Final Project Report: Fine-tuning Distilled Large Language Models for Medical Question Answering

Group Geranimus: Ayush Sharma, Yiwen Wang, Ningyi Ke

Abstract

Large language models(LLMs) have demonstrated their excellent generative capabilities in various natural language processing (NLP) tasks. These models have been fine-tuned to support domain-specific tasks more effectively. However, there remains a significant gap within the Medical Question-Answering (MQA) domain, partly due to the lack of medical-related datasets and the absence of close expert evaluation of the results. In this project, our goal is to explore the fine-tuning of different types of distilled language models for Closed-Book MQA and compare their performance. Specifically, we aim to determine whether the decoder-only architectures (such as GPT-2 and BLOOM) or the encoder-only architecture (such as t5) is better suited for the MQA task.

1 Introduction

Medical resource scarcity has long been a pressing issue in need of remedy. An aging workforce and a global lack of doctors are exacerbated by the already dire situation and making the shortage even worse. According to data from the World Health Organization (WHO), driven by models of population growth, a frightening estimate has been made: by the year 2035, the global shortage of skilled health professionals is expected to reach to an alarming 12.9 million. (World Health Organization, 2009) Overburdened healthcare professionals and patients experiencing delays in accessing necessary medical care represent significant consequences of this scarcity.

Yet, with the advancement of large language models and the accumulation of healthcare-related data, the development of precise question-and-answer systems tailored to the medical domain has become feasible. Large Language Models (LLMs) such as GPT (Generative Pretrained Transformer) are a breakthrough in AI technology that uses deep learning and natural language processing to analyze and generate human-like text. Through complex algorithms and extensive training on various text data, these models can perform various tasks without the need for domain-specific data training. LLMs' great potential has resulted in their broad acceptance in our daily lives, with applications ranging from media to business, entertainment to education. This even provides opportunities for those without a technical background to use these models to draft articles, create music, or write simple programs, helping them improve their creativity, productivity, and learning abilities.

Thus, our motivation is to let healthcare workers benefit from the automated information summarizing and generation power of LLMs, and also let patients have access to reliable and convenient medical inquiry systems. Our objective is to develop closed-book question-answering systems that can answer questions related to diagnosis, medication, treatment, and side effects. This means the system input will be one medical-related question, and the system will need to rely solely on its pre-trained knowledge to answer this question. We plan to build the system by fine-tuning powerful LLMs. Our first motivation is to build a model that can give precise answers when no context is given. To find a suitable base LLM, we plan to experiment with three different types of distilled LMs that have different underlying architectures. That is to say, we also want to compare the performance decoder-only model and encoder-only model for the MQA task.

2 Related Work

Question answering systems are designed to automatically answer questions by using natural language corpus or information stored inside databases. Classified by the implementation paradigm, question answering system can be classified as follows: (Calijorne Soares and Parreiras, 2020)

- **Information Retrieval QA:** It uses searching algorithms to find answers on the Internet or a collection of documents, then filter out the best suitable information as the answer.

- **Natural Language Processing QA:** It uses machine learning models to extract and generate answers based on the information retrieved. This kind of model is a two-staged model, that usually consists of an information retrieval component and an LM that can process the information and then generate answers.

- **Knowledge Base QA:** It uses structured data stored in databases, and uses standard database query language while finding answers.

- **Hybrid QA:** A hybrid QA system will combine the above three kinds of systems for answering.

Multiple research has been conducted in the field of Natural Language Processing QA system design. Here we summarize and introduce some works focused on using LLMs:

(Singhal et al., 2022) first proposed an instruction-tuned PaLM MQA system named Med-PaLM that reaches 67.6% accuracy on the MedQA (US Medical License Exam questions) dataset. To be noticed the MedQA dataset is set to be in the form of multiple choice selection instead of the traditional question answering pair format. More recently, (Singhal et al., 2023) further developed Med-PaLM 2 which leveraged PaLM 2, which is an improved version of PaLM, instruction tuning, and prompt tuning strategies like ensemble refinement. The Med-PaLM 2 scored up to 86.5% on the MedQA dataset.

(Wang et al., 2024) used a two-staged retriever/reader model called Joint Medical LLM and Retrieval Training (JMLR) that combined information retrieval using rank loss and LLM to generate the final answer. JMLR used a BERT-based model to generate document embeddings and then updated the retrieval loss along with the answer generation loss together. JMLR reached 61.3% on the MedQA dataset.

