Improving Performance of Medical Document Querying using Dense Passage Retrieval

Haowei Qiu
McGill University, Computer Science
haowei.qiu@mail.mcgill.ca

Riley Ballachay
McGill University, Computer Science
riley.ballachay@mail.mcgill.ca

Abstract

Large language models (LLMs) have taken the world by storm, particularly as a question-answering tools for general purpose tasks. In more specific tasks, fine-tuning LLMs on specialized domain knowledge has demonstrated and effective way of creating a specialized tool for specialized question-answering tasks. Medical QA, a task that involves answering queries from medical professionals, has been a focal point for researchers. Existing medical QA datasets and models often focus on short questions of approximately 100 words, which may not accurately represent the extensive and specialized medical documentation that physicians encounter. These documents can span dozens of pages and are filled with intricate technical information. The LongHealth dataset was created to address this gap, providing a dataset of 20 patients, with 5000 words per patient of clinical notes and 400 accompanying questions about their content. The dataset is a promising step forward in bringing medical QA models into a clinical setting. This work builds upon that of the original authors by adding a retrieval step to extract relevant passages prior to passing the document to the chat models. It is shown that using top-10 passage retrieval increases model performance by up to 10% and reduces inference time by up to 3.2x. It is the hope that this work demonstrates the efficacy of retrieval in the domain of long medical documents.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capacity in a variety of tasks including language translation, text summarization, and question-answering (Wei et al., 2022). In particular, chat bots like ChatGPT have quickly gained international notoriety as general question-answering and multi-purpose cognitive aid. ChatGPT was pre-trained on a corpus containing Wikipedia, books and news articles. The wide scope and scale of the training data for ChatGPT has allowed for impressive performance in a variety of tasks, however it's performance has been shown to be limited in more specialized domains and tasks, falling far behind state of the art (SOTA) models (Kocoń et al., 2023). To unlock capabilities on specific tasks, pre-trained LLMs may be fine-tuned to specific tasks. With pre-training, ChatGPT has demonstrated capabilities in specific domains including healthcare, education, research, natural sciences and human-machine interaction (Liu et al., 2023). LLMs have been of great interest in healthcare, a field where deep-learning has already shown revolutionary potential (Thirunavukarasu et al., 2023). The plethora of information online and in medical literature makes it difficult for physicians to have up-to-date evidence to apply in their decisions. Question-answering LLMs for healthcare have largely focused on Open-Book information retrieval, where retrieval is used during inference to pull relevant information from medical references (Yagnik et al., 2024). There is a need for more personalized question-answering, addressing the unique physiology and narrative of each patient. Multiple QA models have been developed in this domain, including BioBERT for biomedical text mining (Lee et al., 2019), ClincalBERT for predicting hospital readmission from clinical notes (Huang et al., 2020) and Med-BERT for prediction of disease status from electronic health records (EHR) (Rasmy et al., 2020). Multiple datasets are used for training and evaluation of these models. MedMCQA, a large-scale, Multiple-Choice Question Answering dataset is composed of more than 194k medical entrance exam questions, each averaging 15 tokens in length. MedQA, a benchmark for general medical knowledge, has 61k questions containing patient symptoms of 100 words (Jin et al., 2020). See Figure 1, adapted from (Singhal et al., 2023) for a summary of multiple-choice questions from these datasets and others.

These datasets, and the accompanying models

Name	Count	Description
MedQA (USMLE)	1273	General medical knowledge in US medical licensing exam
PubMedQA	500	Closed-domain question answering given PubMed abstract
MedMCQA	4183	General medical knowledge in Indian medical entrance exams
MMLU-Clinical knowledge	265	Clinical knowledge multiple-choice questions
MMLU Medical genetics	100	Medical genetics multiple-choice questions
MMLU-Anatomy	135	Anatomy multiple-choice questions
MMLU-Professional medicine	272	Professional medicine multiple-choice questions
MMLU-College biology	144	College biology multiple-choice questions
MMLU-College medicine	173	College medicine multiple-choice questions

