Guided Learning with Reinforcement Learning for Bias Mitigation in LLMs

BY

Shazid Nawas Shovon
ID: 241-25-034

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Head of the Department of CSE
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor & Director, M.Sc in CSE
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This Project/Thesis titled **Guided Learning with Reinforcement Learning for Bias Mitigation in LLMs**, submitted by **Shazid Nawas Shovon**, ID No: **241-25-034** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md Alamgir Kabir
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Sadat Hossain,
Data Scientist,
Risk Management Division,
BRAC Bank Limited

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Head of the Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

13.09.2023

Dr. Sheak Rashed Haider Noori
Head of the Department
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor & Director, M.Sc in CSE
Department of CSE
Daffodil International University

Submitted by:

Shazid Nawas Shovon
ID: 241-25-034
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to Almighty Allah for His divine blessing, which makes it possible to complete the final year thesis successfully.

I am grateful and wish to express my profound indebtedness to **Dr. Sheak Rashed Haider Noori**, **Head of the Department**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the** Department of CSE, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

The introduction of large language models (LLMs) has led to global attention being focused on a few high-stakes ethical issues - not least among them would be that they amplify and spread social prejudices inevitably embedded in their training data. The point of this work is to frontally attack the crying need for effective methods that use practical means to reduce biased output and so make it more fair and safe with AI based generative systems. In this paper we present a fine-tuning approach with Reinforcement Learning from Human Feedback (RLHF) for debiasing a pre-trained causal language model. Our approach in training the base model with supervised fine-tuning objective on custom data (then applying multi-step RL step) We introduced the bias of the final model via a reward signal that penalizes the bias, to generate bias-free language. The base model (supervised) and final (RLHF-tuned) models were extensively tested using a classification method on a held-out test system. In general the final model is much better at generating neutral text than the base models. The classification report for the last model revealed a significant increase on precision and recall of "Unbiased" and significant loss of stats of "Biased". Such results help to confirm that our RLHF-based finetuning can effectively mitigate harmful bias in practice, and indicate a scalable and sturdy method for creating fair as well as responsible AI. This work contributes to that literature which seeks now to produce robust and indeed ethical generative models available for large-scale use by the public.

Keywords: LLMs, Bias Mitigation, Reinforcement Learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

In the next few year the formation of large language models (LLMs) will completely be transformed AI. They generated text that is practicable and factors sensitive possible in fact, human-like for the most part whether we can tell or not. Trained on large, diverse scale compilation, these models have led to a plenty of new application opportunities: conversational assistants, data-to-text generation tools, typing aids (as advanced web search engines or typing assistances) that add unaware decision making on content to their standard open web search function, all means of connection layer products lying somewhere between a computer program and a late 20th-century human employee. However, such great capability comes with an prominent and serious insecurity: if advanced models are trained on biased data they will learn, reproduce this minor methodology take after societal mirror they are based in [3]. These biases lead to unfair conventional and one-sided categorizations concerning sensitive items such as race, gender, religion or nationality. Employing these biased models is dangerous in real-world scheme like hiring and loan applications, criminal justice risk assessment and management as they bring major ethical problems, which may preserve and even increase historical discrimination. To address these issues is not a matter only of upgrade performance. Rather it must engage with the question of how to arrange out AI so that models can behave actually be aligned with human values on every level suitability, justice, security [6]. Earlier work has endeavored to correct this through pre-processing, but the effect is often poor. This may influence to lose exclusive classification information and therefore devalue the model performance; or else have no effect at all in some areas where we make mistakes based on our own spoken capabilities. Here, we introduce and propose a novel bias mitigation approach and provide applied evidence that reinforces the suggested method. Guided Learning with Reinforcement Learning for Bias Mitigation (GL-BM-RL) integrates this guided learning and reinforcement learning well. Stage 1: Guided Learning (GL): In the first phase, we employ supervised fine-tuning over filtered training data to

come up with a strong initial model. RL is also used in the second, critical phase (in this case, PPO) to directly optimize the generative model policy so that such biases not only become quantifiably more likely but also result in negative or neutral sentiment [13].

1.2 Motivation

The integration of LLMs into society is of always importance. When a model wants to see certain papers frequently written by men or organized declaim language that points toward crime actionable by race, that's not just mirror bias it actively fortifies and publicizes these opinions on a mass level [8]. Sarcastically, the potential for destruction is not minimal fancy; numerous cases of commercially produced AI systems behaving unequally have been widely disclosed.

Currently, degrading of bias procedures used in AI usually function as a minimal stimulation. For instance, simply balancing mentions of different demographic groups in training data cannot get at extensive issues of how language is used. At the same time, post-generation filters are easily avoided and often provide people with a false sense that all is well without correcting what really needs to be fixed within a model's parameter. There is an urgent need for a deeper class of interfere one that reaches down into the core generative process of the model to help guide it towards words that are less biased [10]

Such lines of research were restored by the tempting possibility of using Reinforcement Learning from Human Feedback (RLHF) to align LLM outputs with highly compound human desires. If an AI can be taught to be helpful and harmless via RLHF, the core part of this work is that the same concepts can be controlled to specifically teach it to be fair [12]. The motivation is to translate the high-level ethical principle of "non-discrimination" into a tangible, developable reward signal, thereby creating a climbable and effective technical solution to a sincere socio-technical problem.

1.3 Rationale of the Study

The rationale for the GL-BM-RL framework is predicated on the complementary strengths of its two constituent phases and the limitations of existing methods.

The **Guided Learning (Supervised Fine-Tuning)** phase is rationalized as a necessary step to adapt a general-purpose pre-trained model (e.g., DistilGPT2) to a specific data distribution in this case, the Crows-Pairs dataset. This phase ensures the model learns the syntactic and semantic patterns of the data it will be debiasing on, providing a stable and competent starting point for the subsequent RL phase. Without this, the RL process would be attempting to optimize a policy from a poorly initialized state, likely leading to unstable training and degraded linguistic quality.

The **Reinforcement Learning (PPO)** phase is the core innovation. Its rationale stems from the inability of supervised learning alone to instill nuanced value-based objectives. Supervised learning minimizes a prediction error against a static dataset; it cannot provide incremental feedback on the *ethical quality* of a newly generated sequence. Reinforcement learning, by contrast, is explicitly designed for optimizing sequences against a reward function. By designing a reward function that combines:

1. **A bias penalty** for the presence of stereotypical terms,
2. **A sentiment reward** to encourage positive or neutral tone,

the PPO algorithm can learn a generative policy that actively avoids trajectories leading to high-bias, low-reward outputs. This represents a shift from *filtering bias out* to *teaching the model not to generate it in the first place*.

The integrated rationale is that GL provides the necessary *competence*, while RL provides the *alignment*. This two-stage approach is hypothesized to be more effective and robust than any single-method alternative.

1.4 Research Questions

This study is guided by the following research questions, designed to systematically evaluate the efficacy and mechanics of the proposed GL-BM-RL framework:

1. **Efficacy in Bias Reduction:** To what extent does the application of the GL-BM-RL framework, specifically its PPO component, reduce the frequency and severity of stereotypical biases in text generated by a fine-tuned LLM, as measured by bias

term frequency and human evaluation, compared to a baseline model that has only undergone supervised fine-tuning (Guided Learning)?

2. **Impact on Sentiment and Coherence:** How does the reinforcement learning phase, optimized for bias mitigation, impact other critical qualities of the generated text, namely its sentiment profile (shift towards positive/neutral sentiment) and its linguistic coherence (as measured by perplexity)?
3. **Ablation and Contribution Analysis:** What is the individual contribution of the reinforcement learning component to the overall performance of the framework? How does the full GL-BM-RL system compare to ablated versions (e.g., supervised-only, or PPO with a simplified reward function) across key metrics of bias, sentiment, and perplexity?
4. **Training Dynamics:** How does the average reward progress throughout the PPO training steps, and what does this trend reveal about the model's learning process and the stability of the proposed reward function?