(Lyu et al., 2024) identified that vanilla fine-tuned LLMs have the problem of generating incomplete, non-factual, or illogical answers. Their hypothesis is that these problems stem from inadequate knowledge awareness during traditional fine-tuning. Thus they proposed a new fine-tune technique called Knowledge-aware fine-tuning(KnowTuning), which uses a two-stage strategy that can help LLM to explicitly extract knowledge triples from answers and implicitly teach them to distinguish reliable and unreliable knowledge. They evaluated the quality of the model by using GPT-4 to rate their answers in completeness, actuality, and logicality on the MedQuAD dataset.

In our project, we chose the second paradigm because we want to leverage modern LLM's previous access to rich online corpus during the pre-trained phase, and also utilize their strong text understanding and generation ability. It is proven that LLMs can get good performance on medical knowledge multiple choice questions with only prompt tuning, thus we are confident that LLMs are equipped with a certain amount of medical knowledge that can support a closed-book question answering system construction. (Liévin et al., 2023) The choice of building a closed-book QA system also considered the model response time and extra bias introduced if using external data sources.

3 Approach

Our project is inspired by methodologies from MedML (Yagnik et al., 2024), which we employ as a conceptual framework. However, all code for our project is being written by us from scratch. This approach allows us to ensure that our implementation is fully original to our project's needs, enhancing its uniqueness and effectiveness.

The example model input question and output answer should look like this:

Question: *Who is at risk for Hairy Cell Leukemia?*

Reference 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. Talk with your doctor if you think you may be at risk. The cause of hairy cell leukemia is unknown. It occurs more often in older men.*

In this project, we aim to build a reliable medical question-and-answer system with high performance. We plan to achieve this by fine-tuning three pre-trained Language Models and comparing their performances.

In terms of Language Models, we will continue to use the same models as the reference paper, which are Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019), Text-to-Text Transfer Transformer(T5) (Raffel et al., 2020), and BigScience Large Open-science Open-access Multilingual Language Model(BLOOM) (BigScience Workshop, 2022). Considering our available computing resources, we decided to use the distilled version of LLMs, which are generated by knowledge-transferring of the original LLMs, allowing them to achieve comparable performance while being more lightweight and suitable for practical fine-tuning.

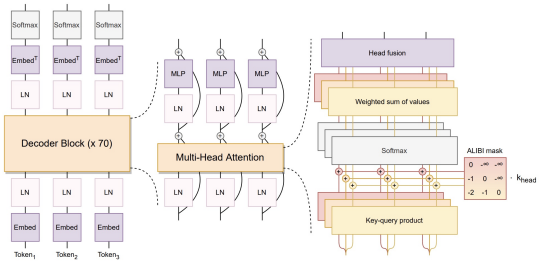

Figure 1: BLOOM Architecture (Scao et al., 2022)

To expand on these three models, GPT-2 is a decoder-only Transformer model, which can help the model focus on generating text and capture subtle differences in human language. t5, on the other hand, is an encoder-decoder Transformer model, which uses the encoders to process input text and capture the relationship between words and text, while the decoder uses this representation to generate the labeled output text as the output. As for Model BLOOM, it is an auto-regressive, decoder-only Transformer model that has a slightly different model architecture compared to GPT-2, it uses ALiBi Positional Embeddings and an extra Embedding LayerNorm followed.

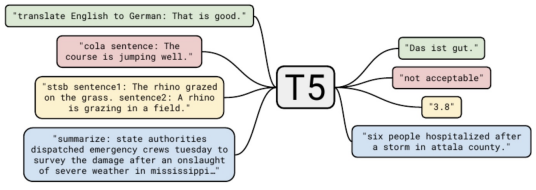

Figure 2: T5: unified text-to-text training example (Raffel et al., 2019)

Regarding our project approach, we intend to fine-tune the different distilled versions of LMs, followed by a comparative analysis between the performance of these fine-tuned models and that of the general model we tested before. Finally, we will compare the performance of the encoder-decoder models and the decoder-only models to find out which model is the most effective for our needs.

In the process of fine-tuning our model, we begin by combining the text from the questions and answers in the dataset into a cohesive text sequence, followed by thorough data preprocessing. We then standardize the length of answers, truncating them to a uniform and optimal size. This will not only help the model better capture the relationship between questions and answers while also reducing the amount of time needed for training. Furthermore, we will modify our tokenization strategies to align with the different architectures of various models, transforming raw text into a structured representation that the model can understand. We will use full fine-tuning that will update all parameters of the model. Finally, we will employ metrics for the evaluation of the effectiveness and check if the model's performance is enhanced.

