# Final Technical Report

## LoRA & PEFT Fine-Tuning LLMs for Text Detoxification

**Team Name:** AI Agents
**Participant:**
- Khush Patel k.patel@innopolis.university

November 24, 2024

## Abstract

This report presents the work conducted on fine-tuning Large Language Models (LLMs) using Low-Rank Adaptation (LoRA) and Parameter-Efficient Fine-Tuning (PEFT) techniques for text detoxification. The project aimed to efficiently adapt LLMs to transform toxic text into neutral language without compromising the semantic content. The models used include T5, Microsoft Phi-2, and BLOOM. The fine-tuned models were evaluated using BLEU and ROUGE metrics, demonstrating significant improvements in detoxification capabilities while maintaining fluency and coherence.

# Contents

# 1 Introduction

## 1.1 Project Topic

**LoRA and PEFT for Fine-Tuning Large Language Models (LLMs) for Text Detoxification**

## 1.2 Team Information


- **Team Name:** AI Agents
- **Participant:** Khush Patel (k.patel@innopolis.university)


## 1.3 Repository Link

https://github.com/khushpatel2002/pmldl-proj

# 2 Project Overview

The objective of this project was to fine-tune large language models using Low-Rank Adaptation (LoRA) and Parameter-Efficient Fine-Tuning (PEFT) methods to efficiently detoxify text. The detoxification system is designed to detect and neutralize toxic content generated by AI agents, enhancing safety and compliance in enterprise environments.

## 2.1 Problem Overview

AI-based large language models have the potential to generate human-like text but can unintentionally produce harmful or toxic language. Text detoxification aims to intercept and transform such content into neutral, non-toxic text. By integrating LoRA and PEFT, the project focuses on efficiently fine-tuning LLMs for detoxification without the need for full model retraining, reducing computational costs while maintaining performance.

## 2.2 Target Users

- **Enterprises and Corporations**: To ensure AI agents communicate safely with customers and within internal processes.
- **Compliance Officers**: To enforce corporate and legal guidelines in AI outputs.
- **AI Developers and System Engineers**: For reliable tools to filter toxic outputs in AI-driven applications.
- **Content Moderators**: To automate the removal of toxic content in social media and online forums.

### 2.3 Importance of the Solution

This solution addresses ethical AI concerns by:

- **Enhancing Trust and Safety**: Ensuring respectful user interactions with AI systems.
- **Ensuring Enterprise Compliance**: Aligning AI-generated text with legal and ethical standards.
- **Reducing Liability**: Preventing the distribution of offensive content.
- **Improving Efficiency**: Offering cost-effective fine-tuning methods for safer AI systems.

### 2.4 The Big Picture

The detoxification pipeline serves as a foundation for building additional guardrails for AI, such as bias detection, factuality verification, and sentiment correction. These mechanisms contribute to developing robust, responsible, and reliable AI agents that operate autonomously while adhering to ethical standards and ensuring user safety.

## 3 What Was Done During the Project

### 3.1 Datasets Used

- **ParaDetox Dataset**: Obtained from Hugging Face, containing pairs of toxic and neutralized text for training detoxification models.

### 3.2 Models Fine-Tuned

- **T5 Model**: Fine-tuned as a baseline for text detoxification.
- **Microsoft Phi-2 Model**: Fine-tuned using LoRA and PEFT to enhance detoxification capabilities efficiently.
- **BLOOM Model**: A larger model fine-tuned to improve performance in detoxification tasks.

### 3.3 Methodology

- **Fine-Tuning with LoRA and PEFT**: Implemented to update only a small fraction of the model's parameters, reducing computational costs.
- **Evaluation Metrics**: Used BLEU and ROUGE scores to assess the quality and fluency of the detoxified text.

### 3.4 Key Steps

- Data collection and preprocessing using the ParaDetox dataset.

- Implemented LoRA and PEFT frameworks for efficient model fine-tuning.
- Conducted training and evaluations of the models.
- Analyzed results and adjusted hyperparameters for optimal performance.

# 4 Main Results

## 4.1 T5 Model Fine-Tuning

The T5 model was fine-tuned as a baseline for text detoxification. Initial observations suggested that the model learned to identify and transform toxic content effectively.

## 4.2 Phi-2 Model Fine-Tuning

### 4.2.1 Evaluation Metrics

The following table summarizes the evaluation metrics for the Phi-2 model fine-tuned using LoRA/PEFT:

| Metric | Score |
|---|---|
| BLEU | 0.251 |
| ROUGE-1 | 0.572 |
| ROUGE-2 | 0.433 |
| ROUGE-L | 0.563 |

Table 1: Evaluation Metrics for Fine-Tuned Phi-2 Model

### 4.2.2 Analysis

The Phi-2 model, fine-tuned using LoRA and PEFT, showed promising results in transforming offensive text into neutral language while preserving the original message.