Figure 1: A summary of benchmark datasets for medical question answering, adapted from (Singhal et al., 2023).

trained to answer their queries, do not account for the complexity, length and redundancy of real-world clinical documentation (Adams et al., 2024). The most recent attempt to address this paucity of the LongHealth dataset, an open-source benchmark consisting of 20 fictional patient cases, with accompanying clinical notes ranging from 2 to 15 pages for a total of 5000-7000 words per case. The benchmark consists of 400 associated questions which fall under one of three categories: 1) The ability of the model to extract information from the text, 2) The ability of the model to extract relevant data when exposed to adversarial information and 3) The ability to refuse to answer questions when the relevant information is missing. The benchmark represents a realistic case of querying that a physician or assistant may want to perform on clinical notes from a patient. The publication tests a variety of models, shown below in Figure 2

Model Name	**Parameters**	**Context Length (Tokens)**
gpt-3.5-turbo-1106[2]	Unknown	16,000
Yi-6B-200K [15]	6B	200,000
Yi-34B-200k [15]	34B	200,000
vicuna-7b-v1.5-16k [16]	7B	16,000
vicuna-13b-v1.5-16k [16]	13B	16,000
longchat-7b-v1.5-32k [17]	7B	32,000
longchat-13b-16k [17]	13B	16,000
Mistral-7B-Instruct-v0.2 [18]	7B	32,000
Mixtral-8x7B-Instruct-v0.1 [19]	45B	32,000
zephyr-7b-beta-16k [20]	7B	16,000

Figure 2: A summary of models tested on the LongHealth dataset (Adams et al., 2024).

While the original authors demonstrated the capacity of benchmark LLMs with large context (16k in GPT-3.5 to 200k in Yi-6B) to adequately adapt to each of the challenges, achieving a maximum accuracy of 77%, 71% and 29% on each of the three tasks, respectively, there remains substantial room to evaluate the efficacy of existing clinical models on this dataset. The large models used in these experiments are impractically large, expensive and inference is slow. A retrieval model, specifically dense passage retrieval (DPR) (Karpukhin et al., 2020), would be a helpful tool to select relevant passage out of the documents and pass only those to the chat model, reducing the necessary context length, inference time and hopefully increasing accuracy.

2 Related work

2.1 Question Answering and LLMs

LLMs have obtained huge interest recently, owing to its demonstrated efficacy in addressing a multitude of natural language processing tasks, including Question-Answering (Bang et al., 2023). The transformer-based architecture, which forms the backbone of many LLMs, enables these models to understand the context of a question and generate a relevant answer (Vaswani et al., 2023). Generative Pretrained Transformer (GPT), one of the largest LLMs, has been used to answer complex questions across a wide range of topics (Brown et al., 2020). Similarly, BERT has been fine-tuned for QA tasks, demonstrating impressive results on the SQuAD dataset (Devlin et al., 2019). Furthermore, the RoBERTa model, a variant of BERT that uses dynamic masking and larger batch sizes, has been used in QA tasks and has achieved competitive results (Liu et al., 2019). In general, these studies all proves the broad applicability and effectiveness of LLMs in QA tasks.

2.2 Long Documents and LLMs

While language models prove accurate for QA retrieval problems in short documents and queries, problems arise as the corpus growths in length. Transformer models including BERT typically handle up to 512 tokens, or 380 words - approximately 2 paragraphs of text. Adequate for the most popular question-answer datasets like SQuAD (Rajpurkar et al., 2016), with average context length of 100 words, it is not sufficient for many applications, where the relevant information can be thousands of words long (Pereira et al., 2022). Prior work has attempted to bridge the gap to larger documents in a variety of ways. The first is increasing the accepted number of tokens to the model. Several 'long-document' models already exist for a variety of problems including LongT5 text-to-text transformer with scalable attention (Guo et al., 2021), Longformer the long document transformer with windowed and global attention (Beltagy et al., 2020) and BigBird, which is similar to LongFormer with random attention (Zaheer et al., 2020). Each of these three models accepts much longer inputs of up to 40k, 16k and 32k tokens, respectively. The

