LLM Experimentation Report

February 7, 2025

1 Introduction

This report documents our experimentation with Large Language Models (LLMs) for generating explanations of mental health conditions and suggesting coping mechanisms. Initially, we used GPT-2, and later, we fine-tuned Falcon 7B in 4-bit Mode (QLoRA) on a specific dataset. We attempted to use Mistral but did not receive access.

2 Dataset and Preprocessing

We used the Kaggle dataset `shahzadahmad0402/depression-and-anxiety-data` to train our model. The preprocessing steps included:

- Cleaning text (removing special characters, stopwords, and performing tokenization).
- Label encoding for classification.
- Splitting into training and validation sets (80-20 split).
- Converting data into a format compatible with transformer-based models.

3 Model Selection Rationale

Initially, we used GPT-2 due to its ease of use and availability. However, due to its limited performance in generating detailed explanations, we moved to Falcon 7B in 4-bit Mode (QLoRA). Falcon 7B was chosen for:

- Its open-source nature.
- Better language generation capabilities compared to GPT-2.
- Efficient fine-tuning with QLoRA, making it feasible to run on limited hardware.

We also explored Mistral, but access restrictions prevented us from using it.

4 Implementation Details

4.1 Option 1: Prompt Engineering

We crafted prompts that effectively generate mental health explanations and coping strategies. Example prompt:

> "Explain the mental health condition 'Anxiety' and suggest coping mechanisms, next steps, and helpful resources."

4.2 Option 2: Fine-Tuning Falcon 7B

We fine-tuned Falcon 7B on our dataset using QLoRA to optimize performance while using minimal GPU resources. The fine-tuning steps included:

- Loading Falcon 7B in 4-bit mode using `bitsandbytes`.
- Training on labeled mental health descriptions.
- Evaluating outputs based on coherence and factual accuracy.

5 How to Run the Inference Script

The inference script is provided as a Jupyter Notebook (.ipynb). To run:

1. Install dependencies: `pip install transformers bitsandbytes accelerate`.
2. Load the trained Falcon 7B model.
3. Run the inference cell and input the condition for which an explanation is needed.

6 CLI Usage Instructions

The CLI is implemented in a Jupyter Notebook cell. Users can input a mental health condition, and the model will generate an explanation and coping strategies.

7 Findings and Conclusion

- GPT-2 generated generic responses but lacked depth.
- Falcon 7B fine-tuned with QLoRA provided more contextually accurate and detailed explanations.
- QLoRA allowed efficient fine-tuning on limited hardware.

Overall, Falcon 7B with QLoRA proved to be the best option for our use case.