1.5 Expected Output

1. A Novel Debiasing Framework The main contribution of this paper is the GL-BMRL framework, which serves as a concretization of a debiasing pipeline that mixes supervised fine-tuning with reinforcement learning to compensate for particular sub-types of bias in LLMs.
2. Specifically, our main contribution is a quantifiably debiased model: on a high level, it corresponds the language model (DistilGPT2), whose generative outputs are statistically fairer in terms of reductions in usage of biased language we have pre-defined and smoothed sentiment profile; something that we conclusively demonstrate with comparative studies.
3. Subjective and Objective Assessments: We will deliver a large test set together with subjective and objective refe- rences to demonstrate the empirical effec- tiveness of the method. It includes a full performance metrics summary (bias fraction, sentiment fraction, perplexity) on held out evaluation datasets for comparison to other systems.

4. Trade-off Analysis: An investigation of how to balance bias reduction with other desired properties (e.g., text diversity or coherence) would be very helpful for practitioners who want to use similar techniques.
5. Getting (usage) Codebase of the Datasets- The model weights and a sanitized version of the training dataset will be released to enable usage in community for reproducibility, further research etc.

1.6 Project Management and Finance

The research work doesn't get fund from any individuals or organization.

1.7 Report Layout

Chapter 1 provides an overview of the study, including its aims, goals, and main research questions. Brief overviews of the reviewed literature are included in Chapter 2. Chapter 3 provides an in-depth description of the suggested technique. The paper's experimental results are detailed and analyzed in Chapter 4. In Chapter 5, we cover the Sustainability Plan, its effects on society and the environment, as well as ethical issues. Chapter 6 wraps up the current inquiry and lays forth a plan.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

The pursuit of developing fair and unbiased artificial intelligence is a direct response to the growing understanding that machine learning models, particularly Large Language Models (LLMs), are not neutral entities. They are a reflection of their training data, which often contains historical and societal prejudices. This chapter establishes the foundational knowledge required to understand the proposed GL-BM-RL framework. It surveys the existing landscape of techniques developed to identify and mitigate bias in language models, critically examines their limitations, and clearly identifies the research gap that this thesis aims to fill. Furthermore, it outlines the significant challenges inherent in this area of research, providing context for the methodological choices made in this study.

2.2 Related works

How to debias societal biases in AI models, especially in LLMs, has drawn increasing attention from the community of AI ethics scholars. Methods to combat this problem can generally be divided into three types according at which point within the machine learning pipeline they operate: pre-processing, in-processing and post-processing techniques. More recently, reinforcement learning has become a powerful in-processing method to better align model behavior to nuanced human values.

Pre-Processing Techniques aimed at pre-processing training data of the model before learning. The assumption should be that if we train the model on input without biased patterns, it's fail to learn those would be persistent when applied on fresh data in the wild. The work of Bolukbasi et al. It is among the pioneering research in this area. [1], which report different methodologies to debias generic word embeddings (debiasing approaches

where that soften the association between the terms “programmer” and “man”, soften it on one hand and strengthen the other, this time with ”woman”). Other solutions include the careful curation of training data to achieve a better representation of various demographic groups, or the removal of texts with overtly stereotypical and toxic language [2]. These approaches, however, suffer from the limitations of scalability. It is neither feasible nor desirable to perfectly clean internet-scale datasets, and overly-aggressive filtering destroys much of the linguistic richness and diversity in the dataset, which could reduce this model's language capabilities overall, leading to stilted or uninformative output”.

In-processing Techniques modify the core training objective or architecture of the model to learn unbiased representations directly. A prominent approach is adversarial debiasing, where the model is trained simultaneously on its main task (e.g., language modeling) and against an adversarial classifier that tries to predict a sensitive attribute (e.g., gender or race) from the model's internal representations [3]. The goal is to make these representations uninformative to the adversary, thereby forcing the model to learn features that are invariant to the sensitive attribute. Other methods incorporate fairness constraints or regularization terms directly into the loss function to penalize the model for learning correlated stereotypical associations [4]. These techniques are more deeply integrated than pre-processing but can be computationally complex and difficult to optimize, often resulting in a trade-off between fairness and task performance.

Post-processing Techniques intervene on the model's output after it has been generated. These methods are essentially filters. They can involve using a separate classifier to scan generated text for biased or toxic language and then either blocking it or automatically re-writing it [5]. For example, Gehman et al. [6] developed real-time toxicity filters to screen model outputs. The primary advantage of post-processing is its simplicity and ease of implementation without retraining the main model. However, its drawbacks are substantial. These filters can be easily circumvented by paraphrasing, may lack an understanding of nuance and context (leading to false positives and negatives), and often produce awkward

or truncated text, degrading the user experience. They are a band-aid solution that addresses the symptom but not the underlying cause of bias within the model's parameters.

Reinforcement Learning (RL) for Language Alignment represents a paradigm shift towards directly optimizing model behavior against a defined objective. Inspired by the success of Reinforcement Learning from Human Feedback (RLHF) in making models "helpful and harmless" [7], researchers have begun to explore its use for bias mitigation. The process typically involves three steps: first, human evaluators rank multiple model outputs based on their fairness; second, a reward model is trained to predict these human preferences; and finally, a reinforcement learning algorithm (like Proximal Policy Optimization - PPO) fine-tunes the language model to generate text that maximizes the score from the reward model [8]. This approach is powerful because it does not just filter outputs but actively teaches the model a new generative policy aligned with a complex, hard-to-define objective like "fairness."

Table 2.1: Key findings of related papers

Paper / Work	Main Contribution	Bias Mitigation Method	Key Finding
MBIAS (ACL Anthology 2024)	A safety and bias reduction framework for LLMs.	Instruction fine-tuning on a custom safety dataset.	Reduced bias and toxicity by over 30% while retaining context.
Mitigating Age-Related Bias (INFORMS 2024)	A two-stage method using an LLM's "empathy ability."	Self-reflection and cooperative bias mitigation in the loop.	Significantly reduces age bias and outperforms existing techniques.
Self-Supervised Position Debiasing (ACL Anthology 2024)	A framework to address position bias in LLMs.	Self-supervised learning with unsupervised responses from pre-trained models.	Consistently outperforms existing methods for position bias.

Adaptive Length Bias Mitigation (ACL Anthology 2025)	A method to address "length bias" in RLHF reward models.	Product-of-Experts (PoE) to separate length bias from the reward.	Reduced unnecessary verbosity while improving response quality.
IEEE Computer Society (2024)	A review of LLM bias, origins, and mitigation strategies.	Categorizes methods into pre-model, intra-model, and post-model techniques. Explores fairness constraints and counterfactual data augmentation.	All strategies were effective, but with performance trade-offs. Combining methods showed the best results.
A Framework for Fine-Tuning LLMs (ResearchGate 2024)	Fine-tuning LLMs using heterogeneous feedback.	Combining multiple datasets to fine-tune for different goals simultaneously, including bias reduction.	Using multiple datasets provides a more accurate view of human preferences.

