. They used the model's output probability of free-text inputs ("N/A" or random token) to adjust the label probability of the original instance. We implemented both methods as described in their original papers.

**CRISPR** (Yang et al., 2024) addressed both label and instruction bias. The method proposed the concept of bias neurons. It identified the neurons that more responsible for bias through gradient-based attribution, and used pruning techniques to modify the weight parameters learned during pre-training. In accordance with the original paper, we sampled 20 instances from the training set to analyze and locate the bias neurons.

**LoRA** (Hu et al., 2021), low-rank adapter fine-tuning, leverages the intrinsic low-rank structure of large language models by introducing a bypass matrix to simulate full-parameter fine-tuning. It is currently one of the most effective and widely used parameter-efficient fine-tuning methods. In our implementation, we utilized the SFTrainer tool from the TRL (Transformers Reinforcement Learning) library developed by HuggingFace. Specifically, we set $k = 8$, $target\_modules = [\text{``}q\_proj\text{''}, \text{``}o\_proj\text{''}, \text{``}k\_proj\text{''}, \text{``}v\_proj\text{''}, \text{``}gate\_proj\text{''}, \text{``}up\_proj\text{''}, \text{``}down\_proj\text{''}]$.

**Full-parameter fine-tuning**, in contrast, involves adjusting all parameters of the language model during training, which requires significantly more computational resources compared to efficient parameter fine-tuning methods. For our experiments, we employed the Trainer tool from the HuggingFace transformers library.

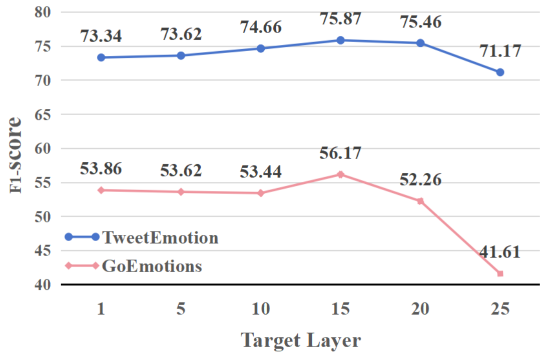
## E Results without Instruction

Table 8 shows the weighted F1 scores of different methods on five fine-grained datasets with no-instruction setting. Our method achieves better results especially on the Gemma2 series models.

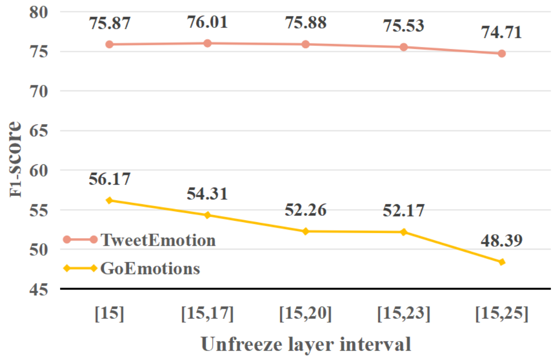
## F Parameter Analysis

First, we conducted parameter analysis experiments on Gemma2 (2B) model to explore the impact of target layer selection and the number of layers trained on task performance. As illustrated in Figure 4a, when the target layer is located in the intermediate layers, task performance exhibits a small peak. However, as the number of layers selected for training increases, performance drops rapidly. In Figure 4b, we present the effect of unfreezing the components $o\_proj$ and $down\_proj$ in layers which after the target. For the TweetEmotion dataset, training the five layers immediately following the target layer has minimal impact on the F1 score, with a slight decline observed thereafter. In contrast, for the GoEmotions dataset, additional training does not yield any performance improvement; instead, it results in a substantial decrease in the F1 score.

Then, we performed several experiments to determine the value of the hyperparameters $\alpha$, $\beta$, $\gamma$. The results are shown in Figure 5.




(a) The impact of target layer selection on performance.




(b) The impact of the number of training layers on performance.

Figure 4: The results of parameter analysis.