second method to solving the long-document QA problem is to use 'retriever' models trained to extract relevant portions of the document prior to performing QA inference using a 'reader'. This allows the same downstream QA model to be used, with more finely engineered inputs. Examples of this technique are Dense Passage Retrieval (Karpukhin et al., 2020) and Retrieval Augmented Generation (Lewis and Kiela, 2020) for open-domain question answering. While long-attention models may work well for questions consisting of a few pages, retriever models are the only solution that scales to cases with hundreds of thousands or millions of input tokens. Furthermore, retriever-reader models reduce unnecessary overhead in cases where relevant information is clustered to one or two short sentences in a long document.

Several solutions to the QA problem for long documents have been proposed, in some way incorporating the long document models described above. Most of the papers searched employed the retriever-reader model. D. Chen et al demonstrated the efficacy of their solution DrQA on the SQuAD dataset, which utilizes an embedding vector lookup for retrieval and a bi-directional attention flow model for 'reading' or, answering the question using the retrieved information (Chen et al., 2017). W. Yang et al combined Anserini, an open-source information toolkit for retrieval (Yang et al., 2017) and BERT to create their model BERTserini, which also proved very effective on the SQuAD benchmark (Yang et al., 2019).

2.3 LLMs with Clinical Applications

While AI has made inroads in healthcare, the application of language modeling to health is still limited (Singhal et al., 2022). The medical field is wide, extremely complicated and the dimension of data is huge. General-purpose embedding models like ClinicalBERT and BioBERT have shown promising results in various healthcare applications. ClinicalBERT, a language model fine-tuned on a large corpus of clinical notes, has been used for tasks such as predicting hospital readmission and understanding patient symptoms (Huang et al., 2020). Its ability to understand the unique language used in clinical settings makes it a valuable tool for healthcare professionals. BioBERT, a domain-specific language model pre-trained on large-scale biomedical corpora, has been successfully applied in several biomedical text mining tasks, including named entity recognition, relation extraction, and question answering (Lee et al., 2019). BioBERT can help in extracting meaningful insights from biomedical literature, aiding in tasks such as drug discovery and disease diagnosis.

Several datasets have derived from this overarching goal, including MedQA, a benchmark for general medical knowledge, with 61k questions containing patient symptoms of 100 words in length and supporting evidence from a medical resource, as well as the true patient diagnosis (Jin et al., 2020). Other large medical datasets exist as unlabelled corpuses available for understanding the context of medical records. The Medical Information Mart for Intensive Care III dataset, for example, contains electronic health records of 39k patients (Johnson et al., 2016)

LongHealth: A Question Answering Benchmark with Long Clinical Documents (Adams et al., 2024) is another significant contribution to the field. This benchmark provides a unique challenge due to the length and complexity of the clinical documents it contains. The dataset tests models' ability to understand and extract relevant information from lengthy, detailed medical texts.

3 Modeling

In our study, we focused on improving the performance of medical document querying. Our approach involved the use of Dense Passage Retrieval (DPR) and two pre-existing models, vicuna-7b-v1.5 and Mistral-7B-Instruct. Our hypothesis was that using a small number of passages extracted from the total document using DPR would perform just as well, if not better, than passing the whole document to the chatbot.

3.1 Baselines

3.1.1 Vicuna-7b-v1.5

The first baseline model we used is vicuna-7b-v1.5. This model represents the original method of passing the entire document to the chatbot for question answering. It is a large language model that has been trained on a diverse range of internet text. Vicuna-7b-v1.5 is an auto-regressive language model based on the transformer architecture. It was developed by LMSYS and fine-tuned from the Llama 2 model. The primary use of Vicuna is research on large language models and chatbots.

3.1.2 Mistral-7B-Instruct

The second baseline model is Mistral-7B-Instruct. Similar to vicuna-7b-v1.5, this model also takes the entire document as input for question answering. Mistral-7B-Instruct is a transformer-based model that has been pre-trained on a large corpus of text from the internet.