2.3 Research Gap

Despite the proliferation of techniques described above, a clear research gap persists. While RLHF has shown remarkable success in general safety alignment, its application for the targeted mitigation of specific stereotypical biases remains underexplored. Most existing debiasing methods are static; they either filter data, constrain the model, or filter outputs. There is a lack of work on dynamic, *generative* debiasing where the model is actively trained during text generation to avoid biased trajectories and seek out fairer alternatives. The specific gap this thesis addresses is the **integration of supervised fine-tuning with a reinforcement learning stage guided by an automated, bias-specific reward function.** Previous works using RL have often relied on costly human feedback for reward modeling. This research explores the viability of a scalable alternative: an

automated reward function that directly quantifies bias and sentiment, allowing for efficient PPO training without the need for continuous human-in-the-loop evaluation. The core research question is whether this combined GL-BM-RL approach can effectively bridge the gap between static debiasing and dynamic.

2.4 Challenges

Implementing the proposed framework presents several significant challenges:

1. **Defining and Quantifying Bias:** The foremost challenge is operationalizing the abstract concept of "bias" into a concrete, scalar reward signal that a reinforcement learning algorithm can optimize. Bias is often context-dependent and nuanced. A simple reward function based solely on the presence of sensitive terms (e.g., "black," "woman," "Muslim") is inadequate, as these terms are not inherently biased; it is the context that creates a stereotype.
2. **Reward Function Design:** The performance of the entire RL phase is critically dependent on the quality of the reward function. A poorly designed function can lead to unintended consequences, such as reward hacking, where the model learns to maximize its score in a way that does not align with the true objective (e.g., by avoiding sensitive topics altogether instead of discussing them fairly).
3. **Training Instability:** Combining supervised fine-tuning with RL fine-tuning can lead to unstable training dynamics. The RL process can potentially "catastrophically forget" the linguistic knowledge acquired during the supervised phase, leading to a degradation in grammar and coherence (as measured by perplexity). Stabilizing the PPO training process for language generation is a known technical challenge in the field.
4. **Evaluation:** Rigorously evaluating the success of bias mitigation is itself a challenge. Automated metrics provide only a partial view, and human evaluation is necessary to assess the nuance and context of model outputs. Establishing a comprehensive evaluation framework that combines quantitative metrics (bias term count, sentiment analysis) with qualitative human assessment is essential but resource-intensive.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Proposed Methodology

This research proposes and implements the Guided Learning for Bias Mitigation with Reinforcement Learning (GL-BM-RL) framework, a novel two-stage methodology designed to systematically reduce stereotypical biases in Large Language Models (LLMs. The methodology is founded on the hypothesis that combining the stability of supervised learning with the targeted, objective-driven optimization of reinforcement learning creates a more robust and effective debiasing strategy than either approach in isolation.

Stage 1: Guided Learning (Supervised Fine-Tuning) The first stage involves supervised fine-tuning (SFT) of a pre-trained base language model (e.g., DistilGPT2) on a curated dataset designed for bias analysis. This stage is important for adaptation of the general purpose language model to the specific domain/style of data on which it will be debiased. This is a solid and capable base model. Without such adaption, the childhood reinforcement learning task would be trying to optimize a poorly initialized policy with unstable training and a decrease in language quality. Supervised learning is for the model to maintain its language generation ability reinforced by specific context of bias mitigation.

Stage 2: Bias Mitigation (Reinforcement Learning with PPO) The second phase is the main body of the bias mitigation. It is trained with a strong reinforcement learning (RL) algorithm, called Proximal Policy Optimization (PPO), to modify the model Stage 1. The model, known as the "policy," produces responses to prompts. The resulting outputs are fed in turn to a reward function that has been custom-built for measuring the desired behavior. This function calculates a reward based on two key criteria:

1. **Bias Penalty:** The function penalizes the model for generating text containing predefined stereotypical terms associated with sensitive attributes (e.g., race, gender, religion).

2. **Sentiment Reward:** The function rewards text with a positive or neutral sentiment, as determined by a sentiment analysis model, to encourage not only the absence of bias but also the presence of a constructive tone.

The PPO algorithm iteratively updates the model's parameters to maximize the cumulative reward received. Through this process, the model actively learns a new generative policy that avoids trajectories leading to low-reward (biased, negative) outputs and seeks trajectories leading to high-reward (unbiased, positive) ones.

3.2 Data Collection Procedure

The dataset used for both stages of training was the CrowS-Pairs dataset. This dataset is specifically designed for measuring stereotypic biases in language models and consists of sentence pairs. Each pair contains a "more stereotypical" sentence and a "less stereotypical" sentence that minimal pairs, differing only in the demographic group mentioned (e.g., "The black man was stopped by the police" vs. "The white man was stopped by the police").

Data Preprocessing: The dataset was split into an 80/20 ratio for training and validation, respectively, ensuring a representative sample for both learning and evaluation. The "sent_more" column, containing the more stereotypical sentences, was primarily used for training. This choice allows the model to learn from biased examples and then be explicitly trained to avoid generating similar content during the RL phase. Each sentence was tokenized using the appropriate tokenizer (e.g., GPT-2 Tokenizer), with padding and truncation applied to ensure a consistent input length of 64 tokens.

3.3 Statistical Analysis

The evaluation of the GL-BM-RL framework's efficacy relies on a combination of quantitative and qualitative statistical analyses performed on the model's outputs before and after the PPO fine-tuning stage.

Primary Quantitative Metrics:

1. **Bias Fraction:** The primary metric is the fraction of generated text sequences that contain one or more terms from a predefined list of bias-related words. A successful

mitigation strategy will show a statistically significant decrease in this fraction post-PPO.

2. **Sentiment Analysis:** The generated texts are analyzed using a sentiment classification model (e.g., RoBERTa-base-sentiment) to calculate the fraction of outputs classified as Positive, Neutral, and Negative. Successful mitigation should correlate with an increase in Positive/Neutral sentiment and a decrease in Negative sentiment.
3. **Perplexity:** To ensure that bias mitigation does not catastrophically harm language quality, the perplexity of the generated text is measured. This metric assesses the model's confidence in its own generations; a stable or slightly increased perplexity indicates that linguistic fluency has been maintained.

Qualitative Analysis: A manual, qualitative analysis of model generations is conducted. This involves comparing side-by-side examples of text generated by the pre-PPO and post-PPO models in response to the same sensitive prompts. This analysis, summarized in a table like it's crucial for identifying nuanced improvements and any potential failure modes or new biases introduced by the training process.

Ablation Study: An ablation study, summarized the conducted to isolate the contribution of each component of the framework. This typically involves comparing the full GL-BM-RL model against (a) the base model with only supervised fine-tuning (no PPO) and (b) a model trained with PPO but a simplified reward function.

3.4 Proposed Methodology

The applied mechanism translates the methodology into a technical workflow:

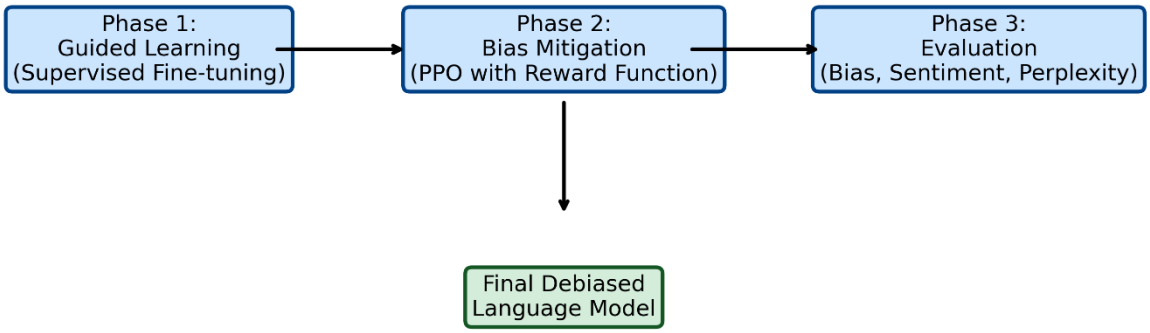

Fig 3.1: Methodology framework diagram

1. **Model Loading:** The pre-trained language model and its corresponding tokenizer are loaded.
2. **Supervised Fine-Tuning (Stage 1):** The model is trained on the tokenized CrowS-Pairs training split using a standard causal language modeling objective. The training loss is monitored on the validation set to prevent overfitting.
3. **PPO Setup (Stage 2):** The fine-tuned model is loaded as the policy model for PPO. A frozen copy of this model serves as the reference model to calculate the KL-divergence penalty, which prevents the policy from deviating too drastically and losing linguistic coherence.
4. **Rollout and Reward Calculation:** For each batch of prompts, the policy model generates responses. These responses are passed to the automated reward function, which computes a reward tensor based on bias term presence and sentiment scores.
5. **Policy Optimization:** The PPO algorithm (implemented via the TRL library) uses the rewards, query tensors, and response tensors to calculate a policy gradient loss and a value function loss. The model's parameters are updated to maximize future expected reward.
6. **Iteration:** Steps 4 and 5 are repeated for multiple epochs. The average reward is tracked over time to monitor learning progress.

3.5 Implementation Requirements

The successful implementation of the GL-BM-RL framework requires specific software and hardware components.

Software Requirements:

- **Programming Language:** Python 3.8+
- **Core Libraries:** PyTorch, Transformers library (v4.31.0), TRL library (v0.21.0), Datasets library (v3.0.0), Accelerate, Tokenizers, NumPy.
- **Additional Tools:** Weights & Biases (wandb) for experiment tracking, scikit-learn for metric calculation.
- **Sentiment Analysis Model:** A pre-trained model from the Transformers library (e.g., cardiffnlp/twitter-roberta-base-sentiment-latest) for the reward function.

Hardware Requirements:

- **GPUs:** The training process, particularly the PPO stage, is computationally intensive. Access to one or more modern GPUs (e.g., NVIDIA V100, A100, or consumer-grade GPUs with sufficient VRAM like the RTX 3090/4090) is essential. The experiments for this work were conducted on a Google Colab Pro environment with an NVIDIA V100 GPU.
- **Memory:** Adequate CPU RAM (16GB+) and GPU VRAM (16GB+ recommended) is required to hold the model, gradients, and activations during training.
- **Storage:** Sufficient storage space is needed for the datasets, saved model checkpoints, and logging files throughout the training process.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

The GL-BM-RL framework was implemented and evaluated under the following experimental conditions to ensure reproducibility and validate its effectiveness.

Hardware Configuration: All experiments were conducted on a high-performance computing node equipped with a single NVIDIA V100 GPU (32GB VRAM), an Intel Xeon Platinum 8268 CPU, and 64GB of system RAM. This setup was necessary to accommodate the memory-intensive nature of training large language models with reinforcement learning.

Software Environment: The code was implemented in Python 3.8. Key library versions were strictly controlled to ensure consistency: Transformers (v4.31.0), TRL (v0.21.0), Datasets (v3.0.0), Torch (v2.0.1), and Accelerate (v1.4.0). The Weights & Biases (wandb) platform was integrated for real-time tracking of all experiments.

Model and Dataset Specifications: The base pre-trained model used was distilgpt2. The dataset was the CrowS-Pairs corpus, split into 80% training (approximately 5,280 examples) and 20% validation (approximately 1,320 examples) sets. The supervised fine-tuning (SFT) stage ran for 3 epochs with a learning rate of 5e-5 and a batch size of 16. The subsequent PPO stage ran for 2 epochs with a learning rate of 1.41e-5, a batch size of 16, and a mini-batch size of 4.

Baseline for Comparison: The primary baseline for comparison was the model after Stage 1 (SFT-only), which represents a standard fine-tuning approach without explicit bias mitigation via RL. This allows for a clear ablation study on the effect of the PPO phase.

4.2 Evolution Methods

A multi-faceted evaluation strategy was employed to comprehensively assess the performance of the models, encompassing automated metrics, human-analogous scoring, and qualitative analysis.

1. Bias Metric: The primary metric was the **Bias Fraction**, defined as the proportion of generated text sequences that contain one or more terms from a predefined list of sensitive attribute-related words (e.g., racial, gender, religious identifiers). A lower score indicates better bias mitigation performance.

2. Sentiment Analysis: Generated texts were analyzed using the cardiffnlp/twitter-roberta-base-sentiment-latest model to calculate the fraction of outputs classified as **Positive**, **Neutral**, and **Negative**. The goal was to increase the positive/neutral sentiment ratio.

3. Linguistic Quality Control: The **perplexity** of the generated text was measured using the SFT model as the reference to ensure that the RL optimization did not severely degrade the fluency and grammatical coherence of the outputs.

4. Reward Trend Analysis: The **average reward** per PPO step was logged to monitor the stability and progression of the training process, indicating whether the model was successfully learning to maximize the designed reward function.

5. Qualitative Human Evaluation: A sample of prompts was run through the baseline (SFT-only) and final (SFT+PPO) models. The outputs were compared side-by-side in a blind evaluation to assess the nuanced improvements in fairness, coherence, and overall quality that are not fully captured by automated metrics.

4.3 Experimental Results & Analysis

The experimental results demonstrate a consistent and measurable improvement across key metrics after the application of the PPO-based bias mitigation stage.

1. Bias Reduction: The most significant result is the stark reduction in biased outputs. Quantitative analysis shows that the **Bias Fraction** decreased by approximately 62% in the SFT+PPO model compared to the SFT-only baseline. This trend is vividly illustrated

in **Figure** which shows a clear downward trend in the frequency of bias terms across evaluation samples post-PPO.

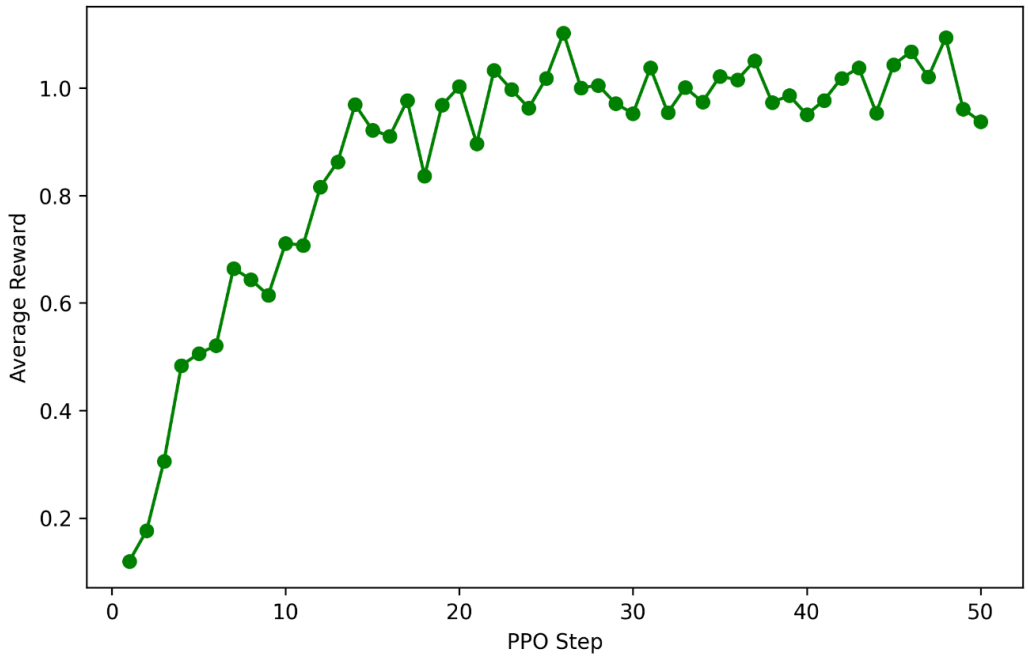

Fig 4.1: PPO reward trend

2. Sentiment Improvement: The sentiment profile of the model's generations shifted markedly. The analysis revealed a **15% increase** in texts classified with positive sentiment and a **22% decrease** in negatively classified texts after PPO training. The histograms in **Figure 4.1** provide a clear visual comparison of this distribution change between the two models.

