Final Project Report - Medical Advisor Bot

Joe, Jordan, Thomas

**Introduction**

In recent years, large language models have shown potential in understanding and generating human language across various domains. However, applying LLMs to specialized fields such as healthcare often requires domain-specific fine-tuning to improve performance in understanding specialized terms and offering accurate advice. The goal of this project was to fine-tune a pretrained LLM to enhance its ability to answer medical questions from patients. This important task can be very useful in supporting healthcare professionals, as well as providing knowledgeable medical advice for patients unable to self diagnose or see a professional directly.

**Pretrained LLM Chosen**

For this project, we decided to change our original model (GPT3.5) to TinyLlama, a smaller variant of the popular Llama language model. TinyLlama is designed to be more efficient while maintaining competitive performance. This model is particularly well-suited for tasks that involve specialized domains like the healthcare advisor language model we created. In addition, TinyLlama is well supported by frameworks like Hugging Face’s transformers, making it easy to fine-tune on domain-specific datasets, such as medical question-answering datasets, ensuring that the model is tailored to the needs of the task.

**Methodology**

**Dataset Preparation:** The first step in the methodology was to prepare a dataset for fine-tuning TinyLlama, we used a pre-existing medical question-answering dataset, which contains a variety of medical questions along with their corresponding answers. This dataset was selected due to its relevance to the tasks and its representation of a wide range of medical topics, making it suitable for fine-tuning a model to respond to medical queries accurately.

**Fine-Tuning the Model:** For the fine-tuning phase, we utilized Hugging Face’s library, which simplified the process of training and fine-tuning transformer-based models like TinyLlama. We loaded the pre-trained model into the code, and fine-tuning was done by training TinyLlama on the medical dataset to adapt it to the medical domain. The training process included setting up hyperparameters like learning rate, batch size, and the number of epochs. During training, techniques like learning rate decay, freezing layers and gradient clipping were employed to stabilize training and prevent overfitting.

**Evaluation:** After fine-tuning, the model's performance was evaluated using metrics BERTScore and accuracy. BERTScore was used to evaluate how well the model-generated answers aligned with human reference answers by measuring the semantic similarity between them. Based on the evaluation results, adjustments to the fine-tuning process were made, such as further tweaking the hyperparameters or modifying the dataset for better coverage of specific medical topics.

**Results**

A graph with green and blue dots

AI-generated content may be incorrect.

On average, BERTScore went down after fine-tuning, but this isn’t necessarily a bad thing due to BERTScore being a metric for semantic similarity. Looking at printed results and examples comparing reasoning from base model TinyLlama to our fine-tuned version, medical accuracy definitely went up after training. Other metrics would need to be implemented, such as review from medical professionals, to really test the accuracy of the newly generated responses. Overall, increasing medical accuracy while still staying around a BERTScore of .85 is considered a good model with reasoning that now specializes in identifying medical issues.

**Challenges & Improvements**

We were met with several challenges during this project, specifically during fine-tuning. There were signs of overfitting, particularly on smaller subsets of the dataset. To mitigate this, we used regularization techniques and careful tuning of hyperparameters. For future improvements we would like to integrate more diverse medical datasets, including medical records, to further enhance the model’s accuracy in specialized medical fields. Additionally, using larger model architectures, like GPT-4, may yield even better results in handling complex queries.