

Build a biomedical chatbot by fine-tuning a pretrained large language model

Group 5
Xi Ouyang 50015835
Jingyuan Miao 50015823

2024-04-26

CONTENTS

1. Content Overview
2. Data Collection & Processing
3. Fine-Tuning & Evaluation
4. Analysis & Reflection

01

Content Overview

Fine-tune a LLama-2-7B-Chat pretrained model with corpora from the biomedical domain to build a biomedical chatbot.

Our target is to build a reliable and helpful chatbot which can be used in our daily life with only a little data. And we have succeeded in building such one.

- An Application example:

Hello doctor,I am having an uncomfortable feeling in between the middle of my spine and left shoulder blade. It seems to get stiff, and my bones pop a lot around there. What am I dealing with? Right now I have no pain or tender near that area still relief but curious on what happened and why I got this shock and that popping feeling so loud it scared me half to death.



Hi,Thanks for your query.From description it seems to be due to muscle spasm.I would suggest you to take muscle relaxant like chlorzoxazone for few days. You can take analgesic like ibuprofen for pain relief. Do ice fomentation and take rest. You can discuss with your doctor about it. You may get blood tests like CBC, ESR, RA factor to rule out connective tissue disorders.I do hope that you have found something helpful and I will be glad to answer any further query.Take care





香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

02

Data Processing

1. The data used for fine-tune: Pile

PubMed Central: The PubMed Central dataset is a subset of the PubMed online repository for biomedical articles run by the United States of America's National Center for Biotechnology Information (NCBI).

PubMed Abstracts: The PubMed Abstracts dataset comprises the abstracts of 30 million publications in the PubMed online repository for biomedical articles.

Through this two dataset, we plan to Improve sensitivity about medicine area and medical Logical thinking ability.

Actually, necessary knowledge that more related to our daily life is already contained in pretrained model, our work is to make the output of the model more useful and reliable.

PCI Alternative Using Sustained Exercise (PAUSE): Rationale and trial design. Cardiovascular disease (CVD) currently claims nearly one million lives yearly in the US, accounting for nearly 40% of all deaths. Coronary artery disease (CAD) accounts for the largest number of these deaths. While efforts aimed at treating CAD in recent decades have concentrated on surgical and catheter-based interventions, limited resources have been directed toward prevention and rehabilitation. CAD is commonly treated using percutaneous coronary intervention (PCI), and this treatment has increased exponentially since its adoption over three decades ago. Recent questions have been raised regarding the cost-effectiveness of PCI, the extent to which PCI is overused, and whether selected patients may benefit from optimal medical therapy in lieu of PCI. ...

2. The data used for fine-tune: ai-medical-chatbot

A set of dialogues between a Patient and a Doctor.

Example:

Patient: ["Hello doctor,I had an emergency surgery six months ago punched through a window right arm cut both arteries cut radial nerve with ulnar damage and cut bicep tendon nerves back now almost full thumbs up the only issue is bicep. My question is there something I can do to fix it like lengthening surgery, or am I stuck? I am a boxer supposed to go Olympics this year. My surgeon focused on the nerve and did not care help. PS pronator is still little and has not fully recovered."]

Doctor:["Hello. If you are saying it is already six months since the surgery and if you are doing regular physiotherapy, still has not gained much of the actions, then it is not worth waiting anymore. There are many muscle transfer surgeries which can give you the desired movements and help you in day today activities. But you have to remember that post such operations you will require time to train the brain and get the full use of such movements. That will take time. Do not want to demotivate you, but tremendous training and hard work might be required to reach the pre trauma stage of your boxing."]

Data processing

The raw texts are composed of different formats, some of them are papers, some are dialogues, reports and others.

We plan to apply a method to transform the raw texts to reading comprehension texts, which turns out to be more useful in model prompts just like practice after reading and improves the ability to answer questions based on the learned knowledge.

The reading comprehension texts can be easily converted to dialogues which achieve our requirement.