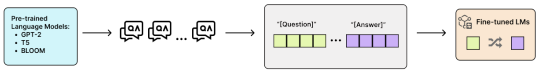

Figure 3: Finetuning Models

4 Experimental

4.1 Dataset Information

In this project, we used the datasets as in the references: MedQuAD (Ben Abacha and Demner-Fushman, 2019) and iCliniq (Regin, 2017), which both contain medical questions and answer pairs. Our objective in using the original datasets is to ensure the availability and reliability of the data, thereby providing a solid foundation for our analysis. In the following sections, we will go deeper into each dataset, showing their respective information content and unique features.

4.1.1 MedQuAD Dataset

The MedQuAD dataset is sourced from 12 websites related to the National Institutes of Health (NIH) in the United States. This is a large repository which contains 47457 entries. Each answer in this dataset has a considerable length, with an average of 198 tokens and a median of 137 tokens. The use of MedQuAD as training data is due to two main criteria. Firstly, it has high-quality content and authoritative support from the National Institutes of Health (NIH) in the United States, en-

suring that information is reliable and scientifically sound. Secondly, this dataset is derived from questions based on medical expertise, which implies that the data is highly specialized and relevant to the professional medical community. One example of MeQuAD is as follows:

Question: *What are the symptoms of Adult Acute Myeloid Leukemia?*

Reference Answer: *Signs and symptoms of adult AML include fever, feeling tired, and easy bruising or bleeding. The early signs and symptoms of AML may be like those caused by the flu or other common diseases. Check with your doctor if you have any of the following: Fever. Shortness of breath. Easy bruising or bleeding. Petechiae (flat, pinpoint spots under the skin caused by bleeding). Weakness or feeling tired. Weight loss or loss of appetite.*

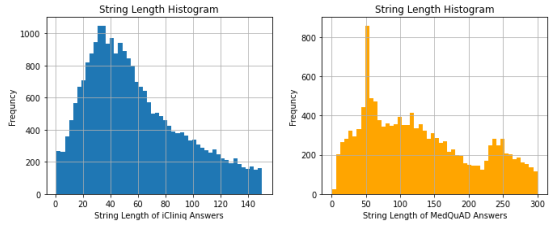

Figure 4: Distribution of the answer's length for MedQuAD and iCliniq datasets

4.1.2 iCliniq Dataset

Compared to the MedQuAD, iCliniq dataset includes a wider range of sources, including eHealth Forum, iCliniq itself, Question Doctors, and WebMD. Multiple sources provide the iCliniq dataset with rich diversity, containing a total of 29752 question and answer pairs. The answers in this dataset are relatively concise, with an average of 91 tokens and a median of 60. The uniqueness of iCliniq comes from its comprehensive and diverse dataset, which includes a wide range of medical inquiries and dialogues similar to actual medical consultations. This not only expands the dataset, but also reflects the discussion and communication that people may encounter in real-life medical scenarios. One example of iCliniq is as follows:

Question: *Can I take zotex while on cipro?*

Reference Answer: *This is a great question and very common for pharmacists to get during cough cold and allergy season. It is fine to take these two medications together. If you ever have any questions about possible drug interactions or other questions come and ask the pharmacist. we are here to help.*

4.1.3 Combined Dataset

In order to test these models' ability to handle both medical expertise knowledge and conversation-like medical question-answering, we build a new combined dataset by mixing MedQuAD and iCliniq together and randomly shuffling all samples.

4.2 Evaluation Metrics

In terms of evaluation metrics, we will evaluate the performance of LLMs in medical question-answering tasks using two evaluation metrics. First, we will use the SacreBLEU score (Post, 2018) to test how well the generated answer matches the reference answer. Second, we will utilize the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score (Lin, 2004) to evaluate how well the generated answer expresses the key information contained in the reference answer, hence determining if the model effectively transmits the information as predicted.

4.3 Training Details

- **Data preprocessing:**
  - Answer Truncation: To maintain computational efficiency and focus, answers were truncated to a maximum length of 300 words.
  - Special Token Concatenation: For the distilled GPT2 model, questions and answers were concatenated using a special [response] token, with additional [startoftext] and [endoftext] tokens marking the beginning and end of each QA pair. For the distilled BLOOM model, [question] is added before each question, and [answer] is added before each answer for the training data. As for testing data, [question] is added before the question, and [answer] is appended to the end of the question.
  - Tokenization: The dataset was tokenized using different tokenizers designed for different models, incorporating the aforementioned special tokens for better model comprehension.

