Fine-tuning LLaDA-8B-Instruct for Improved Question Answering and Bias Mitigation using Reward-Guided LoRA

Marc Saouda Prateek Gautam Phillip Nelson

June 27, 2025

Abstract

We investigated the impact of fine-tuning a LLaDA-8B-Instruct model using a Low-Rank Adaptation (LoRA) adapter on multiple-choice question (MCQ) data, aiming to reduce gender bias and improve answer quality. Our experiments, conducted with varying dataset sizes (1k, 2k, 3k, 4k) and temperature settings (0, 0.7, 1) for generation, demonstrate that this approach leads to improved performance in both open-ended and MCQ question answering. The fine-tuning process, framed within a REINFORCE-based reinforcement learning (RL) framework, utilized a reward signal combining a bias metric (GenBitMetrics) and a semantic preservation term (KL divergence) to guide the LoRA adapter. Notably, the generated answers exhibited a reduced use of pronouns, particularly gendered ones, enhancing clarity, coherence, and neutrality. The fine-tuning process proved to be cost-effective and rapid, yielding significant improvements in generating semantically equivalent but less biased text.

1 Introduction

Large language models (LLMs) are known to reflect societal biases present in their training data, resulting in generated outputs that may perpetuate stereotypes or offensive content. Existing solutions often rely on post-hoc filtering or fine-tuning, which can degrade fluency or fail to generalize. This report details our fine-tuning approach, which leverages a LoRA adapter on the LLaDA-8B-Instruct model, using MCQ data to mitigate gender bias and enhance question-answering capabilities. We analyze the resulting improvements, focusing on the reduction of pronoun usage and overall answer quality. The core computational task is the generation of semantically equivalent but less biased versions of input text, formulated as a conditional text generation problem solved via reinforcement learning.

2 Methods

2.1 Model and Fine-tuning Architecture

We employed the LLaDA-8B-Instruct model as our base LLM. To adapt this model for bias mitigation, we utilized Low-Rank Adaptation (LoRA). LoRA fine-tuning allows for efficient adaptation of large models by training only a small number of adapter parameters (target modules: q_proj, k_proj, v_proj, $attn\_out$) while keeping the pre-trained model weights frozen. This approach significantly reduces computational overhead. The LoRA adapter, configured with parameters like r (rank) and $lora\_alpha$, acts as a lightweight, updatable policy network.

2.2 Dataset and Task Formulation

The fine-tuning data consisted of subsets (1k, 2k, 3k, 4k samples) drawn from approximately 150k multiple-choice prompts. These prompts, generated using Expected Parrot, feature a combination of gendered and gender-neutral answer choices. The task was formulated as conditional text generation: given a structured multi-choice prompt potentially containing gender bias, the model's goal is to generate or favor gender-neutral options. This was framed within a reinforcement learning (RL) paradigm.

2.3 Training Process and Reinforcement Learning Framework

We implement a REINFORCE-style algorithm to fine-tune only the LoRA adapter, holding the base model frozen and setting $cfg\_scale = 0$ throughout to isolate the effect of parameter updates.

1. **Text generation (no CFG):**
   - Input prompts *input_ids* (length L_p) are padded to (L_p + *gen_length*) with a special mask token.
   - We denoise in *steps* iterations (e.g. 128), in *block_length* segments (e.g. 32 tokens per block).
   - At each iteration, we compute logits $\ell \in \mathbb{R}^{B\times T\times V}$ from the LoRA model.
   - To encourage exploration, we add Gumbel noise at temperature $T > 0$:

$$\tilde{\ell} = \exp(\ell)/\big(-\ln(U)\big)^T, \quad U \sim \mathrm{Uniform}(0,1).$$

   - We then select the next token by $\arg\max$ over $\tilde{\ell}$ for each masked position.
   - A "low_confidence" remasking strategy re-masks the least confident predictions for the next iteration.
2. **Reward computation:** For each generated sequence in a microbatch of size M:
   - Decode the generated tokens to text and compute its gender bias

$$bias = \big|\mathrm{GenBitMetrics}(text)\big|.$$

   - Extract the LoRA model's logits ℓ^{lora} and the base model's logits ℓ^{base} on the *same* generated sequence.
   - Compute token-wise KL divergence and average over generated positions:

$$KL = \frac{1}{L_g}\sum_{t=1}^{L_g}\sum_{v=1}^{V} p_{t,v}^{\mathrm{lora}}\,\ln\frac{p_{t,v}^{\mathrm{lora}}}{p_{t,v}^{\mathrm{base}}},$$

     where L_g is the generated length.
   - Combine into a scalar reward:

$$R = -\big(bias + \lambda_{KL}\cdot KL\big), \quad \lambda_{KL} = 0.1.$$

3. **Policy (LoRA) update:**
   - For each microbatch, sum the log-probabilities of the generated tokens under the LoRA model:

$$\log p = \sum_{t=1}^{L_g} \ln p^{\mathrm{lora}}(x_t \mid x_{<t}).$$

   - Form the REINFORCE loss

$$\mathcal{L} = -\mathbb{E}\big[\log p\cdot R\big], \quad \text{approximated by } \tfrac{1}{M}\sum_{i=1}^{M} -\log p_i\, R_i.$$