Both of the two models were imported and used in our study. They served as a fair comparison for our proposed method.

3.2 Proposed model

Our research study propose the use of Dense Passage Retrieval (DPR) model in long medical document question answering task, aimed to retrieve pertinent passages from extensive documents given a question. The model operates by transforming both the questions and answers into a high-dimensional space. The goal is to ensure that the embeddings of a question and its corresponding answer are close to each other, while the embeddings of a question and unrelated passages are distant.

During the training phase, the model employs a contrastive loss function. This function encourages the model to generate similar embeddings for a question and its corresponding answer, while generating dissimilar embeddings for a question and unrelated passages. This is achieved by iterating over the training data in batches, transforming the answers and questions into embeddings, performing in-batch negative sampling to create a similarity matrix, and calculating the contrastive loss on this matrix.

In the evaluation phase, the model is provided with a question and a set of documents. It transforms the question and each document into the high-dimensional space, and retrieves the most relevant passages for the question based on the cosine similarity between their embeddings. The performance of the model on each task is evaluated based on its recall and accuracy at different cutoffs.

During the inference phase, the model is provided with a new question and a set of documents. It transforms the question and each document into the high-dimensional space, and retrieves the most relevant passages for the question. The retrieved passages are then used to generate a response to the question.

The model's performance is further evaluated on a more challenging dataset consisting of long clinical documents and complex multiple-choice questions. This provides a comprehensive evaluation of the model's performance and its potential applications in the healthcare domain.

Dataset	Avg Question Length	Avg Passage Length	# of Questions
medication	7.7	9.5	5683
relation	8.4	9.7	8000

Table 1: Statistics from the emrQA dataset used in this project. Length is average length in words.

4 Dataset and evaluation

4.1 emrQA

A separate dataset is used to train on the extraction task prior to evaluating on the LongHealth dataset. The ermQA dataset (Pampari et al., 2018), an NLP-augmented dataset derived from the i2b2 relations dataset (Uzuner et al., 2011), was decided upon. The emrQA questions are derived from electronic health records, however are much shorter and simpler than those from LongHealth. The two longest of the five emrQA datasets, relation and medication were used, and one question corresponding to each 'answer' or retrieved passage was used from the dataset. See a summary of the final emrQA dataset used in Table 1, and a sample question from the relation dataset in Figure 3. Note that the average question and answer lengths are significantly shorter than those from LongHealth, less than 10 words each on average.

Figure 3: Example of a question from emQA medication dataset

```
Question: Has this patient ever been treated
    with levothyroxine sodium

Passage: POTENTIALLY SERIOUS INTERACTION:
    DIGOXIN & LEVOTHYROXINE SODIUM
        Reason for override: will monitor
```

4.2 LongHealth

The LongHealth dataset is slightly different in composition from emrQA. 20 patients with 20 questions each are provided, with documents ranging from 5000-6000 words. While emrQA is a strict retrieval task, LongHealth is posed as a multiple choice question-answering task, so five choices are provided for each question. The different formats were coalesced during training of DPR. See Figure 2 for a summary of the average question

Dataset	Avg Question Length	Avg Passage Length	# Questions
LongHealth	14.8	16.9	400

Table 2: Statistics from the Longhealth dataset used in this project, averaged over all 20 patients. Length is average length in words.

and passage length, as well as the total number of questions. Each patient has 20 questions, and has a unique medical condition. For an example of an example question from the LongHealth dataset, please see Figure 4.