- **Model Training and Configuration**

  For the fine-tuning of 3 models on 3 different datasets, we used different sets of training configurations. The core training hyperparameters are listed in Table 1.

Models	$GPT-2$	$BLOOM$	$T5$
batch size	16	2	N/A
epochs	2	2	2
learning rate	5e-4	5e-6	N/A
warm-up	1e2	2%	N/A

Table 1: Training Hyperparameter Configuration

- For distilled GPT2, AdamW was used as an optimizer, accompanied by a linear schedule with a warmup for the learning rate. Model performance was evaluated after each epoch, noting the average training loss as an indicator of model learning. The entire fine-tuning process of the distilgpt2 model spanned approximately 5 to 6 hours, with each epoch taking about 1 to 1.5 hours to complete.
- For distilled BLOOM, Adamfactor was used as an optimizer, which is a commonly used optimizer in NLP tasks. A warm-up ratio of 2% instead of fixed warm-up steps was used to make the training compatible with datasets of different sizes. The model performance was evaluated every 200 steps to monitor the training process. Since the distilled BLOOM has the largest parameter sizes, the fine-tuning process will take over 6 hours to complete for one epoch on the largest dataset, thus the epoch is cut to 1 when fine-tuning on the combined dataset.
- For distilled T5, we employed the SimpleT5 framework to fine-tune our T5 model, which automates the tuning of key hyperparameters. This framework has been pre-optimized with experimentally validated settings, eliminating the need for manual adjustments and ensuring stable performance. We used the default configuration of SimpleT5 (marked as 'N/A' in Table 1) to help simplify the process and focus on exploring model applications rather than hyperparameters.

5 Results

In this study, we fine-tuned three different pre-trained models: GPT-2, BLOOM, and T5, trained on three datasets (iCliniq, MedQuAD, and combined dataset), and obtained their performance results. In this section, we will first present the quantitative results of these models in the form of SacreBLEU and ROUGE scores. ROUGE-1 and ROUGE-L scores are selected because ROUGE-1 checked content overlap by matching individual words, while ROUGE-L evaluates the fluency and structure by considering the longest common subsequence, both could provide a more comprehensive view for our quantitative analysis. Then, we will conduct a qualitative analysis by presenting several examples of outputs generated by each model.

5.1 Quantitative Results

5.1.1 Pre- and Post-Fine-Tuning Performances

In this section, we compare the performance of models before and after fine-tuning to guarantee the effectiveness of our fine-tuning methods and further discuss the changes in these scores.

Models	$SacreBLEU$	$ROUGE-1$	$ROUGE-L$
(Pre) GPT-2	0.213	0.116	0.084
(Post) GPT-2	0.244	0.081	0.08
(Pre) BLOOM	0.623	0.096	0.012
(Post) BLOOM	0.725	0.086	0.069
(Pre) T5	0.465	0.113	0.016
(Post) T5	0.670	0.093	0.080

Table 2: Scores of Pre- and Post-Fine-Tuning Models for iCliniq Dataset

Models	$SacreBLEU$	$ROUGE-1$	$ROUGE-L$
(Pre) GPT-2	0.717	0.076	0.067
(Post) GPT-2	6.27	0.208	0.175
(Pre) BLOOM	1.676	0.139	0.025
(Post) BLOOM	14.109	0.267	0.219
(Pre) T5	0.006	0.058	0.023
(Post) T5	9.382	0.244	0.203

Table 3: Scores of Pre- and Post-Fine-Tuning Models for MedQuAD Dataset

Models	$SacreBLEU$	$ROUGE-1$	$ROUGE-L$
(Pre) GPT-2	0.891	0.120	0.086
(Post) GPT-2	4.12	0.117	0.103
(Pre) BLOOM	0.765	0.089	0.014
(Post) BLOOM	4.325	0.135	0.113
(Pre) T5	0.159	0.093	0.018
(Post) T5	4.569	0.139	0.120

Table 4: Scores of Pre- and Post-Fine-Tuning Models for Combined Dataset

According to the tables above, a clear pattern

appears when comparing the performance scores of different models on different datasets before and after fine-tuning. The enhanced performance of BLOOM and T5 for the iCliniq dataset (Table 2) shows a slight yet meaningful boost in the SacreBLEU and ROUGE-L metrics. However, GPT-2 only shows a slight increase in the SacreBLEU score and decreases in its ROUGE-1 and ROUGE-L scores.

