Building and evaluating end-to-end Medical OpenQA Systems with ColBERTv2

Laura Uzcategui
laura@uzcategui.dev

Young Don Ko
youngdon.ko@gmail.com

Abstract

Open domain question answering (OpenQA) for the medical domain has been gaining more research attention, mainly due to the advancements of bigger and more powerful Language Models which generate massive improvements in terms of downstream tasks in NLP. As part of this work, we introduce evaluation of different medical domain datasets for Question Answering (QA) by building two settings for end-to-end OpenQA leveraging ColBERTv2 as information retrieval (IR) mechanism for passage search. First we evaluated MedQuAD dataset, initially developed for solving QA with Question Entailment Recognition, our system aims to resolve the task by working with Few-Shot OpenQA with Autoregressive Language Models. Second we evaluated MedQA dataset, a multi-choice OpenQA dataset for solving medical board exams. In this case we developed an end-to-end OpenQA system with multiple Language Models and further fine-tuning on the training data. During experimentation and evaluation on the test data, we have found that working with ColBERTv2 as our IR mechanism we obtained 3% improvement over different baselines, encouraging the further development of more OpenQA systems that incorporate this component to be able to help with passage retrieval and potentially contribute towards performance improvements.

1 Introduction

Currently, many of the Natural Language Understanding (NLU) and Natural Language Processing (NLP) advancements coming up everyday involve evolving the system components with respect to solving downstream tasks like Standard QA and its variations like OpenQA, Few-shot OpenQA, among others. In particular, we have observed multiple approaches developed for OpenQA, as an example reading comprehension models have evolved with systems like Dense Passage Retrieval by (Karpukhin et al., 2020) and Contextualized Late interaction retrieval by (Khattab and Zaharia, 2020) aiming to retrieve relevant passages of text from a big collection of documents that can be passed to Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019). Thus, having the common goal of helping improving performance and generalization of downstream tasks like QA where most of the time domain-specific areas like education, research, and specifically healthcare could make it even more challenging in terms of domain adaptation.

Healthcare applications has started to benefit and obtain gains from this advancements. For example, usually physicians, nurses among others perform lookups for medical records, similar cases or information that could potentially serve as a basis for a new treatment or resolve a new case seen in another patient. However, one has to acknowledge that working on this domain has its limitations such as, availability of data, lack of resources, domain-knowledge and privacy/ethic regulations, our belief is that providing better and accurate systems might help reduce such limitations.

The motivation for this project is derived from the limitations mentioned above, specifically how we can contribute by improving the development of Open-domain QA (OpenQA) systems, where a question is posed to the system, the IR system will retrieve a passage from a large collection of documents and pass it to a reader model with the hope to find the right answer.

Our core hypothesis is centered around the premise that while basic IR mechanisms such as Term Frequency - Inverse Document Frequency (TF-IDF) and Best Match 25 (BM25) (Robertson and Zaragoza, 2009) might work well, we propose that by making use and leverage the power of efficient late interaction retrieval mechanisms like the one implemented in ColBERTv2 (Santhanam et al., 2022) will potentially obtain gains in terms of enriching search, getting better results and obtain an answer improving not only the performance of state-of-the-art systems that use IR techniques but also giving a boost to Question Answering task particularly in the medical domain, which is a difficult domain to tackle.

As part of the work done, our contributions are: we have built an end-to-end OpenQA system making ColBERTv2 the core Information retrieval mechanism using late interaction with distillation with the hope to contribute towards having better passages to be used with different transformer Language Models (LM) fine-tuned over the training data. We believe these combination of the retriever and reader models are well suited for the task of obtaining answers to questions in the medical domain. Additionally, we have chosen 2 medical datasets to perform evaluations on the system and were able to obtain 3% gain in accuracy over the baselines reported using the same IR mechanisms those datasets have used.

2 Related Work

Multiple approaches that pertain to solve QA task has been studied and implemented, over the next sections we describe the 3 main areas of focus for our research: First, finding medical domain datasets that would help us drive data-centric development and evaluation of OpenQA systems. Second, the study of Information Retrieval systems specifically ColBERTv2 as the core component of the system to be developed. Third, we focused on searching reader models that better adjust the task, such as Language Models and Autoregressive models and how to apply fine-tuning to build baselines models that would help to perform evaluation on the medical domain and improve the way we retrieve and/or generate better answers.

2.1 Medical domain datasets

The list for bio-medical datasets collected and categorized by (Jin et al., 2022) is wide and well formed. After studying multiple of those datasets with a focus on approach and format, we have selected two of those as our main candidates for development and evaluation of our system, the selection criteria was contingent to the usage of information retrieval.