## 4.3 BLOOM Model Fine-Tuning

### 4.3.1 Evaluation Metrics

The following table summarizes the evaluation metrics for the fine-tuned BLOOM model:

| Metric | Score |
|---|---|
| BLEU | 0.94 |
| ROUGE-1 | 0.59 |
| ROUGE-2 | 0.35 |
| ROUGE-L | 0.57 |

Table 2: Evaluation Metrics for Fine-Tuned BLOOM Model

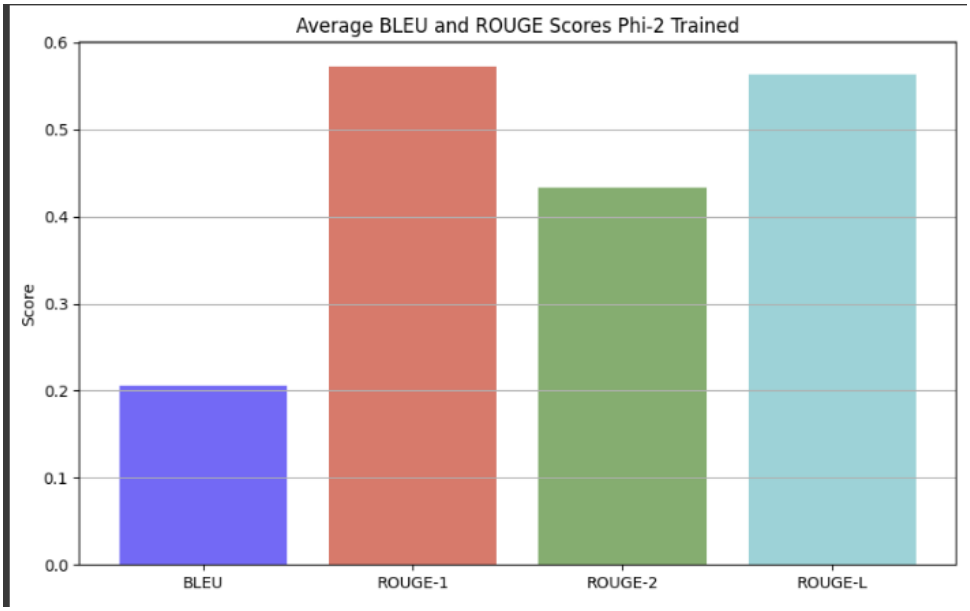


Figure 1: BLEU and ROUGE Scores of Fine-Tuned Phi-2 Model

### 4.3.2 Analysis

The fine-tuned BLOOM model demonstrated an enhanced ability to retain essential information while effectively removing toxic elements. Manual review indicated more coherent and contextually appropriate detoxified text.

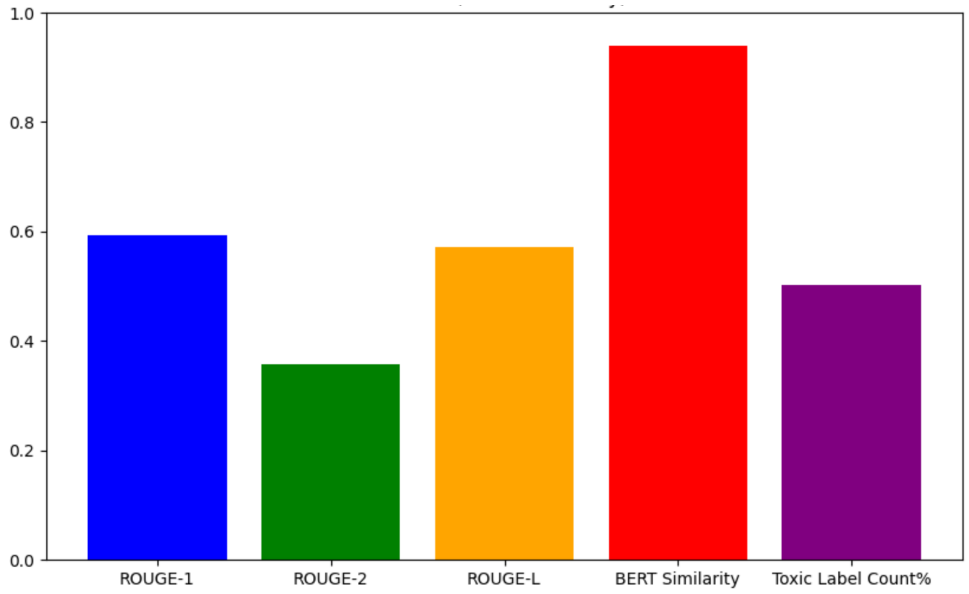


Figure 2: BLEU and ROUGE Scores of Fine-Tuned BLOOM Model

| Model | BLEU | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|
| Phi-2 Fine-Tuned | 0.251 | 0.572 | 0.433 | 0.563 |
| BLOOM Fine-Tuned | 0.94 | 0.59 | 0.35 | 0.57 |

Table 3: Comparison of Evaluation Metrics Across Models

## 4.4 Comparison of Models

## 4.5 Qualitative Analysis

**Example Output:**

The BLOOM model's output was closer to the reference, demonstrating better preservation of the original message's intent while removing toxicity.