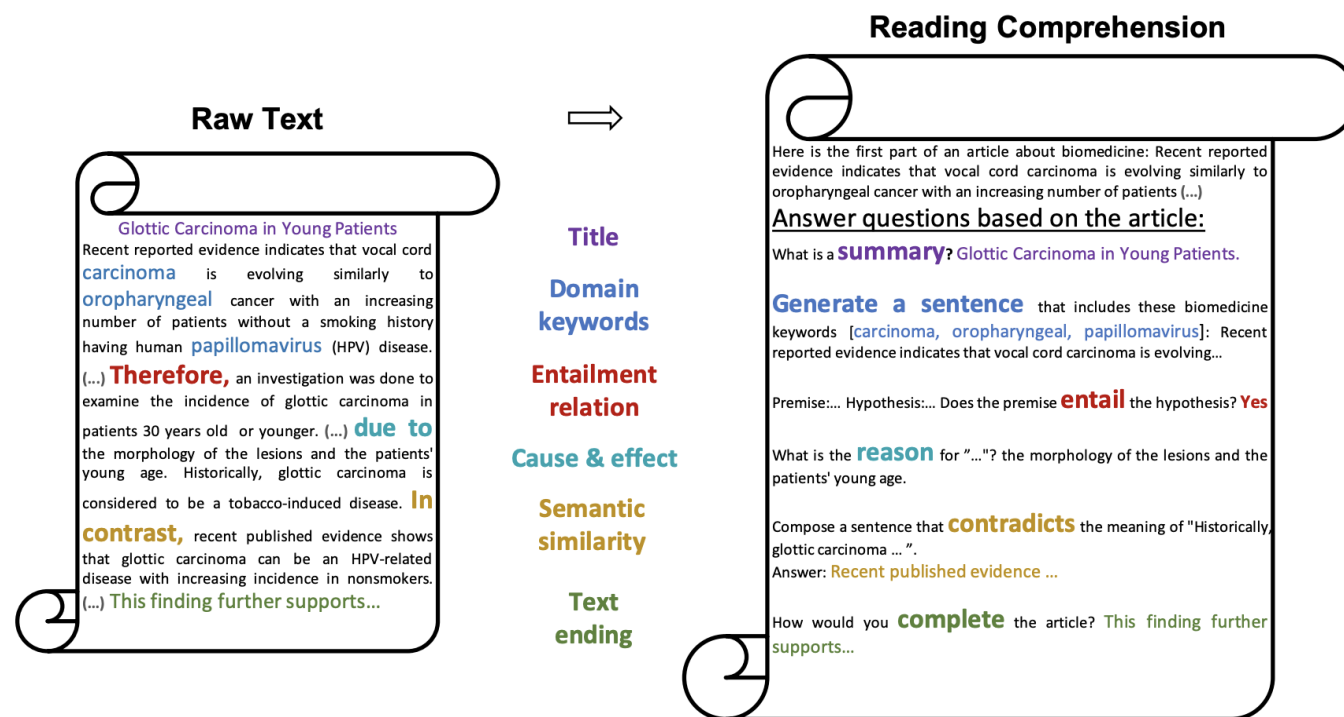


Figure 2: A simplified example of a reading comprehension text, wherein the raw text is followed by a series of tasks constructed from it, including Summarization (purple), Word-to-Text (blue), Natural Language Inference (red), Commonsense Reasoning (teal), Paraphrase Detection (yellow), and Text Completion (green). The complete version is in Appendix G.

Reference paper: [Adapting Large Language Models via Reading Comprehension](#)

Data processing

Before

[Construction of a deprivation index by Basic Healthcare Area in Aragon using Population and Housing Census 2011].
The measurement of inequalities using composite indicators facilitates the prioritization and implementation of public health actions. The most commonly...

After

```
{"conversations": [{"from": "human", "value": "Review the initial portion of an article discussing  
biomedicine: The measurement of inequalities using composite indicators facilitates the  
prioritization ....."},  
{"from": "gpt", "value": "Spearman correlation coefficients were used to evaluate the relationship  
between socioeconomic indicators and Standardized Mortality Ratios (SMR)."},  
{"from": "human", "value": " Here are some biomedicine words: socioeconomic, correlated,  
indicators. Tell me a sentence that describes them in the article."},  
{"from": "gpt", "value": "17 socioeconomic indicators were significantly correlated with SMR."}]}
```

Data for evaluation

MedQA-USMLE-4- options

Prompt: A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon ...

Which of the following is the best treatment for this patient?

Input: { "A": "Ampicillin", "B": "Ceftriaxone", "C": "Doxycycline", "D": "Nitrofurantoin" }

Label: "D":Nitrofurantoin

Load data for LLAMA Factory

```
"med_chatbot": {  
  "file_name": "ai-medical-  
chatbot_train.json",  
  "columns": {  
    "prompt": "instruction",  
    "response": "output"  
  }  
},
```

```
"pile_med": {  
  "file_name": "pile_med_100000.json",  
  "formatting": "sharegpt",  
  "columns": {  
    "messages": "conversations"  
  },  
  "tags": {  
    "role_tag": "from",  
    "content_tag": "value",  
    "user_tag": "human",  
    "assistant_tag": "gpt"  
  }  
},
```



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

03

Fine-Tuning & evaluation

At this stage, we use LoRA (Low-Rank Adaptation) to fine-tune the model. The key hyperparameters are as follows:

LoRA rank The rank of LoRA matrices.	LoRA Alpha Lora scaling coefficient.	LoRA Dropout Dropout ratio of LoRA weights.
8	16	0.1

The key training hyperparameters are as follows:

Cutoff length Max tokens in input sequence.	Learning rate Initial learning rate for AdamW.	Epochs Total number of training epochs to perform.	Max samples Maximum samples per dataset.	Compute type Whether to use mixed precision training (fp16 or bf16).
5120	5e-5	1.0	100000	<input checked="" type="radio"/> fp16 <input type="radio"/> bf16 <input type="radio"/> fp32
Batch size Number of samples processed on each GPU.	Gradient accumulation Number of steps for gradient accumulation.	LR scheduler Name of the learning rate scheduler.	Maximum gradient norm Norm for gradient clipping.	Val size Proportion of data in the dev set.
2	8	cosine	1.0	0

By using LoRA, we train 4,194,304 parameters, which are only 0.0622% of all 6,742,609,920 trainable parameters:

```
trainable params: 4194304 || all params: 6742609920 || trainable%: 0.0622
```

We use a subset of “pile” dataset in biomedical domain (100,000 samples in total) to fine-tune the base model. The training results are as follows (training on a single A800 GPU)

```
"epoch": 1.0,  
"train_loss": 0.07294103698730468,  
"train_runtime": 39890.1742,  
"train_samples_per_second": 2.507,  
"train_steps_per_second": 0.157
```


We evaluate the domain knowledge of models on MedQA testset. To facilitate the evaluation, we further fine-tuned both base and trained model with 100 samples from MedQA trainingset to convert the models' output to a certain format.

base:

```
"predict_bleu-4": 41.86678342498036,  
"predict_rouge-1": 50.47627729772192,  
"predict_rouge-2": 32.96360652003142,  
"predict_rouge-l": 50.149619481539666,  
"predict_runtime": 259.8225,  
"predict_samples_per_second": 4.899,  
"predict_steps_per_second": 0.616
```

The accuracy of base model is 28.75%



fine-tuned:

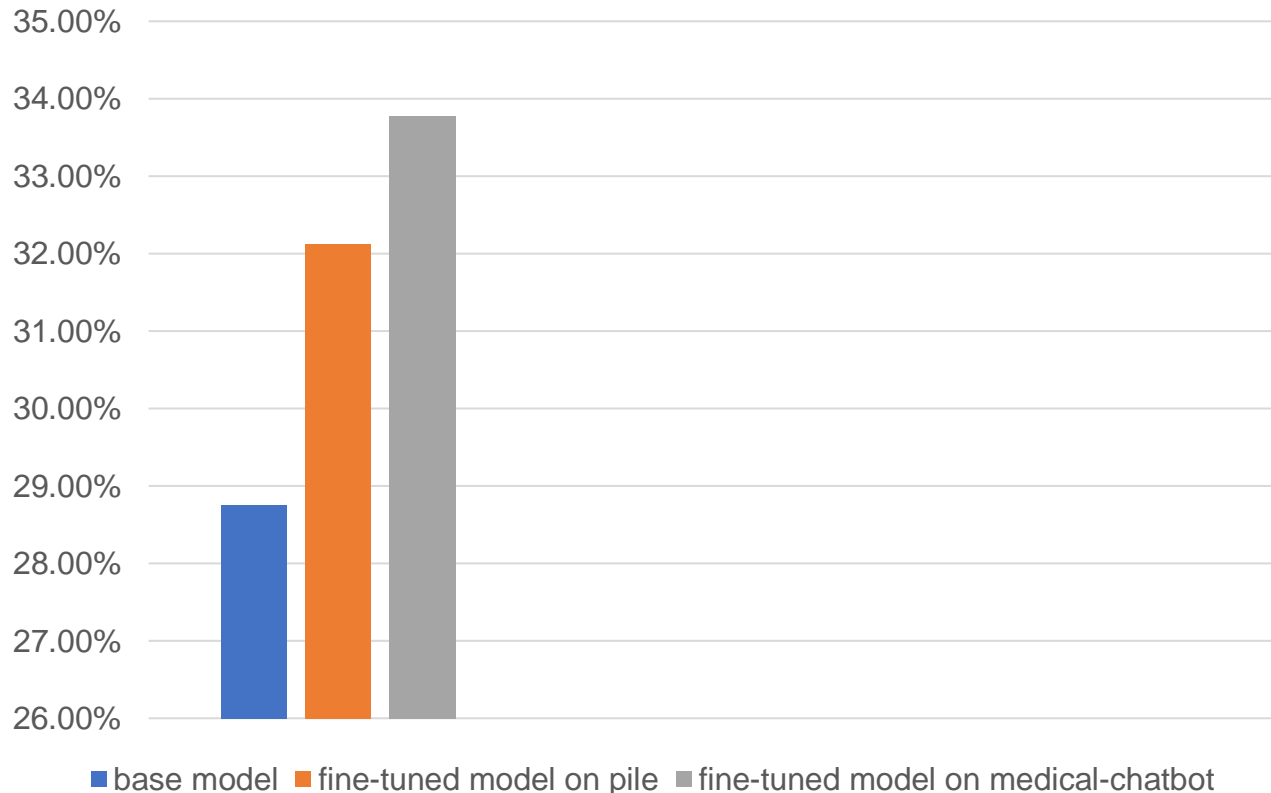
```
"predict_bleu-4": 45.27716292223095,  
"predict_rouge-1": 52.96794658287509,  
"predict_rouge-2": 36.1075499607227,  
"predict_rouge-l": 52.52535797329144,  
"predict_runtime": 266.4192,  
"predict_samples_per_second": 4.778,  
"predict_steps_per_second": 0.601
```