Figure 4: Example of a question from LongHealth dataset

```
Question: What abnormality was noted in Mr.
    Nilsson's ECG on 11/20/2021?

Answers:
a: Bradycardia,
b: ST-elevation myocardial infarction
c: Atrial fibrillation
d: Ventricular tachycardia
e: Intraventricular conduction disorder <--
    correct

Passage: 2-12.92 kU/L
  GLDH                          3.7 U/L
      \<6.4 U/L
  Gamma-GT                      89 U/L
      8-61 U/L
  LDH                           184 U/L
      135-250 U/L
  Parathyroid Hormone           55.0 pg/mL
      15.0-65.0 pg/mL
  25-OH-Vitamin D3              10.9 ng/mL
      50.0-150.0 ng/mL
```

4.3 Metrics

4.3.1 Cosine Similarity

Dense passage retrieval searches for passages in the document which are minimally distant from the query string, which in this case is the question. See below the equation for cosine similarity, where $\mathbf{A}$ and $\mathbf{B}$ are vectors with length of the embedding length.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

4.3.2 Contrastive Loss

The training of DPR seeks to minimize the cosine distance between the query and passage, while maximizing the distance from multiple other false examples. See below this criterion, multi-class N-pair loss, where $f(\mathbf{x}^+)$ denotes a positive example embedding and $f(\mathbf{x}^-)$ denotes a negative example embedding (Sohn, 2016).

$$\mathcal{L}_{\text{N-pair}} = -\log \frac{\exp(f(\mathbf{x})^\top f(\mathbf{x}^+))}{\exp(f(\mathbf{x})^\top f(\mathbf{x}^+)) + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^\top f(\mathbf{x}_i^-))}$$

5 Experiments

Experiments are all carried out in Python. Pytorch is used for all modeling (Paszke et al., 2017). All pre-trained models are accessed using HuggingFace model repository (Wolf et al., 2020). Code and data may be accessed through the DrLongHealth repository [1]. One NVIDIA A40 with 40 GB VRAM and 9 vCPU is used for training of the DPR and two A40 with 80 GB VRAM are used for the question-answering tasks. Training and inference are performed using RunPod virtual machines (Runpod, 2024).

5.1 Training

Training for DPR began with the pre-trained biomedical embedding model BioBERT, accessed with HuggingFace (Lee et al., 2019). The model is trained over 128 epochs with the Adam optimizer, with learning rate of $1e-4$. In-batch negative sampling is used with contrastive loss. A training/-validation split of 0.8/0.2 is used. The model with highest validation accuracy is stored and used for inference [2].

5.2 Question-Answering

All three separate multiple-choice question-answering tasks from LongHealth are performed. Each patient's documents are broken into sentences smaller than 128 words using the Python Natural Toolkit (Bird et al., 2009), which form the basis of the corpus the DPR model searches over. The question is concatenated with the multiple choice answers, and forms the query string. The question and answer are subsequently embedded using the trained DPR models. Question-answering inference is then performed using the top 10 and 25 passages, as well as the entire document from the patient (this is the basis of comparison). This inference is performed on one of two models from the original paper: Mistral-7B-Instruct and Vicuna-7b-v1.5, which were selected as the candidate high-performing and low-performing models in the origi-

[1] https://github.com/rballachay/DrLongHealth

[2] https://github.com/rballachay/DrLongHealth/blob/main/train_dpr.py

nal experiment. See Figure 5 below for how the top-n passages reduces the length of the query passed to the LLM model in task 2.

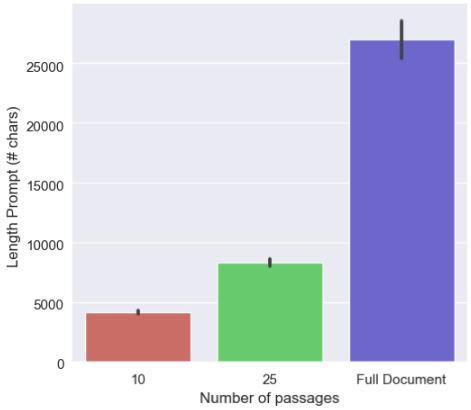

Figure 5: Length of documents passed to LLM by number of passages retrieved. Average over all 400 questions is shown, bar represents standard deviation.

5.2.1 Task 1: Simple Information Retrieval

The first task is to respond to affirmative questions where the information is located in one or more locations in the text. Only the documents relevant to the patient are used.

5.2.2 Task 2: Complex Information Retrieval

The second task is to respond to similar questions as the first task, but where documents from other patients who may have similar medications or symptoms, are included in the documents.