   - Backpropagate only through the LoRA adapter parameters.
   - Apply AdamW with learning rate 2×10^{-5} and gradient accumulation over microbatches.

Forward perturbations via random token masking during the iterative generation process were intended to simulate partial observability, enhancing policy generalization under uncertainty. The implementation leveraged Huggingface Transformers and PEFT libraries.

3 Results

The fine-tuned models showed a notable improvement in generating answers with less pronoun usage, particularly a reduction in gender-specific pronouns, across all dataset sizes and temperature settings. This directly contributed to mitigating gender bias in the outputs, as measured by GenBitMetrics. Both open-ended and MCQ answers benefited from this, resulting in more precise, coherent, and neutral responses. The KL divergence component of the reward helped maintain semantic coherence with the original intent of the prompts.

4 Discussion

The reduction in pronoun usage, especially gendered pronouns, indicates that fine-tuning on MCQ data with a reward function directly penalizing bias (via GenBitMetrics) and semantic drift (via KL divergence) successfully encourages the model to produce more direct, explicit, and neutral answers. This RL-based approach allows the model to learn a policy that balances these two objectives. The efficiency of LoRA makes this a practical approach for targeted improvements in LLMs. While the reward function effectively guided the model, it is acknowledged as a proxy for real-world bias, and future work could explore more sophisticated reward models.

5 Examples

5.1 Example 1

Question: [Insert Question Here]
Before Fine-tuning: [Insert Answer Here]
After Fine-tuning: [Insert Answer Here]

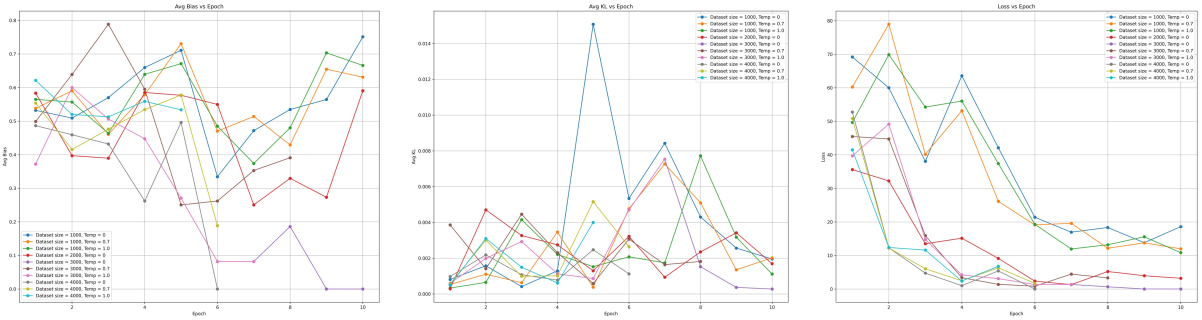
5.2 Example 2

Question: [Insert Question Here]
Before Fine-tuning: [Insert Answer Here]
After Fine-tuning: [Insert Answer Here]

5.3 Example 3

Question: [Insert Question Here]
Before Fine-tuning: [Insert Answer Here]
After Fine-tuning: [Insert Answer Here]

6 Plots



(a) Losses During Fine-tuning (b) Bias Distributions (c) KL Divergence Distributions

Figure 1: Overview of key metrics during and after fine-tuning: loss trends, bias distribution, and KL divergence.

7 Conclusion

Our experiments demonstrate that fine-tuning LLaDA-8B-Instruct with a LoRA adapter, guided by a REINFORCE-based RL algorithm with a multi-objective reward function (GenBitMetrics for bias and KL divergence for semantic preservation), is an effective strategy for mitigating gender bias and improving question-answering performance. The resulting models exhibit reduced pronoun usage and generate higher-quality, more neutral, and semantically consistent answers. The speed and cost-effectiveness of this approach make it a valuable tool for adapting language models to specific tasks and ethical considerations.