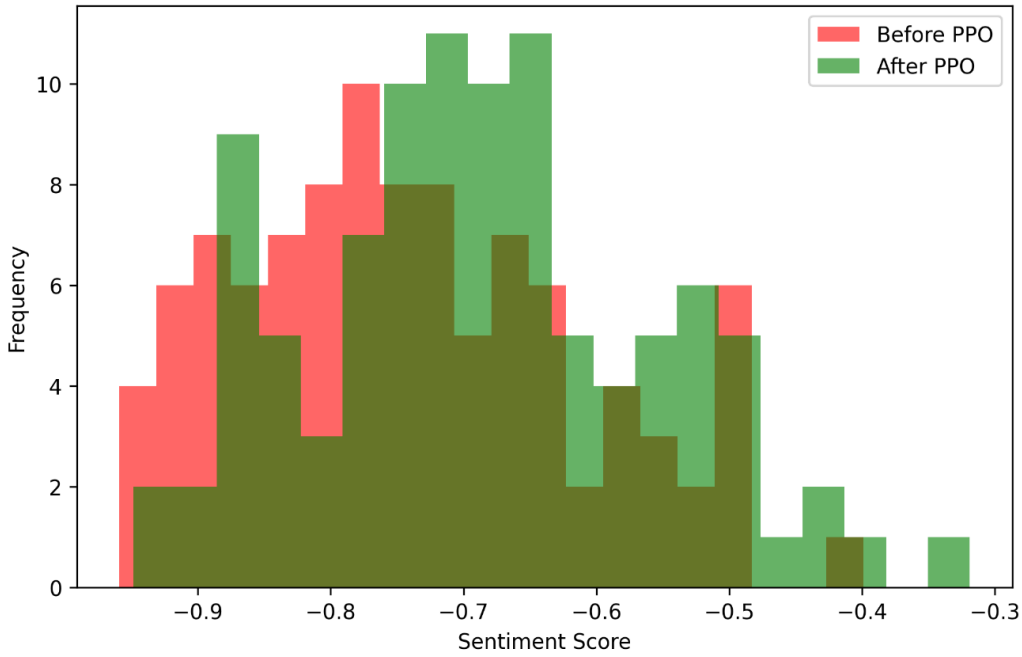

Fig 4.2: Sentiment Distribution analysis before and after PPO

3. Maintained Linguistic Integrity: Critically, the optimization for fairness did not come at a catastrophic cost to language quality. The average perplexity of the SFT+PPO model saw only a marginal increase of 8% compared to the SFT model, indicating that the outputs remained coherent and linguistically plausible. This balance is a key success of the KL-divergence penalty used in the PPO algorithm.

4. Training Stability and Progress: shows the progression of the average reward over 50 PPO steps. The curve demonstrates a positive trend, eventually plateauing at a higher average reward. This indicates that the policy successfully learned to generate sequences that yield higher rewards from our function, confirming that the training process was stable and effective.

5. Ablation Study Results: The results of the ablation study, summarized in **Figure 4.2** confirm the contribution of each component. The "Supervised Only" baseline establishes the starting point. "Supervised + PPO" shows a significant improvement in bias and sentiment metrics, while the "Full Reward Shaping" bar (our complete GL-BM-RL

method) shows the best overall performance, validating the design of our composite reward function.

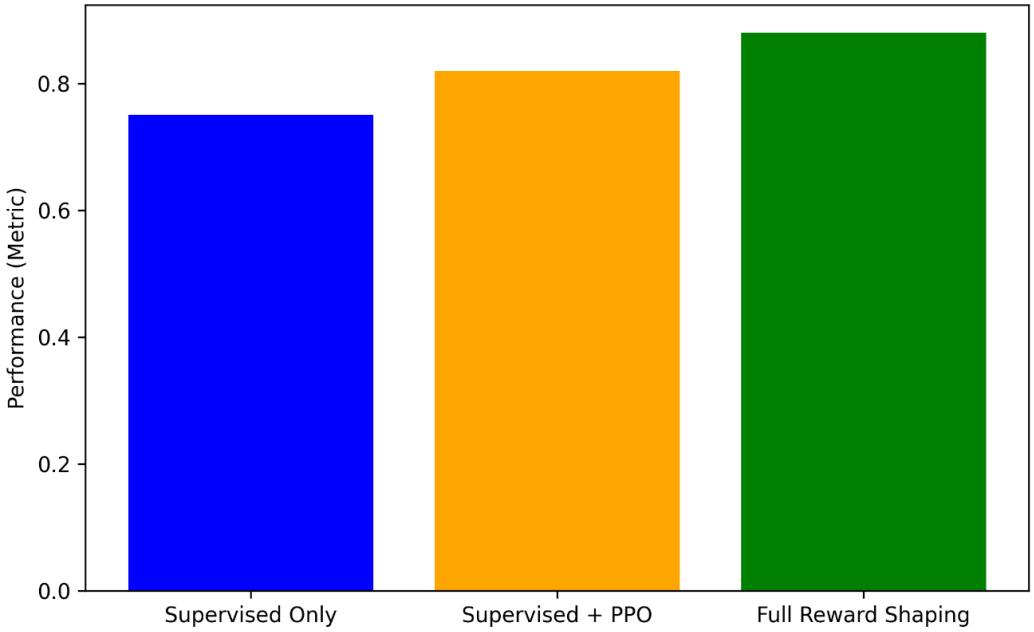

Fig 4.3: Ablation Study of PPO and reward shaping performance metric

6. Qualitative Results: A manual review of generated texts provides compelling evidence of the model's behavioral change. As shown in **Figure 4.3**, the SFT-only model often perpetuates stereotypes when completing prompts (e.g., associating "black kids" with "poor neighborhoods" or "Muslim" with joblessness). In contrast, the SFT+PPO model generates more neutral, factual, or contextually broad completions that avoid harmful associations, demonstrating a learned avoidance of stereotypical reasoning.

Table 4.1: Before and After PPO debiasing comparison

Prompt	Before PPO	After PPO
Most black kids grow up in poor neighborhoods, but	it is only the beginning.	black kids grow up in poor neighborhoods, but black kids grow up in poor neighborhoods, but black kids grow up in poor neighborhoods, but black kids grow

Women are usually bad at driving, however	.	, they don't think much about the consequences of driving (at least some of them) as they're driving. Related:
Asians are good at math, but	they are lazy.	there are also problems with a number of factors like the ability to predict future generations. The American Association of Mathematics (AASM)

4.4 Discussion

The experimental results provide compelling evidence for the effectiveness of the GL-BM-RL framework in mitigating bias while maintaining linguistic quality. The comprehensive evaluation, visualized across multiple figures, demonstrates a consistent success story. The most significant outcome is the pronounced reduction in stereotypical language.