When observing the MedQuAD dataset (Table 3), the performances of all 3 models after tuning demonstrate a significant increase in SacreBLEU, and also some improvements in ROUGE-1 and ROUGE-L scores. This indicates the models obtain a better ability to generate more accurate and contextually relevant outputs after fine-tuning.

Finally, for the Combined dataset (Table 4), all 3 models continue the trend of improvement after fine-tuning, with a sharp increase in SacreBLEU scores. At the same time, the ROUGE-1 and ROUGE-L scores of both models are steadily increasing for BLOOM and T5.

Overall, these results indicate that our fine-tuning has a significant positive impact on model performance on all three datasets, with the degree of improvement differing depending on the model and dataset.

5.1.2 Evaluation on the iCliniq Dataset

The first dataset, iCliniq, showed significant differences in performance between models in Table 5. Among the three models, Bloom achieved the highest SacreBLEU score of 0.725, significantly better than T5 and GPT-2, with scores of 0.670 and 0.244, respectively. In terms of ROUGE-1 and ROUGE-L scores, T5 leads with scores of 0.093 and 0.080, both higher than Bloom and GPT-2. This dataset highlights Bloom's better ability to get some fragments of words to be the same as the reference answer, while T5 performs better in capturing the reference main information. But generally speaking, all three models did not give satisfying answers on the iCliniq dataset.

Models	*SacreBLEU*	*ROUGE − 1*	*ROUGE − L*
GPT-2	0.244	0.081	0.08
BLOOM	0.725	0.086	0.069
T5	0.670	0.093	0.080

Table 5: Model Evaluation on iClniq Dataset

5.1.3 Evaluation on the MedQuAD Dataset

In general, all three models' performances are better compared to their performance on the iCliniq dataset. The evaluation results of the MedQuAD dataset emphasize the advantages of the Bloom model in answering medical facts, which consistently outperforms other models in all metrics (see Table 6). BLOOM's SacreBLEU score is 14.109, which is more than twice the GPT-2 and T5 scores. Similarly, Bloom has the highest ROUGE-1 and ROUGE-L scores, demonstrating its robustness in generating content that appears inside the reference answer.

Models	*SacreBLEU*	*ROUGE − 1*	*ROUGE − L*
GPT-2	6.27	0.208	0.175
BLOOM	14.109	0.267	0.219
T5	9.382	0.244	0.203

Table 6: Model Evaluation on MedQuAD Dataset

5.1.4 Evaluation on the Combined Dataset

The combined dataset integrates iCliniq and MedQuAD's datasets to provide comprehensive challenges. Here, T5 slightly leads BLOOM and GPT-2 with the highest score among all indicators (see Table 7). The SacreBLEU score of T5 is 4.569, and the ROUGE scores of 0.139 (ROUGE-1) and 0.120 (ROUGE-L) are the best. Since the scores of the three models are relatively close, we cannot assert which model performs better on this dataset.

Models	*SacreBLEU*	*ROUGE − 1*	*ROUGE − L*
GPT-2	4.12	0.117	0.103
BLOOM	4.325	0.125	0.113
T5	4.569	0.139	0.120

Table 7: Model Evaluation on Combined Dataset

5.2 Qualitative Results

In this section, we discuss some sample questions and the answers generated by these three models. Although quantitative metrics like SacreBLEU and ROUGE scores help track overall performance trends in an automated way, reviewing questions and predicted answers in each dataset can provide a deep insight into each model's strengths and limitations from another angle. For the iCliniq dataset, the example of prediction answers we obtained from distilled BLOOM and distilled T5 are as follows:

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 Answer: *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.*

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

From the above example, it can be observed that there is a significant difference in response between the two models. The reference answer suggested users seek help from a medical provider at the beginning and determine if the pain problem is related to the ears, but did not ask users to give up seeking advice from dentists. BLOOM, which scored highest in SacreBLEU, produced a lengthy response but with repetitive segments, indicating that potential verbosity does not add clarity for the users. T5 scored slightly lower in SacreBLEU and provided a shorter answer compared to BLOOM. For any ear or jaw problems, it is recommended to see a dentist. However, this conciseness may be considered too precise and lack detailed explanations, making it difficult for users to make appropriate decisions due to a lack of reference materials.

