ACADEMIC REPORT
PRESENTATION

Fine-tune Distilled Large Language Models for Medical Question Answering

Group Geranimus

Ayush Sharma, Ningyi Ke, Yiwen Wang

2024·04·05

CONTENTES

#01

Problem Statement

Motivation

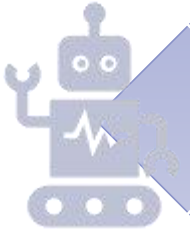
Patients' view

The growing need for reliable automated medical information providers vs. Limited accessibility to healthcare professionals online.

Healthcare Professionals' view

The lengthy process of accumulating experience-based knowledge and the pressure of staying updated with the latest medical literature.

Technical Feasibility



Large Language Models are trained using abundant online materials, including medical-related information, and they possess outstanding generative abilities for question-answering tasks.

Ref: Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions?.

Project Objective

Goal

Develop a **closed-book question answering system** that can answer questions related to diagnosis, medication, treatment, and side effect.

Input

A medical-related question: Who is at risk for Hairy Cell Leukemia?

Output

An accurate and reliable answer: Gender and age may affect the risk of hairy cell leukemia. Anything that increases your chance of getting a disease is called a risk factor. Having a risk factor does not mean that you will get cancer; not having risk factors doesn’t mean that you will not get cancer.

#02

Dataset

Dataset: MedQuAD & iCliniq

	MedQuAD	iCliniq
data	Question and answer pairs	Question and answer pairs
sources	12 National Institution of Health(NIH) websites	eHealth Forum, iCliniq, Question Doctors, and WebMD
size	47,457	29,752
answer length	Mean: 198 tokens Median: 137 tokens	Mean: 91 tokens Median: 60 tokens
characteristic	• High-quality contents supported by the authority backing of the NIH. • **Consists of medical expertise questions**.	• Rich data sources contribute to the diversity and comprehensiveness of the data. • **Consists of conversation-like medical consultations**.

Ref: Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering BMC Bioinform., 20(1):511:1–511:23.; Lasse Regin Nielsen. 2017. Medical Question Answer Data.

Dataset: MedQuAD & iCliniq

	MedQuAD	iCliniq
Example question and answer	How is SCOT deficiency diagnosed?	Will brushing my teeth with baking soda and sea salt increase my sodium intake? . I'm on a low sodium diet
	Diagnosis of SCOT deficiency is made in people showing the signs and symptoms of the condition and who have absent or reduced SCOT enzyme activity.	If you are careful not to swallow it and rinse well afterwards it shouldn't count as sodium intake. check with your dentist on how frequently you should be doing this

Ref: Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering BMC Bioinform., 20(1):511:1–511:23.; Lasse Regin Nielsen. 2017. Medical Question Answer Data.

#03
Approach

Approach: Fine-tune Distilled LLMs

Data Processing

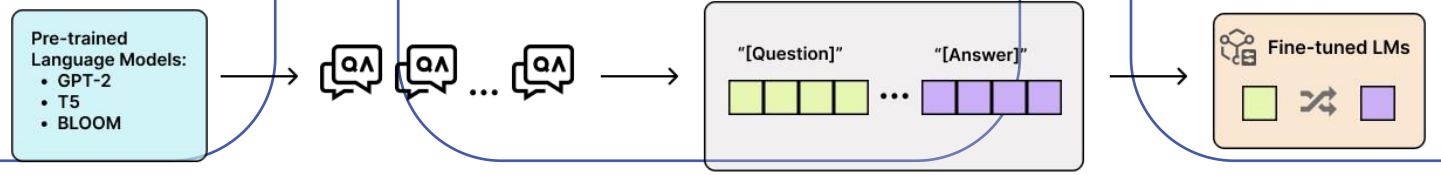
- Remove special tokens and extra spaces.
- Cut Answer length to 300 tokens.
- Concatenation
- Tokenization

Full Fine-tuning

- **Encoder-decoder Structure**
  - ➢ Distilled T5
- **Decoder-only Structure**
  - ➢ Distilled GPT2
  - ➢ Distilled BLOOM

Evaluation

- **Quantitative**
  - ➢ SacreBLEU
  - ➢ ROUGE
- **Qualitative**
  - ➢ Generated answer analysis



Ref: Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. MedLM: Exploring Language Models for Medical Question Answering Systems.

#04

Results and Analysis

Results: Quantitative Analysis

	sacreBLEU	Rouge-1	Rouge-L
Iclinq			
GPT-2	0.244	0.081	0.08
Bloom	0.725	0.086	0.069
T5	0.670	0.093	0.080
MedQuAD			
GPT-2	6.027	0.208	0.175
Bloom	14.109	0.267	0.219
T5	9.382	0.244	0.203

	sacreBLEU	Rouge-1	Rouge-L
Combined: Iclinq + MedQuAD			
GPT-2	4.12	0.117	0.103
Bloom	4.325	0.135	0.113
T5	4.569	0.139	0.120

Results: Qualitative Analysis on iCiniq

Question: my ear hurts and so does my teeth/jaw on the same side do i go see a doctor first or a dentist?

Reference Answer
you may want to start with your medical provider so that you can be treated if there is a ear-related problem however seeing a dentist first would be okay too. dental and jaw issues like tmj disorders can cause pain to be "referred" to the ear.

BLOOM generated
i would go see a dentist first. i am not sure if you have a tooth problem or not but if you have a tooth problem you should see a dentist. a tooth problem can be caused by a tooth eruption or a tooth fracture. a tooth eruption can be caused by a tooth being too small or too hard to be erupted…

Model repeats itself

T5 generated
if you have any problems with your ear or jaw it is best to see a dentist.

Model being too precise and not doing further elaboration

Results: Qualitative Analysis on MedQuAD

Question: How many people are affected by Pallister-Hall syndrome ?

Reference Answer
This condition is very rare; its prevalence is unknown.

GPT2 generated
PallisterHall syndrome is a rare disorder in which the body is unable to process the body's own tissues and organs. It is estimated to occur in about 1 in 1 in 2 in 2.5 to 2. The body. The exact incidence of PallisterHall syndrome is 1 in 1 in 2. The exact number of affected individuals. The frequency of the body is 1 in 2. The frequency of the body is 2

Model generates meaningless sentence, hallucinating, and repeats itself

T5 generated
PallisterHall syndrome is a rare disorder that affects about 1 in 100,000 people worldwide.

Model may be hallucinating about the percentage

#05
Conclusion

Conclusion

Key Findings:

- Distilled **T5** performed better on the **iCliniq** dataset than the other two models. This is likely because the encoder component can better capture the context of human questions, enabling it to generate more reasonable and precise responses.
- Distilled **BLOOM** outperformed the other two models on the **MedQuAD** dataset. This may be due to differences in the amount of medical expertise knowledge that these three models were trained on.
- **Decoder-only models suffer from repetition**, whereas encoder-decoder model does not. The encoder component provides representation information to the decoder, which helps T5 generate more coherent and non-repetitive responses.
- **All three models hallucinate**, which makes the responses less reliable. An open-book QA system may be more suitable in medical QA task.

THANKS

For Your Attention