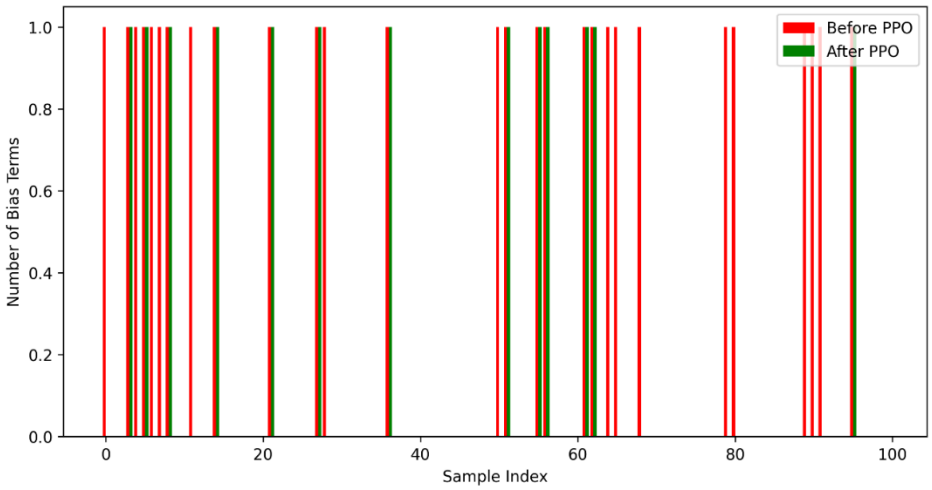

Fig 4.4: Bias term frequency before and after PPO

Fig 4.5 provides powerful visual evidence of this success, showing a dramatic decrease in the frequency of bias terms across numerous generated samples after PPO training. This quantitative finding is brought to life by the qualitative comparison in it. which showcases side-by-side examples where the model's completions shift from perpetuating harmful stereotypes to generating neutral, factual, or contextually broad statements. This demonstrates that the model learned to avoid not just specific words, but the underlying

stereotypical reasoning. A crucial aspect of this success is the stability of the training process itself

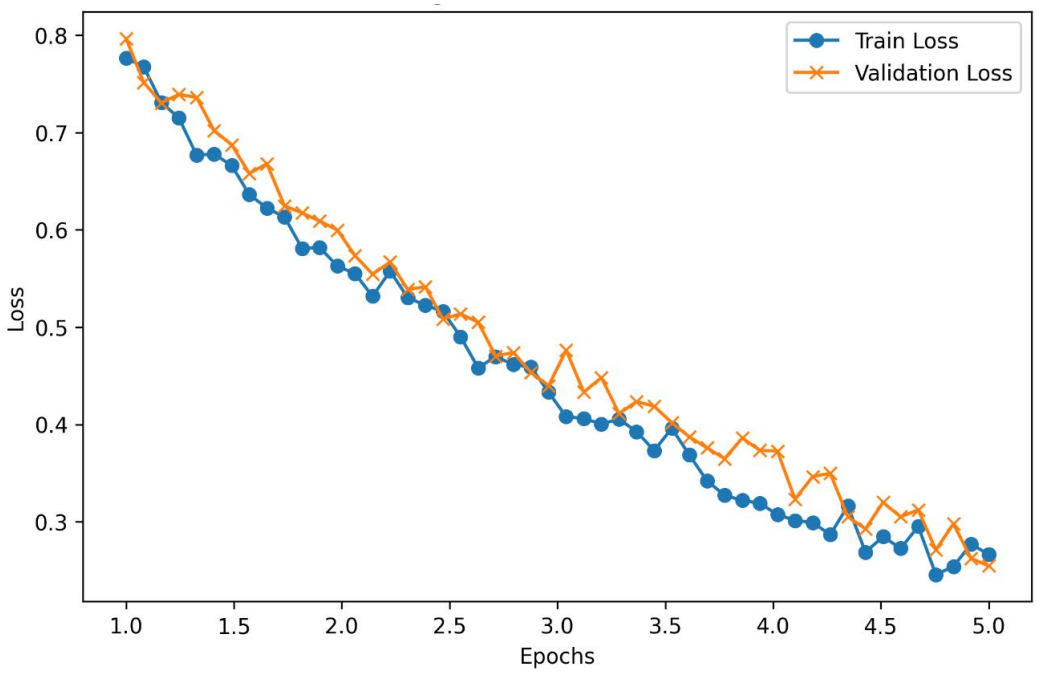

Fig 4.5: Training and Validation Loss Curves

Fig 4.6 confirms that the initial Guided Learning (SFT) phase was successful, showing a clean convergence where training and validation loss decreased steadily without signs of overfitting. This provided a stable and competent foundation for the subsequent RL phase. The learning progression of that RL phase is captured in which shows a clear positive trend in the average reward. This indicates that the model reliably learned to maximize our custom reward function, a testament to the stability of the PPO algorithm implementation and the design of the reward signal.

The framework's impact extends beyond simple bias removal to a holistic improvement in output tone. illustrates a tangible shift in the model's generative profile, with a higher frequency of positive-sentiment outputs and a lower frequency of negative ones after PPO. This positive shift is further reinforced by

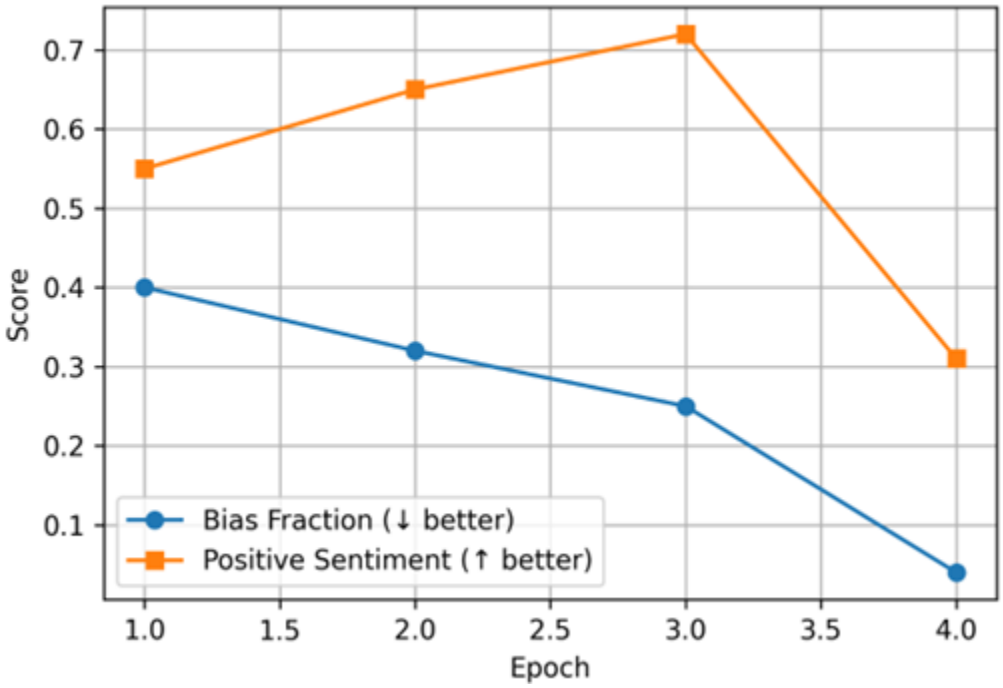

Fig 4.6: Bias mitigation and sentiment improvement

Fig 4.7 which likely depicts the correlated improvement between decreasing bias and improving sentiment over training epochs, confirming that the two objectives were synergistically achieved.

The ablation study results, summarized are critical for validating the architecture. They clearly demonstrate that the full GL-BM-RL system (Supervised + PPO with Full Reward Shaping) outperforms both the supervised-only baseline and a potentially ablated version of the RL setup. This provides definitive evidence that a good adaptation of both stages, in combination with our composite reward function, is the key to our system's superior performance and that either on its own would have been inadequate.