For the MedQuAD dataset, here are another two examples from distilled GPT-2 and distilled T5:

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

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

GPT-2 generated Answer: *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...*

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

The reference answer briefly discussed the rarity of this illness, but there is no particular prevalence data. The GPT-2's SacreBLEU and ROUGE scores are moderate, but the generated answer revealed that the GPT-2 has hallucinatory behavior, such as a prediction answer that seems plausible but is actually fictitious. Additionally, the answer is repetitive, which makes it impossible to guarantee the accuracy of the generated answer and may cause confusion for users looking for medical information. Another obvious problem of GPT-2 is the meaningless sentences and syntax errors generated, which makes the answer hard for humans to understand. T5 produced an approximate response for the prevalence rate and gave numerical values. This can give surface-level precision, but it also raises questions about potentially misleading statistical data generated by the model, which also reflects the "hallucinating" nature sometimes seen in language models where precise data is not given.

6 Analysis

In examining the performance of three distilled LLMs on the iCliniq and MedQuAD datasets, our analysis yields distinct insights into their capabilities for MQA.

The T5 model, with its encoder-decoder architecture, delivered the highest performance on the iCliniq dataset. This is evidenced by its leading ROUGE scores. As the questions in the iCliniq dataset are based on human context, we believe the performance of the T5 model is likely due to the ability of the encoder component to better capture the human context of the questions, which helps it provide better responses compared to other models.

For the MedQuAD dataset, BLOOM demonstrated superior performance. We reckon this might be due to the differences in the amount of medical expertise data present in the datasets, on which these models were pre-trained. We believe BLOOM is trained on more medical-related materials than the other two models.

In the case of the combined dataset, the difference in quantitative performance among GPT-2, BLOOM, and T5 on the combined dataset of iCliniq and MedQuAD is so marginal that it's difficult to conclusively determine which model performs best. But when we take the qualitative results into consideration, we found that T5 generates more correct and related answers. So in general,

T5 should be the most suitable model for the MQA task.

When we compared these fine-tuned models against their base models, we observed that the base models performed worse than the fine-tuned models, indicating that fine-tuning indeed provided a performance boost. This highlights the benefit of tailoring the models to specific domains for enhanced closed generative Question answer capabilities. However, even with these improvements from the base models, limitations were still evident. All three models had a tendency to hallucinate and provide users with incorrect data. Specifically, T5 and BLOOM demonstrated tendencies toward generating incoherent or repetitive responses, underscoring concerns about their reliability for medical question answering. The amount of time and computing resources used to perform full fine-tuning is also considerable, more light-weight fine-tuning techniques like LoRA can be used for further experiments.

7 Conclusion

This study evaluated the capabilities of three distilled LLMs in the context of MQA across two distinct datasets: iCliniq and MedQuAD. Our findings reveal that the T5 model, leveraging its encoder-decoder architecture, demonstrated superior performance on the iCliniq dataset which contains human context-based questions, likely due to its enhanced ability to capture and interpret the human context of queries. In contrast, BLOOM exhibited exceptional performance on the MedQuAD dataset, suggesting a potential advantage stemming from its pre-training on a dataset containing medically relevant data. Our analysis highlights the impact of fine-tuning on model performance. Despite the increase in improvements, we also saw its limitations, where the model had a high tendency to hallucinate, and the GPT-2 and BLOOM were also prone to generating incoherent or repetitive responses. These issues compromise the reliability of the models for delivering accurate medical information, a critical aspect when considering their application in the healthcare domain. Given these constraints, open book QA emerges as a more reliable approach in the realm of medical question answering. By leveraging external information sources during the answering process, open-book QA can potentially mitigate the risk of inaccuracies inherent in closed generative QA models.

8 Future Work

Future work may include using another quantitative analysis to better measure the quality of the generated answers. Usually, BLEU and ROUGE scores are designed for translation tasks and text summarization tasks respectively. A better way is to use LLMs, for instance, GPT-4, for comparison of how good the generated answer compared to the reference answer. Another possible future work is to use other kinds of fine-tuning techniques on T5 to further improve the QA system's performance.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

Marco Antonio Calijorne Soares and Fernando Silva Parreiras. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. Knowtuning: Knowledge-aware fine-tuning for large language models.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Lasse Regin. 2017. Medical question answer data.

Teven Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander Rush, Stella Biderman, Albert Webson, Pawan Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Moral, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability.

World Health Organization. 2009. Qa on the health workforce crisis.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. *MedLM: Exploring Language Models for Medical Question Answering Systems*. Available at arXiv:2401.11389 [cs.CL].