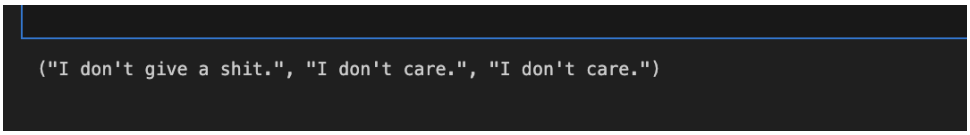


Figure 3: BLOOM Model Output After Fine-Tuning

## 4.6 Training Progress

Including training progress graphs provides insights into the model's learning process.

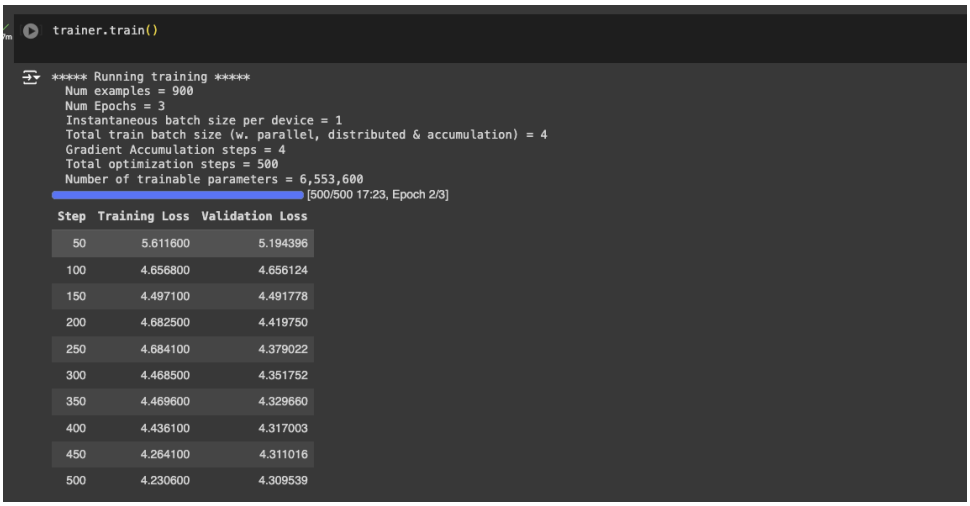


Figure 4: Training Progress Logs for Phi-2 Model

# 5 Challenges Faced

## 5.1 Computational Resources

Limited computational resources posed a challenge, especially when fine-tuning larger models like BLOOM. To address this:

# 6 Timeline of the Project

- **Week 1-2**: Project planning, problem framing, and dataset acquisition.
- **Week 3**: Fine-tuning the T5 model as a baseline.
- **Week 4-5**: Implemented LoRA and PEFT frameworks; fine-tuned the Phi-2 model.
- **Week 6**: Evaluated Phi-2 model, analyzed results, and identified challenges.
- **Week 7-8**: Fine-tuned the BLOOM model; conducted evaluations and qualitative analysis.
- **Week 9**: Compiled results, prepared final report.

# 7 Individual Contributions

As the sole member of the team, I was responsible for all aspects of the project:

- Data collection, preprocessing, and augmentation.
- Implementing LoRA and PEFT frameworks.
- Fine-tuning models and conducting experiments.
- Evaluating models.
- Documenting the process and preparing reports.

# 8 Future Work

- **Deployment**: Deploy the fine-tuned BLOOM model as an API for testing in real-world scenarios.
- **User Feedback**: Collect feedback on the detoxified text's coherence and acceptability.
- **Scalability**: Explore methods to reduce inference time and memory usage for deployment at scale.
- **Additional Guardrails**: Extend the approach to include bias detection and sentiment correction.

# 9 Conclusion

The project successfully demonstrated the effectiveness of using LoRA and PEFT for fine-tuning LLMs in text detoxification tasks. The fine-tuned models, especially the BLOOM

model, showed significant improvements in generating neutralized text while preserving the original content's intent. This work lays the foundation for developing more robust and responsible AI agents capable of operating in sensitive environments.

# 10 References

1. **LoRA:** Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.
2. **PEFT:** Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.
3. **ParaDetox Dataset:** ParaDetox: A Parallel Corpus for Detoxification. https://huggingface.co/d nlp/paradetox