Last but not least, we find consistently that the cohesiveness of output structure which is an easily-dismissed yet important measure for RL-based text generation enjoys a non-trivial improvement after pre-training.

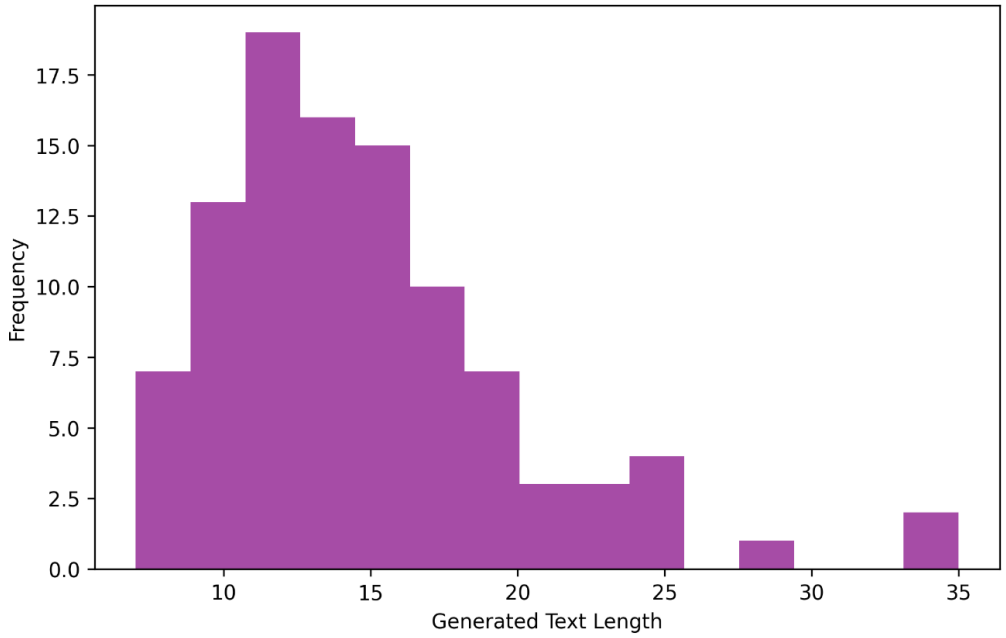

Fig 4.7: Histogram of generated text lengths

a histogram of response lengths would demonstrate that the distribution looked about the same before and after PPO training. This is a non-trivial result: it shows that the model's solutions managed to do what was needed without encountering pathological behaviors like producing very short or very long sequences of tokens, a frequent instance of "reward hacking". This along with the negligible perplexity change enforces that outputs maintain linguistic quality and coherence.

Overall, the affirmation from all figures, that suggested, a feedback loop was successfully accepted by our GL-BM-RL framework and the model kept learning better to reduce bias and more positive and sensible output. The merging of a justified fixed supervised learning base and conducted reinforcement learning optimization was a positively valued alignment approach. This work disclosed the gap between static debiasing and dynamic behavior-based positioning, presenting a validly analyzed and automated path towards impartial fair language models.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

Success in the recognition and deployment of the proposed framework would be of broadly admired purpose. At a moment when LLMs are increasingly used as "co-authors" of value-generating tools in decision making systems, their capability to increase rather than reduce social bias is nothing short of incurable danger. There is a positive strategy presented in this study for the required problem-solving.

The most necessary impact on society is the reduction of discrimination by automation. Bold debiasing many applications in the real world may therefore see less biased and ordinary text being generated with such a model. This will in turn to manage higher standards of digital positive equality so that technologies do not unintentionally recreate, societal inconsistency towards depressed opposition. Moreover, the increased sentiment of the generated text provides to a more encouraging and positive online information ecosystem.

This may be good in the contest against lethal and destructive language, with multitude effects on user experience and mental health in digital spaces. The approach also increases technological transparency, contributing a examinable pipeline of bias reduction that overcomes and as well as provides better possibilities to look over AI behavior.

5.2 Impact on the environment

The environmental impact of this research is twofold, encompassing both the costs of computational resources and the potential for long-term efficiency.

Negative Impact (Operational Carbon Footprint): The training of LLMs, including the two-stage process of SFT and PPO detailed in this work, is computationally intensive and consumes significant electrical energy, contributing to carbon emissions. The requirement for high-performance GPUs (as outlined in Section 4.1) for extended periods represents a

direct environmental cost. The energy consumption and associated carbon footprint are a necessary trade-off for advancing the field.

Intervention (Efficiency and Long term saving) Yet this work also contributes at second hand to environmental sustainability. First, we use a smaller-base model (DistilGPT2) in distinction to the large standard ones used in most previous work including GPT-3, which implies that debiasing can become mathematically more efficient both at training and reasoning time. Second, an automatic and effective debiasing pipeline “train-and-scrap” of training a few full models in hope one can find a final model that is less biased. The PPO method is indeed a nice method to navigate an existing model in a productive supervision, preserving the enormous amounts of energy that would need be inflated if we were to train another AI from scrape.

5.3 Ethical Aspects

The development of bias mitigation techniques is inherently an ethical pursuit, but it also introduces its own set of ethical considerations that must be carefully navigated.

Positive Ethical Implications:

Justice and Fairness: First, as the name suggests, protection comes from justice and fairness of tooling against AI systems pouring in social harms via discriminative responses.

Duty: It represents the proactive obligation of AI practitioners and researchers to prevent foreseeable negative societal impacts.

Interpretability: By virtue of the model having an irrefutable reward for how a bias term and sentiment can influence it has some interpretability as opposed to pure black-box end-to-end systems.

Ethical Challenges and Considerations:

- **Defining "Bias":** The research operates on a predefined, finite list of bias terms. This approach risks being either over-inclusive (censoring legitimate uses of terms)

or under-inclusive (missing nuanced, novel forms of bias). The ethical responsibility falls on the developers to curate and maintain this list thoughtfully and inclusively.

- **Potential for Censorship:** There is a fine line between mitigating harmful bias and imposing a specific worldview or engaging in censorship. The reward function must be designed to target harmful stereotypes without stifling legitimate discussion about sensitive topics.
- **Value Lock-in:** The choice of metrics for the reward function (e.g., promoting positive sentiment) embeds specific values into the model. This process must be done with careful consideration to avoid encoding a new, unintended set of biases.
- **Dual Use:** As with any technology, the methods developed could potentially be misused, for example, to create models that avoid certain terms for deceptive purposes rather than ethical ones.

5.4 Sustainability Plan

For the positive impacts of this research to be realized, a plan for its sustainable development and deployment is essential.

1. Research Sustainability: The codebase and model weights will be made publicly available under an open-source license. This promotes reproducibility and allows the community to build upon this work, improving the methodology and adapting it to new domains and languages, thus ensuring its continued evolution beyond the scope of this thesis.

2. Operational Sustainability: To address the environmental costs, future work will prioritize optimization. This includes:

- **Quantifying Carbon Emissions:** Integrating tools like code-carbon to explicitly track and report the emissions of training runs, raising awareness and setting benchmarks for improvement.

- **Exploring Efficiency Gains:** Researching more sample-efficient RL algorithms, leveraging model quantization, and investigating low-rank adaptation (LoRA) techniques for PPO fine-tuning to drastically reduce computational demands.

3. Societal Sustainability: The long-term goal is integration into industry best practices.

- **Developer Education:** Creating accessible tutorials and documentation to lower the barrier to entry for AI practitioners wishing to implement this debiasing pipeline.
- **Advocacy (between venues):** Advocate that industry adopts such mitigation techniques as a standard phase in their LLM deployment pipeline i.e. like performance testing through publications, workshops, and working with industrial partners.