5.2.3 Task 3: Information Negation

The third task is to correctly identify when a question may not be answered given the documents. The same questions are posed and documents from other patients are supplied instead.

6 Results

6.1 DPR Training

The DPR model was trained over 128 epochs, however the lowest validation accuracy was observed after 15 epochs, so that model was retained. See in Figure 6 the traing loss curve over the first 15 epochs.

The out-of-the-box BioBERT model performed very well at the retrieval task, and improvements on the LongHealth dataset were minor over training on the emrQA dataset. Over training, the top-10 retrieval accuracy on the LongHealth dataset increased from 38% to near 44%, while the top-100

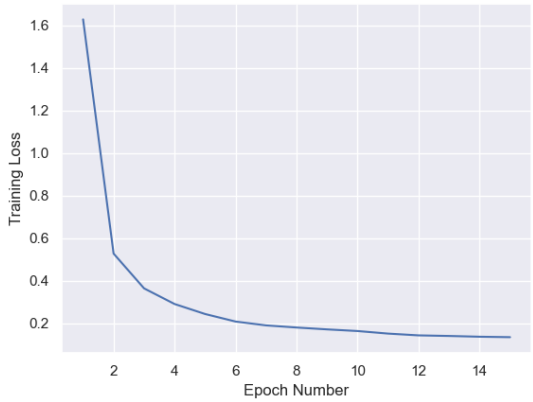

Figure 6: Average training loss by epoch during training of the DPR model on the emrQA dataset.

increased from 75% to 83%. See Figure 7 for full accuracy results of the retriever on LongHealth.

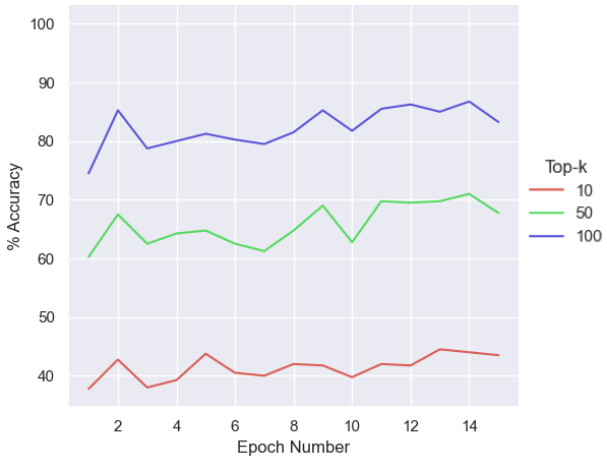

Figure 7: Top-k accuracy of DPR retrieval task on the LongHealth dataset over each epoch of training on the emrQA dataset.

6.2 Question-Answering

400 questions are passed to the two QA LLM models with 10/25 passages or the full document, and the multiple choice answer is parsed from the response. Only a single replicate was performed due to computational restraints, and the low variation observed over replicates in the original paper. See the full accuracy results for all three tasks in Figure 8. Accuracy of the vicuna-7b-v1.5 model is significantly improved by extraction of the top 10 and 25 passages, from a baseline of 31% in Task 1 to 42% and 44% with 10 and 25 passages, respectively. This is mirrored in Task 2, where vicuna-7b-v1.5 accuracy improves from 28% to 38% for top 10

and 25 passages. A slight decrease in accuracy is observed with Mistral-7B-instruct, decreasing from 62% to just below 59%.

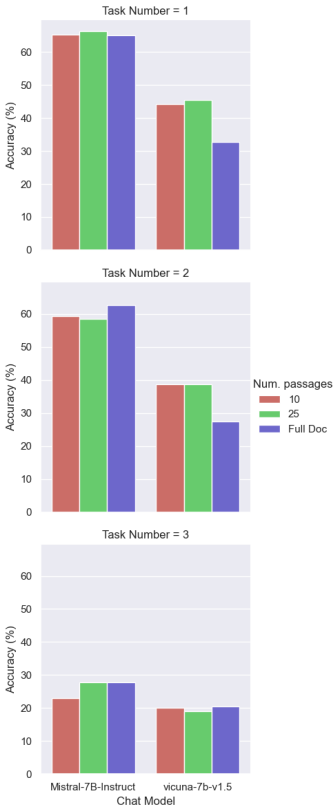

Figure 8: Multiple-choice QA accuracy over 400 questions for prompts with 10 and 25 DPR-retrieved passages as well as the full document. Each row corresponds to one of the tasks described in the experiment section.