Starting from Question Entailment Recognition approach from (Ben Abacha and Demner-Fushman, 2019) where a new QA methodology was proposed to improve medical QA systems by working with Terrier[1] as their retriever for questions-answers pairs and a logistic regression classifier to choose the questions that could entail an answer from the Original question posed to the system.

In terms of Open-domain datasets, (Jin et al., 2021) presented MedQA a new large scale dataset where the format of the answers is represented as multiple-choice, and evaluated with an end-to-end OpenQA system using Elasticsearch[2] as their Retriever and on the Reader side they used transformer Models fine-tuned on the data that was retrieved and the questions-answer pairs. A most basic approach was taken by (Alzubi et al., 2021) as they built their own retriever using TF-IDF mechanism with Cosine Similarity, and transformer models with a ranker to select the answer among possible candidates.

2.2 Information Retrieval Models

Information Retrieval has been a crucial task in pretty much many of the actions we perform in our daily lives, for example when we do search in Google, we are implicitly using BERT or searching for a song in Spotify or a product in Amazon, most likely we are also using an IR technique. Weighting and scoring techniques such as TF-IDF, BM25 among others have proven to be good for IR basic tasks, although those suffer from limitations such as not being context-aware or using only co-occurrence patterns among others.

Interesting progress has been shown around using Neural Retrieval, Dense Passage Retrieval as efficient and effective mechanisms to obtain good results regarding search, yet usually those systems encode queries and documents into single-vector representations, as as an alternative (Santhanam et al., 2022) proposed ColBERTv2 as an improvement over the original version of ColBERT where (Khattab and Zaharia, 2020) demonstrated how using late interaction mechanisms with multi-vector representations could be very efficient and provide a retriever that could generalize well to out-of-domain data and minimize memory footprint.

Results has shown that ColBERTv2 evaluated against other mechanisms that use distillation or any other pre-training method, achieves superior quality by gaining around 1.5 points approximately over the best systems such as SPLADEv2 (Formal et al., 2021) and RocketQAv2 (Ren et al., 2021).

[1] http://terrier.org/

[2] https://www.elastic.co/

2.3 Reader Models

Among the different approaches to answer extraction for a question, we found that Eleuther (Black et al., 2021) and GPT-3 (Brown et al., 2020) as Autoregressive models used in the Few-Shot OpenQA setting could achieve results that improve over 3% over Standard QA systems. We have followed this path as an alternative to (Ben Abacha and Demner-Fushman, 2019) logistic regression classifier used over MedQuAD in order to perform evaluation. In terms of OpenQA setting, studying models such as BERT by (Devlin et al., 2019) and its variations such as BioBERT introduced by (Lee et al., 2020) as a pre-trained language model for biomedical domain with the hypothesis that could reach great results in this domain-specific area by re-using BERT weights as their base model. In contrast to BERT based models, we tried using BioRoBERTa-base by (Gururangan et al., 2020) based on RoBERTa (Liu et al., 2019) as it is trained using Byte-Pair Encoding (BPE) and specifically pre-trained on 2.68M of scientific papers from Semantic Scholar.

Yet another interesting approach recently published and that got state-of-the-art results with bioNLP tasks such as Question-Answering is LinkBERT presented by (Yasunaga et al., 2022) a pretraining method that leverage document link knowledge to be able to embed knowledge into LMs and perform better predictions and bridge the gap between documents and using vanilla objectives in next sentence prediction used in BERT.

3 Data

To evaluate our end-to-end Medical OpenQA systems, we selected 2 publicly available datasets to built upon it. The criteria for selecting those was based in how well those aligned to our our hypotheses given that were developed with the aim to evaluate systems where information retrieval was being incorporated and to be able to find/retrieve the answers to a given question.

3.1 MedQuAD

Starting by MedQuAD, developed with the aim to evaluate an Information Retrieval with Question Entailment Recognition system, where (Ben Abacha and Demner-Fushman, 2019) proposed it as a way to improve QA systems with the hypothesis that given a User Question (PQ) they could select Hypothesis Questions (HQ) using Natural Language Inference that could help to get an answer to the consumer health questions.

MedQuAD is a set of 47k Question-Answer pairs collected from12 U.S. National Institute of Health (NIH)[3] websites. To build this dataset, they hand-crafted patterns to be able to generate the question-answer pairs, additionally, they have added metadata such as the focus of the set of questions. A taxonomy of Question types can be observed at Table 1, defined after manual evaluation of 1.7k questions.

