

Milestone 3 report

Kirollos Luka 52-4363

Youssef maged 52-13803

A Comparative Study on Fine-Tuning Strategies for Medical Chatbots: LoRA-Based Adaptation of BioGPT vs. Conventional Fine-Tuning of T5

Study Overview

- **Objective:**

To compare two fine-tuning strategies for building a medical chatbot and evaluate which approach leads to better performance and efficiency.
- **Models Used:**
 - **BioGPT:**
 - Domain-specific language model
 - 1.5 billion parameters
 - Fine-tuned using **LoRA (Low-Rank Adaptation)** for efficient, resource-friendly training
 - **T5-base:**
 - General-purpose language model
 - Fine-tuned conventionally on the same medical dataset
- **Approach:**
 - BioGPT is adapted using a novel fine-tuning method (LoRA) that reduces training complexity.
 - T5-base undergoes traditional fine-tuning to serve as a baseline for comparison.
- **Goal:**

To determine which model and fine-tuning strategy—domain-specific with LoRA or general-purpose with conventional fine-tuning—delivers superior results in terms of:

 - Accuracy
 - Training efficiency
 - Practical applicability for medical chatbot development

Model training overview.

Dataset:

For both models we used the same dataset which can be found on [kaggle](#).

Approach:

We chose a fine-tuning approach instead of a Retrieval-Augmented Generation (RAG) method because, for a medical chatbot, some answers might not be easily retrieved from a fixed database. Additionally, we couldn't find a dataset that included both questions and their relevant context, which is essential for effective retrieval. Therefore, the chatbot needed to infer information based on its pretrained knowledge, making fine-tuning a more suitable option.

Models choice:

The first model we used is based on BioGPT, which is built upon the GPT-2 architecture and contains approximately 1.5 billion parameters. We chose BioGPT for this project because it is specifically pretrained on biomedical literature, making it well-suited for understanding and generating medical language. Its size strikes a good balance between complexity and feasibility for fine-tuning on our available hardware. Although we initially attempted to train a larger model with 7 billion parameters to potentially improve performance, we were unable to do so due to resource constraints, such as limited GPU memory and computational capacity. Therefore, the 1.5 billion parameter BioGPT model was selected as the most practical and effective choice for our medical chatbot.

The second model we experimented with was the T5 model, starting with the T5-base version due to its strong performance across various NLP tasks. We considered using the larger T5-large model to potentially improve results, but resource and time constraints made this impractical. Traditional fine-tuning of larger models like T5-large requires significantly more computational power and longer training times, which exceeded our available capabilities. As a result, we proceeded with T5-base as the more feasible option for this project.

BioGPT

Model architecture

BioGPT uses a decoder-only architecture, meaning it generates text one word at a time based on what it has seen so far. This makes it fast and efficient, but it might miss some details when understanding complex inputs.

T5 uses an encoder-decoder design, where the encoder first reads and understands the whole input, and then the decoder generates the output. This helps T5 handle tasks that need deeper understanding, but it takes more time and computing power.

In short, decoder-only models are quicker but sometimes less thorough, while encoder-decoder models are better at understanding context but need more resources.

BioGPT training

We trained the BioGPT model using a parameter-efficient fine-tuning approach called LoRA (Low-Rank Adaptation). Starting with the pretrained BioGPT model (based on GPT-2 architecture with 1.5 billion parameters), we applied 4-bit quantization to reduce memory usage and make training feasible on limited hardware. The model was prepared specifically for k-bit training to enable this efficient fine-tuning.

We configured LoRA to target key modules within the model's attention and feed-forward layers, allowing us to update a small subset of parameters while keeping the majority fixed. This significantly reduces the computational cost and memory footprint of fine-tuning.

Our training data was reformatted to include patient and doctor dialogue context, which was tokenized using the BioGPT tokenizer. Training was run for 3 epochs with a batch size of 4 and gradient accumulation to effectively increase batch size while managing GPU memory. Techniques like mixed precision (fp16) and gradient checkpointing were used to further optimize resource usage.

We trained the model using the SFTTrainer, designed to work seamlessly with the LoRA approach and causal language modeling tasks. After training, the fine-tuned model and tokenizer were saved for later use, and the validation dataset was stored separately.

BioGPT results

Metric	Score
ROUGE-1	0.2372
ROUGE-2	0.0404
ROUGE-L	0.1373
BLEU	0.0322
METEOR	0.1903

These evaluation scores appear low on the surface, but they can be misleading given the nature of the task. Metrics like ROUGE, BLEU, and METEOR are primarily designed to assess similarity to a reference text, which works well for tasks like summarization or translation. However, for open-ended question-answering in a medical context, these metrics don't fully capture the quality or relevance of the answers.