The results support the hypothesis that retrieving a subset of the relevant information for the question-answering model is more beneficial than passing the whole document. Computational intensity, both in time and memory, is known to scale with the length of the prompt input to the model. Recent work by (Blankemeier et al., 2023) demonstrated a linear correlation between context length and inference time in a question-answering task. This is supported by the observed speedup when performing inference using a subset of the passages, as the average speedup-per-question was 1.8x for mistral-7B and 3.2x for vicuna-7b, including DPR inference time. See Figure 9 for the full inference time results.

Generally, accuracy with retrieved passages remained close to the original document, with a small increase observed in the vicuna-7b model. As touched upon in the introduction, the information

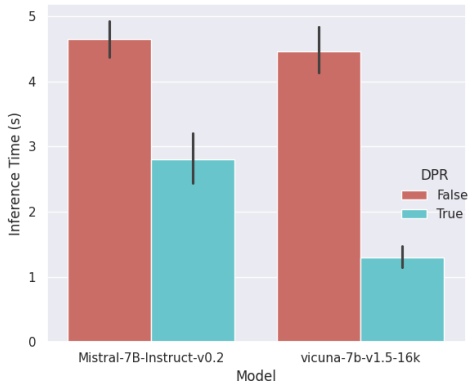

Figure 9: Average Inference time of LLM models with and without DPR extraction. 'DPR' condition refers to top-10 passages selected. Inference time includes DPR inference in the case that it is included.

relevant to each question is concentrated in one or few short segments of the full document, and the most important task of the chat model is to find and extract this information. The high accuracy of the top-n passages indicates that the trained DPR model is accurately identifying and retrieving information from these passages. To demonstrate this concentration of passages, the top-5 passages from each of the 20 questions are plotted for each patient, as they were in the original paper, and the results can be seen in Figure 10.

The concentration of these passages throughout the document contains the original passages for each question, but appears to include passages which were not represented in the original version of this figure. The reason for this is that the authors of LongHealth only included one of the references necessary to answer a question, however many of the documents repeat information, sometimes with more explicit terms. A good example of this is the Breast Cancer Patient, who had a question relating to a BCRA testing. While the authors used the first reference of this as the 'correct' retrieval passage, there are multiple other references in the document that were not indicated. See Figure 11 for an example of how a question may return multiple high-similarity passages that do not answer the question directly.

In reviewing the results for Task 3, a flaw in the design of Task 3 became apparent, which the original authors may have overlooked. It seems that the model was able to determine that there was not enough information available to answer the question, however it selected a different multiple