Development of this dataset and the system itself was evaluated against TREC LiveQA 2017[4] task obtaining an improvement of almost over 29.8% over the best system.

Question Type	Qty	Included Information
Diseases	16	Research, causes, treatment, prognosis,symptoms, inheritance, genetic changes, etc.
Drugs	20	Medicine interaction,food interaction,effectiveness contraindications.
Information	1	Procedures, exams, and treatment.

Table 1: MedQuAD description by question type and quantity.

3.2 MedQA

MedQA[5] is a dataset built with the aim to contribute to having an Open Domain Question Answering Dataset from Medical Exams Boards within 3 different Regions: US, Taiwan and Mainland China and motivate the research community to build more OpenQA systems in the medical domain. The distribution of question-answer pairs collected for the English subset is described in Table 2 where they applied removal of duplicates and split the data based on questions with 80% for training, 10% development and 10% test set.

Subset	Training	Development	Test
USMLE	10178	1272	1273

Table 2: Split of Question-Answer pairs for English MedQA dataset.

In order to evaluate the dataset and see how hard would be to in fact respond to the questions pro-

[3] https://www.nih.gov/

[4] https://trec.nist.gov/data/qa/2017_LiveQA.html

[5] https://github.com/jind11/MedQA

vided over medical exams, the authors have built an end-to-end OpenQA system, using Elasticsearch as Retriever with an index over a collection of 18 English books used for US and Mainland China and 33 simplified Chinese books, those books are officially used by students when they want to present medical examination to get their license.

For the purpose of our project we have worked only with the English subset of the data (USLME), our reasoning over this dataset is that as it is designed to work as OpenQA then will fit our goal of evaluating ColBERTv2 as an efficient retriever to build those kind of systems. The format of this dataset is multiple-choice and it is described with an example on Apendix A.

4 Models

Our end-to-end system is divided in two sections based on the different approaches used to work with the different datasets. First we chose to work with Few-Shot OpenQA setting for MedQuAD and OpenQA setting for MedQA dataset, from there we started studying and selecting models that align best with the setting.

4.1 Few-Shot OpenQA Setting

As part of our first evaluation we have built and end-to-end system with MedQuAD. Our design is composed by a regular Few-Shot setting with the components that can be observed over Figure 1.

On the retriever side, we used a pre-trained ColBERTv2 checkpoint trained on MS MARCO[6] passage ranking dataset, as it has been suggested by (Santhanam et al., 2022) that ColBERTv2 performs well with out-of-domain data we did not think it was necessary to do further training on the retriever. The parameters of the model are based on HuggingFace BERT transformer model with 12 attention heads and 12 hidden layers, with a vocab size of 30522. The tokenizer used was BertTokenizer with max model token length of 512 token per sequence.

In terms of indexing, 2 indexes has been created to build baselines to compare against our system. One index for Questions with the training split of MedQuAD, to be able to search questions that will be similar to the Original Question posed to the system, the second index was built with a collection of medical books produced during the development of MedQA, as we presumed it will contain enough knowledge to help us build a good retriever.

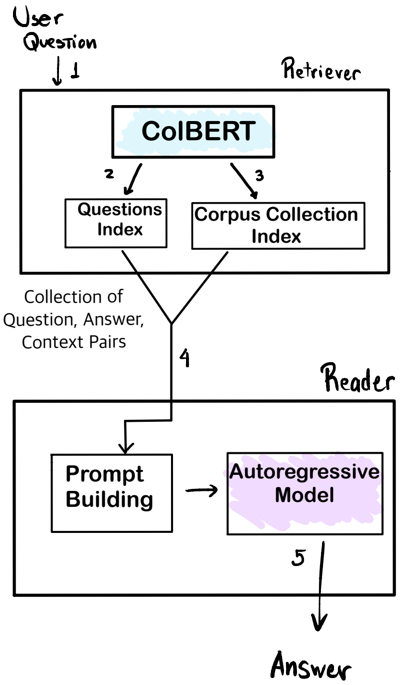

Figure 1: Few Shot OpenQA Setting

On the reader side, we have chosen Eleuther as the Autoregressive Language model that would obtain the answer. The parameters we have utilized are the pre-trained model with 1.3B parameters (*gpt-neo-1.3B*). The model was trained over 380B tokens over 362k steps. The tokenizer settings were padding from the left and padding/truncation to the longest sequence, the maximum length per sequence was set to 2k tokens.