Medical QA often allows for variation—different doctors may provide different but equally valid responses to the same patient case. As a result, the model might generate accurate and appropriate responses that differ in wording or structure from the reference, leading to lower automatic scores despite good performance. This highlights the limitation of relying solely on these metrics for evaluation in complex, open-domain tasks like medical dialogue.

Prompt : Hey doctor, My eyes feel very dry and there is frequent mucus that is coming out of my eye. What could be the reason?

Response : Hello, Thanks for your query. I have gone through your query and here are my thoughts on your concerns. The symptoms of dry eye are due to tear deficiency. You can use artificial tears or eye drops. But if you do not get rid of mucus, then it will become more difficult to clear the eye. It will cause discomfort in

your eyes. Also, it may lead to vision loss. If you have any other problems, then you should consult a physician for further evaluation

The generated response to the prompt about dry eyes and mucus shows that the model has learned to provide structured, medically relevant information. It begins with a polite and professional greeting and gives a plausible explanation—tear deficiency—as a cause for the symptoms, suggesting the use of artificial tears and advising further consultation if the issue persists. This reflects an understanding of general medical advice that would be expected in such a context.

However, certain consistent patterns in the model's responses are directly influenced by the training dataset. For instance, responses often begin with a generic opening like *"Thanks for your query. I have gone through your query and here are my thoughts..."* This repeated phrasing is a result of patterns seen frequently in the training examples. In some cases, the model also includes a specific doctor's name—likely one present in the original dataset—which may not be appropriate when generating generalized answers.

Despite these quirks, the core content of the responses is usually coherent and medically relevant. The model often provides sensible advice that aligns with basic clinical knowledge, which suggests that the fine-tuning process was successful in helping the model learn appropriate patterns of communication and reasoning in the medical QA domain.

Chatbot creation

This notebook demonstrates the deployment of a fine-tuned BioGPT model as a medical chatbot using Gradio and LangChain. The workflow involves loading a LoRA-adapted version of BioGPT, setting up a Hugging Face pipeline for text generation, and wrapping it into a conversational interface. A conversational memory is included using LangChain's ConversationBufferMemory, and responses are formatted using prompt templates to mimic a doctor-patient dialogue.

The chatbot performs reasonably well in generating medically relevant responses and maintains a professional tone. It successfully loads the fine-tuned model, merges LoRA weights, and integrates it with a simple UI for interaction.

Downsides & Limitations:

1. Limited Conversational Memory:

One major limitation stems from the use of BioGPT (based on GPT-2), which is a decoder-only architecture not originally trained for multi-turn conversation.

As a result, its ability to leverage context from past interactions is weak. While LangChain's memory stores recent interactions, the model does not deeply integrate or reason over previous messages.

2. Small Context Window:

BioGPT, like GPT-2, has a limited context window (~1024 tokens). This restricts the length and depth of the conversation the model can handle effectively. Long interactions or complex multi-turn medical cases may lead to loss of earlier context and cause irrelevant or repetitive responses.

3. Lack of True Understanding of Dialogue Flow:

GPT-2's pretraining did not include dialogue optimization, so it doesn't naturally excel at keeping track of who's speaking or adapting to user feedback mid-conversation. This limits its utility in cases where precise, turn-by-turn understanding is crucial.

Despite these constraints, the chatbot provides a functional proof of concept for deploying fine-tuned medical models.

T5-Base

Model architecture

As mentioned above, T5-base is an encoder-decoder model in which the encoder processes the input sequence and compresses it into a context vector, while the decoder takes the context vector and generates the output sequence. The advantage of this architecture is that it can handle variable-length input and output sequences, and it also effectively captures and maintains context from the input sequence. However, this architecture disadvantage is that the early input elements may get diluted by the time they reach the decoder and also requires significant data and computational resources.

T5-Base Training

We used a basic fine-tuning approach for the specific task training for the model, we used to freeze the first 9 layers out of 12 of the encoder and freeze the embedding layers also. That reduces the training time since the frozen layers don't require gradient computation, and decreases the memory usage during training. The Lower layers of the encoder are used to capture the fundamental language patterns and the generic features that transfer well across the tasks, and freezing them preserves the pre-trained knowledge. This also prevents overfitting, as it reduces the number of trainable parameters. We kept the higher layers of the encoder unfrozen, as it allows the model to adapt to the medical domain terminology and patterns while maintaining the general language understanding in the lower layers.