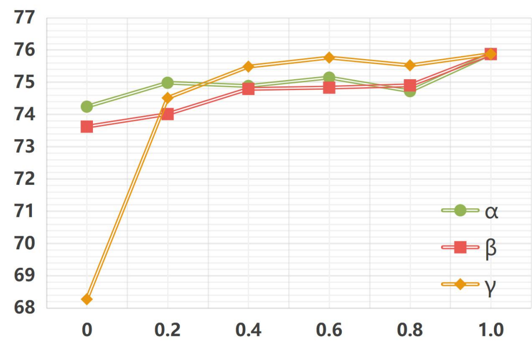
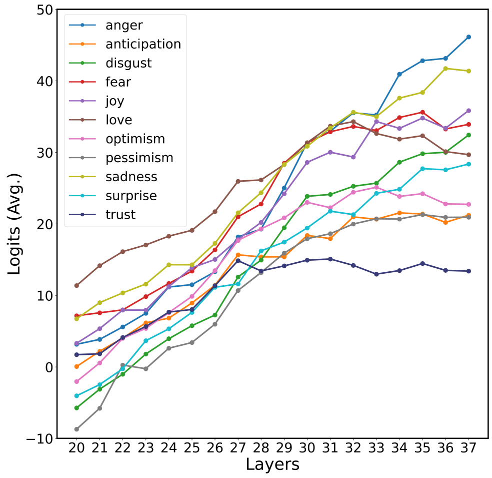


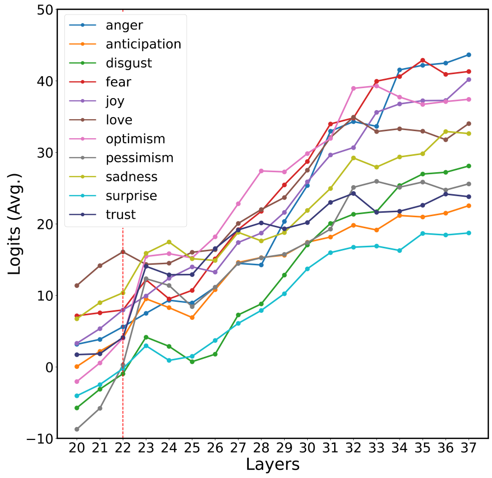

Figure 5: Hyperparameter analysis.

## G Visualisation

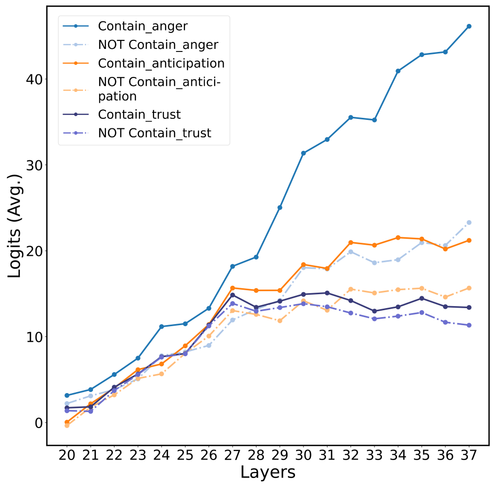
Figures 6 (a-d) illustrate the impact of fine-tuning the Gemma2 (9B) model with our method on label bias mitigation. The results demonstrate a significant improvement in the model's output logits and its ability to discriminate low-frequency labels, with a notable reduction in the gap between high-frequency and low-frequency labels.



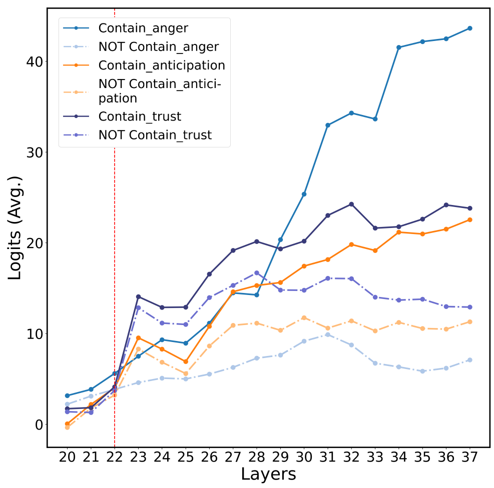

(a) $Contain$ on the original model.




(b) $Contain$ on the fine-tuned model using our method.




(c) Distinguish ability of specific labels on the original model.




(d) Distinguish ability of specific labels on the fine-tuned model using our method.