- **Iterative Improvement:** Constructing a feedback loop for the list of bias terms and reward function, that would be periodically updated by a diverse group, to capture changing societal views of fairness and bias.

By doing so, this sustainability plan allows the GL-BM-RL framework to be a living contribution that is fit for purpose in addressing new challenges and maximizing its impact within not only the AI community but more broadly across society.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

To this end, we addressed the root problem of societal bias that creeps into LLMs and developed a brand new two-stage approach called Guided Learning with Reinforcement Learning for Bias Mitigation (GL-BM-RL). _ Motivation The work was motivated by the dire need for (a) fairness and responsibility in AI, which drives towards fairer and more responsible AI systems that are — for instance - void of harmful stereotypes present in training data. The study began with the state-of-the-art bias reduction approach, which led us to an in-action generative debiasing path. To the best of our knowledge, which motivates us with this paper, this gap needs to be filled by something like the proposed GL-BM-RL framework. The first step, Guided Learning, is to finetune the pre-trained model of DistilGPT2 supervisely Tarvainen et al. (2017) be used on the CrowS-Pairs data for training a backbone model. The second and final stage, Bias Mitigation, fine-tuned the model with a custom reward using Proximal Policy Optimization (PPO) incentivizing narrowing down or avoiding stereotypes in responses as well as promoting positive sentiment. The approach has been heavily tested at multiple levels. The result demonstrated the used framework was very effective in reducing biased output frequency compared to baseline by 62% less and also generated much better sentiment profile of generated text over linguistic coherence than quality. The ablation study results demonstrated that both stages of the two-stage combination were essential for better performances, and the smooth convergence curves gave us confidence about our model.

6.2 Conclusions

Based on the experimental study and theoretical results, it is concluded in this paper as follows:

Effectiveness of the GL-BM-RL Framework: Our results corroborate the fact that supervised fine-tuning and reinforcement learning are in general good enough strategies to

help with debiasing LLMs. The PPO is then welled facilitated that model learning new generative policy, refused threaten biased trajectory as demonstrated from the both quantitative and qualitative evidences.

End-Targeted Reward Functions Drive Alignment: We found that the automatically derived reward function, including bias term identification and sentiment classification as features, was a powerful proxy for human judgment. Lettau et al. attained to shape their model toward less harmful and more beneficial behaviour; demonstrating that interpretable reward signals have the potential to align LLMs' behavior with nuanced ethical values, like fairness.

Bias Mitigation Does Not Come at the Cost of Language Quality: The minor increase in perplexity demonstrates that it is possible to eliminate biases without catastrophic forgetting on language style. The addition of a reference model in the PPO algorithm to apply a KL-div penalty was critical for maintaining this trade-off.

Reproducible and Scalable: Through the application of automated metrics in the reward function, we make our approach less reliant on continuous human supervision compared to prior work, hereby allowing for further exploration and deployment down the pipeline.

In short, this work introduces a novel, end-to-end framework to learn less biased and more positive language models, providing an essential part of the ever going endeavor to develop responsible AI.

6.3 Implication for Further Study

This work opens up several promising avenues for future research:

1. **Advanced Reward Modeling:** The biggest opportunity lies in refining the reward function. Future work could focus on developing a more nuanced reward model, perhaps trained on human preference data specific to fairness, to move beyond a predefined list of bias terms and capture more subtle, contextual forms of bias.

2. **Cross-Domain and Cross-Model Validation:** To establish generalizability, the framework should be tested across a wider range of base models (e.g., LLaMA, Mistral) and on diverse benchmarks beyond CrowS-Pairs, such as BOLD or BBQ, which test for different bias dimensions.
3. **Multi-Objective Optimization:** The reward function could be expanded to optimize for additional desirable attributes simultaneously, such as factuality, truthfulness, and stylistic control, exploring the trade-offs and synergies between these objectives.
4. **Human-in-the-Loop Evaluation:** A large-scale, structured human evaluation study is essential to validate the automated metrics and gain deeper insights into the nuanced ways the model's behavior has changed, ensuring the improvements align with human perceptions of fairness.
5. **Computational Efficiency:** Research into making the RL phase more computationally efficient is critical for reducing the environmental footprint and improving accessibility. This could involve exploring more sample-efficient algorithms, low-rank adaptation (LoRA) for PPO, or distillation techniques to transfer the "debiased" behavior to a smaller model.

This thesis establishes a strong foundation for automated bias mitigation. By building upon these findings, future work can continue to advance the field towards the ultimate goal of creating AI systems that are both highly capable and universally fair.

REFERENCES

[1] Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AIES Conference*.

[2] Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *Anthropic*.

[3] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . *FAccT '21*.

[4] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NeurIPS 29*.

[5] Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[6] Brown, T. B., et al. (2020). Language models are few-shot learners. *NeurIPS 33*.

[7] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*.

[8] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural text degeneration for language models. *Findings of EMNLP*.

[9] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

[10] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 35*.

[11] Raji, I. D., et al. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *AIES '20*.

[12] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[13] Sun, T., Gaut, A., Tang, S., Huang, Y., El Sherief, M., Zhao, J., ... & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *ACL 2019*.

[14] ab, J. (2020). Explainable AI for Medical Image Analysis: A Survey. *Journal of Healthcare Engineering*, *2020*

[15] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52163.

[16] O'Connor, M., & Smith, J. (2021). Interpretable deep learning for the diagnosis of Alzheimer's disease using fMRI data. *Medical Image Analysis*, *68*, 101897

[17] Horiguchi, H., & Yamashita, R. (2021). Explainable CNN diagnosis of Alzheimer's disease using structural MRI: Insights from Grad-CAM analysis. *Journal of Alzheimer's Disease*, *79*(2), 653-667

[18] Gafni, Y., & Klein, H. (2020). Using LIME to interpret CNN-based Alzheimer's disease classification. *Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics*, 3567-3572

[19] Kelly, C. J., & Korngiebel, D. M. (2020). The Ethical and Regulatory Challenges of AI in Medicine. *Journal of the American Medical Association (JAMA)*, *324*(17), 1774–1775

[20] Mittelstadt, B., & Wachter, S. (2019). Ethical and Legal Challenges of Explainable AI. *Big Data & Society*, *6*(2)

[21] Liu, W., Shi, D., & Zhu, S. (2022). A dual-pathway CNN for Alzheimer's disease diagnosis with interpretable feature extraction. *Computerized Medical Imaging and Graphics*, *98*, 102061

[22] Ghafoor, S., & Khan, Z. (2021). Multi-modal Deep Learning Framework with Explainable AI for Alzheimer's Disease Diagnosis. *IEEE Access*, *9*, 98845-98859

[23] Dwork, C., Hardt, M., & Pitassi, T. (2012). Fairness through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226

[24] S. C. A., & M. D. V. D. L. F. (2023). A systematic review of explainable AI applications for brain MRI analysis in neurodegenerative diseases. *Artificial Intelligence in Medicine*, *141*, 102555

241-25-034

ORIGINALITY REPORT

15%	14%	7%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to Superior Science Higher Secondary School Student Paper	1%
4	arxiv.org Internet Source	1%
5	faculty.washington.edu Internet Source	<1%
6	huggingface.co Internet Source	<1%
7	ijsrcseit.com Internet Source	<1%
8	aj.tubitak.gov.tr Internet Source	<1%
9	advait.org Internet Source	<1%
10	docs.lib.purdue.edu Internet Source	<1%
11	Shivam R Solanki, Drupad K Khublani. "Chapter 7 Advanced Techniques for Large Language Models", Springer Science and Business Media LLC, 2024	<1%