The reason we didn't freeze any decoder layer, as we needed the model to adapt fully to the task-specific output patterns and the new response styles, and try to produce appropriate medical advice.

T5-base model contains approximately 220 million parameters, and we used the same dataset we used with the BioGPT, but we trained it only on 40k entry from the dataset with 20% of the data as a test dataset. This gave us a balanced fine-tuning process, as we had limited time and computation power for the training through our fine-tuning approach.

We used Seq2Seq trainer as it is designed for sequence-to-sequence tasks, it is optimized for encoder-decoder models as T5-base model we are using. After training, we save the tokenizer and the model for further testing.

T5 results

Metric	Score
ROUGE-1	0.1310
ROUGE-2	0.0171
ROUGE-L	0.0825
BLEU	0.0072
METEOR	0.0852

As we mentioned before, these metrics could be misleading, especially for open-QA tasks like our task.

Prompt: I have a severe headache and sensitivity to light. What could this be?

Response: HiThanks for using healthcare magicI think you have a severe headache due to light exposure. If it is mildly irritating, then please consult a psychiatrist online who can help you find out more information about migraine treatment options available here. Hope this helps! Regarding Headache: For further queries consult an internal medicine physician online

The generated response for asking about headache due to light exposure. The model showed a little knowledge in the medical field, as in the response it started with a polite opening learned from the data “thanks for using healthcare magic”. Then stated the reason for the headache and suggested visiting a doctor to help in the treatment. But at the end of the response there is a common sentence the model often uses in the other prompts even if it is irrelevant to the prompt, which is “for further queries consult an internal medicine physician online”. This may be due to repeating this sentence in the training process. However, most of the generated answers are relevant to the prompt, but it may provide incorrect medical advice and

wrong medical facts. That shows that fine-tuning successfully made a general uses model like T5 to learn and generate medical answers for our task.

Chatbot creation

For the chatbot generation we used the same approach used for the BioGPT, but we only changed, loading the model and the tokenizer and the pipeline used for the bot usage.

Downsides & Limitations:

1. Limited Examples: 40,000 samples insufficient for medical domain complexity.
2. Small Context window: 512 Tokens only is preventing the model from handling detailed medical histories, and could cause losing the context of the conversation and misleading responses.
3. Base Model parameters: T5-Base Parameters is so small (~220M) which leads to bad pattern recognition and making the model to struggle with complex medical relationships.

Comparison and conclusion

Feature	BioGPT (LoRA Fine-Tuned)	T5-Base (Conventional Fine-Tuning)
Architecture	Decoder-only (GPT-2 based)	Encoder-decoder
Pretraining Domain	Biomedical literature	General-purpose
Parameters	1.5 billion	220 million
Fine-Tuning Method	LoRA (Low-Rank Adaptation) with 4-bit quantization	Partial freezing (first 9 encoder layers frozen)
Training Efficiency	High (resource-friendly, low memory usage)	Moderate (higher resource demands)
Training Data Size	180K samples	40k samples

Strengths	<ul style="list-style-type: none"> - Domain-specific knowledge - Efficient training - Medically coherent responses 	<ul style="list-style-type: none"> - Better contextual understanding - Handles variable-length sequences
Weaknesses	<ul style="list-style-type: none"> - Weak conversational memory - Repetitive phrasing - Limited dialogue optimization 	<ul style="list-style-type: none"> - Smaller model size limits complexity - Occasional incorrect advice - Repetitive patterns
Automatic Metrics	Low ROUGE/BLEU (misleading for QA)	Low ROUGE/BLEU (misleading for QA)
Qualitative Results	Professional tone, medically plausible answers	Coherent but sometimes irrelevant/incorrect advice
Best For	Resource-constrained, domain-specific applications	Tasks needing deeper context (with more data/resources)

Conclusion

The study compared two fine-tuning strategies for developing a medical chatbot: LoRA-based adaptation of BioGPT and conventional fine-tuning of T5-Base. Key findings include:

1. BioGPT with LoRA demonstrated superior domain-specific performance due to its biomedical pretraining and efficient fine-tuning. It produced medically relevant responses but was limited by its decoder-only architecture and context window.
2. T5-Base showed adaptability to the medical domain but struggled with accuracy and coherence, partly due to its smaller size and limited training data. Its encoder-decoder architecture offered better contextual understanding but at higher computational costs.