Figure 6: (a-b) compare the $Contain$ of each label and (c-d) compare the distinguish ability of specific labels on the Gemma2 (9B) model before and after correction.

## H Templates

In Table 9, we show all the templates used in our experiments and the corresponding label names of the datasets.

| Datasets | Template | Label Name |
|---|---|---|
| TweetEmotion<br>(Mohammad et al., 2018) | Review: [review]<br>Emotion: [label] | anger, anticipation, disgust,<br>fear, joy, love, optimism,<br>pessimism, sadness, surprise,<br>trust |
| | Instruction: Select the right emotion words for the given Review from Choices.<br>Choices: [options]<br>Review: [review]<br>Emotion: [label] | |
| GoEmotions<br>(Demszky et al., 2020) | Review: [review]<br>Emotion: [label] | admiration, amusement, anger, annoyance,<br>approval, caring, confusion, curiosity,<br>desire, disappointment, disapproval, disgust,<br>embarrassment, excitement, fear, gratitude,<br>grief, joy, love, nervousness, optimism,<br>pride, realization, relief, remorse,<br>sadness, surprise, neutral |
| | Instruction: Select the right emotion words for the given Review from Choices.<br>Choices: [options]<br>Review: [review]<br>Emotion: [label] | |
| Empathic<br>Dialogues<br>(Rashkin et al., 2019) | Utterance: [utterance]<br>Emotion: [label] | afraid, angry, annoyed, ashamed, anticipating,<br>anxious, apprehensive, confident, caring, content,<br>disappointed, disgusted, devastated, embarrassed,<br>excited, faithful, furious, grateful, guilty,<br>hopeful, impressed, jealous, joyful, lonely,<br>nostalgic, proud, prepared, sentimental,<br>sad, surprised, terrified, trusting |
| | Instruction: Select the right emotion word for the given Utterance from Choices.<br>Choices: [options]<br>Utterance: [utterance]<br>Emotion: [label] | |
| TweetHate<br>(Sachdeva et al., 2022) | Review: [review]<br>Emotion: [label] | hate_gender, hate_race, hate_sexuality,<br>hate_religion, hate_origin, hate_disability,<br>hate_age, not_hate |
| | Instruction: Select the right emotion label for the given Review from Choices.<br>Choices: [options]<br>Review: [review]<br>Answer: [label] | |
| SST-5<br>(Socher et al., 2013) | Review: [review]<br>Sentiment: [label] | very positive, positive, neutral, negative,<br>very negative |
| | Instruction: Select the right sentiment label for the given Review from Choices.<br>Choices: [options]<br>Review: [review]<br>Answer: [label] | |
| SST-2<br>(Socher et al., 2013) | Instruction: Select the right sentiment word for the given Review from Choices.<br>Choices: [options]<br>Review: [review]<br>Emotion: [label] | positive, negative |
| BBQ-Age<br>(Parrish et al., 2022) | Instruction: For the subsequent context and question, decide on the most appropriate answer from the given options.<br>Context: [context]<br>Question: [question]<br>Options: [option]<br>Answer:[label] | The young adult, The sixty-something,<br>The middle-aged woman, The high schooler,<br>The elderly man, Can't answer, Undetermined, ... |
| BBQ-SES<br>(Parrish et al., 2022) | | The judge, The psychologist, The bartender,<br>The low-income students, The farm worker,<br>The physician, Karen, Not enough info, ... |
| BBQ-Disability<br>(Parrish et al., 2022) | | The friend who's deaf, The person with PTSD,<br>The disabled woman, The person with anxiety,<br>The blind, Not known, Can't be determined, ... |
| BBQ-Gender<br>(Parrish et al., 2022) | | The boy, The girl, The cisgender woman,<br>The transgender man, Steven, Carol,<br>The cis woman, Not enough information,... |
| AGNews<br>(Zhang et al., 2015) | Instruction: Instruction: Classify the following news article from the given Choices.<br>Choices: [option]<br>Text: [news]<br>Category: [label] | world, sports, business, technology |
| RTE<br>(Dagan et al., 2006) | Instruction: Context: [sentence1]<br>Question: [sentence2] True or False?<br>Answer: [label] | True, False |

Table 9: Template of all 12 datasets used in our experiments.