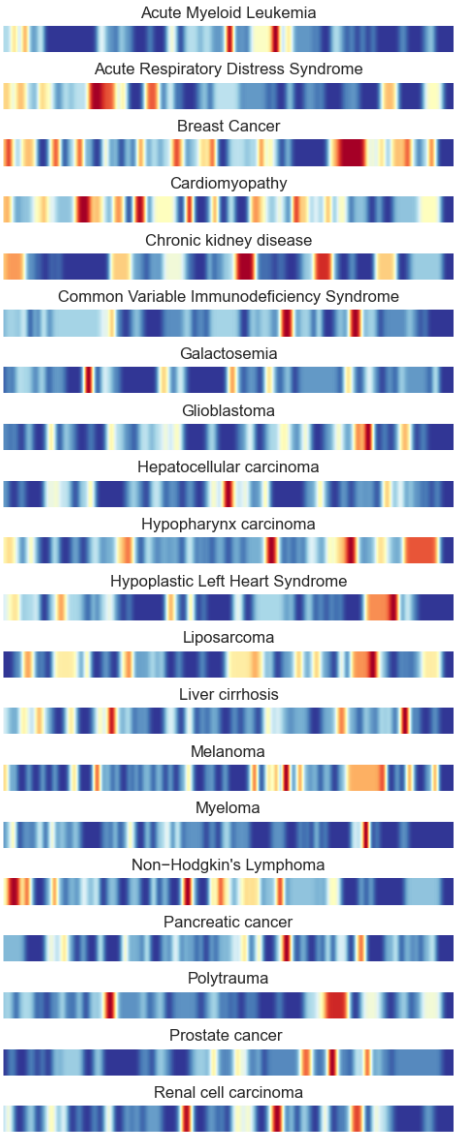

Figure 10: Top-5 passages retrieved from each patients documents plotted as a density map. Blue indicates low retrieval frequency and darker red indicates high.

choice answer than F: Question cannot be answered with provided documents. Patient 2, Question 1 is a good example of this phenomenon, see this in Figure 12.

One likely explanation for this is that in the prompt for the chat model, which wasn't modified from the original in Task 3, where there was an additional multiple choice option present. See the portion of the prompt in question in Figure 13. To improve the accuracy of this task in the future, it is recommended that the prompt is slightly modified in the future for Task 3 to include the option F and a line about choosing that option if it isn't confident.

Figure 11: Example of retrieved passages from Breast Cancer Patient

```
Question: Based on the tumor board decision from
     07/12/2019, what is the recommended course
    of action if Mrs. Linda Mayer's BRCA testing
     returns negative?

'Correct' Passage: Tumor board decision from
    07/13/2019 ... If BRCA testing returns
    negative: Proceed with a selective excision

'Other' Passage: **Recommendation**: If BRCA
    negative, SE left mamma after ultrasound-FNM

'Other' Passage: BRCA testing confirmed a BRCA2
    mutation,\nwarranting bilateral subcutaneous
     mastectomy
```

Figure 12: Example of chat model incorrectly choosing MC response, while understanding

```
Question: What would be the cumulative dose of
    Pembrolizumab in Mrs. Jane Done adjuvant
    chemotherapy?

Correct: F, Question cannot be answered with
    provided documents

Answer: The correct answer is B: 1.8g.
(Note: The document does not provide information
     on the total cumulative dose of
    Pembrolizumab given to Mrs. Jane Done during
     adjuvant chemotherapy.)
```

7 Conclusion

The creation of the LongHealth dataset represents a large step forward for clinical usability of LLMs. Several long-context models generated promising results in preliminary analysis on this dataset. This work has shown that the preliminary results may be significantly expanded upon by addition of a retrieval step prior to LLM inference. Accuracy and inference time were markedly improved by the addition of a DPR retrieval step prior to inference. In addition, the shortening of the question context length opens up the possibility of using other more specialized medical QA models with shorter context length.

8 Contributions

Riley implemented the models and ran the experiments. Haowei found data sources, generated data sets and some results. Both authors contributed equally to writing the final manuscript.

Figure 13: Part of Model Prompt to change for task 3

```
Please answer using the following format:
1. Begin your answer with the phrase "The
    correct answer is".
2. State the letter of the correct option (e.g.,
    A, B, C, D, E).
3. Follow the letter with a colon and the exact
    text of the option you chose.
4. Make sure your answer is a single, concise
    sentence.
```

References

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2024. Longhealth: A question answering benchmark with long clinical documents.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Louis Blankemeier, Jason Fries, Robert Tinn, Joseph Preston, Nigam Shah, and Akshay Chaudhari. 2023. Efficient diagnosis assignment using unstructured clinical notes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 485–494.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *CoRR*, abs/2112.07916.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Sci Data 3, 160035*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. 2022. Visconde: Multi-document qa with gpt-3 and neural reranking.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction.

Runpod. 2024. Runpod Documentation. `https://docs.runpod.io/`. [Online; accessed 10 April 2024].

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.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, and Gabriel Pila. 2024. Medlm: Exploring language models for medical question answering systems.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.