Putting all the pieces together, an original question (OQ) (1) was posed to the system, it will perform a search of the most similar questions (SQ) (2) from the retriever and a relevant pasagge (3), and pass it to a processor that will build the prompt (4) with the similar question-answer pairs plus the original question and pass it to the Model (5) to generate the answer, once the answer was retrieved it was passed to an evaluator function that will compute the Macro-F1 score.

4.2 OpenQA setting

For this setting, we have built an end-to-end OpenQA system that will allow us to select the right choice out of the possible choices available per question in MedQA dataset. The diagram with the main components of the system are described in the Figure 2.

[6] https://microsoft.github.io/msmarco/

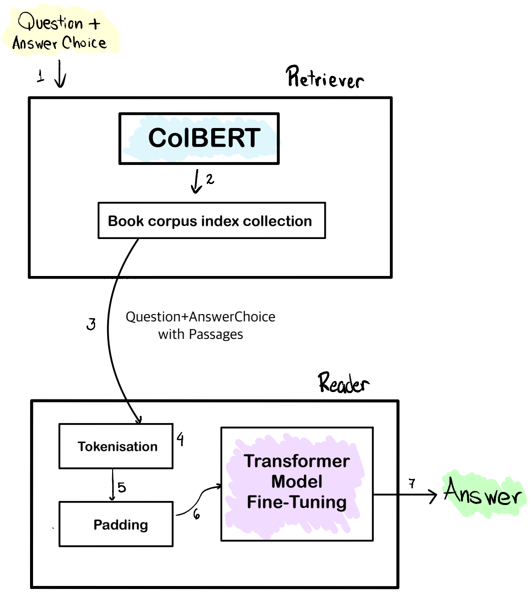

Figure 2: OpenQA Setting

The retriever configuration was exactly the same as described over section 2.2, with the difference that only one index was created with the collection of 18 english medical books distributed publicly by the authors of MedQA.

In the case of the reader, we have utilized multiple variations of HuggingFace pre-trained Language Models models. Those models were chosen accordingly to be aligned with the experiments performed during the development of MedQA and will serve as the baselines for our end-to-end system. The list of models is over Table 3, each of this models was fine-tuned further on the training split of MedQA dataset.

A full end-to-end flow as shown in Figure 2 is as follows, first a concatenation of the Question and Answer Choice is posed to the system (1), then a search is performed over the index (2), a HuggingFace dataset is created with all the question-answerChoice and its correspondent set of passages retrieved (3). The dataset is the pre-proccesed by a Tokenizer (4) and Padding function (5) and passed to perform fine-tuning over a Reader Model (6) that will classify and obtain the answer choice with the highest probability (7), this model is evaluated using Accuracy metric.

Model	weights
BERT	bert-base-cased bert-base-uncased
DistilBERT	distilbert-base-cased distilbert-base-uncased
BioMed-RoBERTa	*From: allenai* biomed_roberta_base
BioBERT	*From: dmis-lab* biobert-base-cased-v1.1-squad biobert-large-cased-v1.1-squad

Table 3: OpenQA Reader Models.

5 Experiments

As part of our experimentation, based on the dataset that was being evaluated a different end-to-end system was implemented. The work was divided between data processing, building baselines for the retriever and for the reader model, and perform evaluation based on the metric chosen for each setting.

5.1 MedQuAD Few-Shot OpenQA Experiments & Results

5.1.1 Data Processing

For this phase of the project, MedQuAD data collection was done by coding a data processor in Python that takes the original XML files in the dataset github repository[7] and processed to obtain question-answer pairs in SQuAD format. We were able to obtain only 15k answer-pairs due to copyright constraints. Further processing was applied to split the data in subsets into 68% for training, 17% and 15% for development and testing.

5.1.2 Indexing

Our baselines were built using the same retrievers as in MedQuAD using Terrier to perform information retrieval. For passage retrieval we have the baselines expressed over Table 4.

ID	Baseline Name
#1	TF-IDF with Terrier.
#2	TF-IDF and Inp_expB2 combined
Ours	ColBERTv2 with Contextualized Late interaction.

Table 4: MedQuAD Few-Shot OpenQA Avg Macro-F1 Results.

Additionally, we have built an index based on the training split with the questions, in order to retrieve the similar questions to the original question initially posed to the system.

[7] https://github.com/abachaa/MedQuAD

5.1.3 Reader

After retrieving the passages and question-answer pairs for each of the baselines and our retriever, we code a prompt builder in Python to construct and batch the prompts that will be passed to the Eleuther model. In order to evaluate the system, we have used Average Macro-F1 scores over the predicted answers. Results can be observed in the Table. 5 for each of the baselines along with our system.