The accuracy of fine-tuned model is 32.13%

The MedQA testset consists of four-choice questions. Since in each sample, A, B, C, and D represent different contents, they cannot be viewed as four classes to use the “F1 score” as an evaluation metric. Therefore, we have used “accuracy” as the metric here.

Since the improvement by fine-tuning with the subset of pile is very limited, we turn to use another dataset named “ai-medical-chatbot”, which contains several dialogues between patients and doctors. We randomly choose 80,000 samples from this dataset to fine-tune the base model, and the comparison among these 3 models’ performance on MedQA testset are as follows:

Accuracy on MedQA testset



Compared to that trained on pile, fine-tuned model on ai-medical-chatbot dataset achieves higher accuracy. What’s more, the latter model takes less time to train than the former one (3h, while 11.1h the former one). This may imply the ai-medical-chatbot dataset is a better choice to equip model with biomedical domain knowledge, at least for this MedQA task.

Besides the domain knowledge, we also concern the models' text generation ability, can they generate natural and content-rich text in biomedical chatting. Here is an example among these models.

Query: Hello doctor,I am having an uncomfortable feeling in between the middle of my spine and left shoulder blade. It seems to get stiff, and my bones pop a lot around there, and it is very uncomfortable to sit in specific ways feels like my bones shift, well the other night it seemed as if my shoulder would pop out of place causing more stiffness after a while of dealing with it. I ended up feeling a popping sensation, where the stiffness was an instant relief. It was so shocking it put my anxiety through the roof. I thought I was going to die because my whole body was shaking and tingling. What am I dealing with? Right now I have no pain or tender near that area still relief but curious on what happened and why I got this shock and that popping feeling so loud it scared me half to death.

Label: Hello. The popping and discomfort what you felt is either because of improper scapulothoracic mobility or may be due to facetjoint irritation in the neck with intervertebral hypermobility. Which could have caused due to stiffness in the chest muscles, shoulder stiffness, etc. I would recommend you to do a lot of icing in the chest and back muscles. Maintain your neck and upper back posture by doing some back retraction work. Gently stretch your neck and chest. Gently massage your chest and shoulders to relieve tension in the muscles. I hope this is helpful for now. Kindly revert in case you need any help in this regard.

Base model: Based on the symptoms you've described, it's important to note that you should consult with a medical professional for proper evaluation and diagnosis. The sensation of stiffness, popping, and shifting in your spine and shoulder blade could be related to several conditions, including musculoskeletal issues, nerve compression, or even a more serious underlying medical condition. As a helpful and respectful assistant, I must advise you to seek medical attention to rule out any potential health concerns. Please prioritize your well-being and seek medical attention if your symptoms persist or worsen over time. If you have any further questions or concerns, please feel free to ask.

Fine-tuned model on medical-chatbot: Hello, It could be due to a pinched nerve. As of now you can use analgesics/anti-inflammatory combination like aceclofenac/serratiopeptidase for symptomatic relief. If symptoms persist, it is better to consult a physician and get evaluated. Hope I have answered your query. Let me know if I can assist you further.

Compared to the base model, the fine-tuned model can give a clearer diagnosis of symptoms and more specific medical advice, rather than just suggesting that users consult a medical professional.



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

04

Analysis & Reflection

After analyze the results, we wonder why the model fine-tuned on the subset of pile not improve significantly. After discussion, we came up with several possible causes and will conduct further research towards them:

1. The tasks in the training data may have little relevance to the tasks at the time of evaluation.
2. There may be still too little training data and too little diversity to “inject” sufficient domain knowledge into the model.
3. The number of training epochs, learning rate and other hyperparameter Settings may still have room for further optimization



Thanks