Scoring	Retriever	Avg. Macro-F1
TF-IDF	Terrier	0.03
TF-IDF+BM25	Terrier	0.015
Ours: MaxSim	ColBERTv2	**0.07**

Table 5: MedQuAD Few-Shot OpenQA Avg Macro-F1 Results.

Those results were not as we would have expected, presumably due to the selected collection for passage retrieval and a misalignment with the questions as it was towards consumer health questions and the collection was based on medical books used for examination of students. Still when comparing results we observed improvement of 3% and 5.5% when using ColBERTv2 over Terrier baselines respectively.

5.2 MedQA OpenQA Experiments & Results

5.2.1 Data Processing

In terms of data processing, we have 2 sources of data to work with, the 18 english books and the set of questions for each data split, both of those are publicly available at MedQA github repository[8]. For the books collection, we created a Python data processor as a modified version of the script provided by the authors of MedQA, to be able to create batches of sentences that will be fed to the retriever of index creation.

As for the dataset itself, we have built a Python data processor to consume the question-answer pairs along with the choices for each subset split and save it into NamedTuple data structures that would allow us to work with easily. After the data from the retriever was collected, we have created HuggingFace datasets[9] for each split in order to pass it to the Reader Models for further fine-tuning and inference for evaluation on the test set.

5.2.2 Indexing

From the point of view of indexing, part of our baselines were chosen to be able to compare MedQA retriever and our proposed retriever ColBERTv2, additionally we have built different indexes named by round in which the main difference was the preprocessing technique applied to the the data before indexing. After testing different rounds and due to time constraint limitation, we only utilized round 3 as it deemed the best results in terms of passage size without being truncated at indexing time.

Retriever (Scoring)	Round	Pre-Processing
Elasticsearch (BM25)	-	Snowball stemming
ColBERTv2 (MaxSim)	1	Only alphanums
ColBERTv2 (MaxSim)	2	Only alphanums. Delete duplicates.
ColBERTv2 (MaxSim)	3	Same as round 2. Smaller doc size

Table 6: Retriever Baselines by round and preprocessing.

The best ColBERTv2 setting to create the index and worked out best based on the average document size was to set $nbits = 2$ (encode each dimension with 2 bits) , $doc\_max\_len = 300$ (maximum size of the document), $query\_maxlen = 300$ (maximum size by query), everything else was left by default. After ColBERTv2 Index creation, we wanted to obtain good quality search results, after trying multiple differente alternative for creating the Searcher we followed ColBERTv2 authors recommendation in their repository [10] and set parameters to $ncells = 4$, $centroid\_score\_threshold = 0.4$ and $ndocs = 4096$.

5.2.3 Readers

After performing retrieval for each of the baselines, we experimented with different Reader models from HuggingFace specifically using the AutoModelForMultipleChoice and AutoTokenizer from Auto Classes module[11] with the respective model weights chosen for each of our experiments.

Before fine-tuning, some data processing has to be done for each split of the MedQA dataset, that will allow us to set the format in which the reader will be expecting the data for tokenization

[8] https://github.com/jind11/MedQA

[9] https://huggingface.co/docs/datasets

[10] https://github.com/stanford-futuredata/ColBERT/

[11] HuggingFace Autoclass https://huggingface.co/docs/transformers/model_doc/auto

and padding. For each question, we will take each context retrieved for a particular choice of data and form a batch of contexts that will go after the CLS classifier token, then a second batch of sentences was formed concatenating the Question and a Choice of answer, in this way the softmax classifier set at the top of the CLS output will be able to predict the probabilities for each choice and we take the maximum probability.

The chosen metric to do evaluation during finetuning and inference is Accuracy, as we wanted to compared how well each model does during classification to select the right choice of answer. Furthermore we wanted to be aligned to do perform the same evaluation as MedQA system so that we could have a point of comparison with the scores obtained during (Jin et al., 2021). In terms of finetuning, we performed two variations based on the number of epochs (3 & 8), and found that the best setting was to set the number of epochs to 8 due to model size and the size of the data used for the fine-tune process.

Following evaluation on the test set, we grouped the results based on the reader model we utilized and compared with the retriever used in MedQA and ours (ColBERTv2), those results can be observed in Table 7

	Retriever			
	Elastic		ColBERTv2	
Model	Dev	Test	Dev	Test
BERT$_{uncased}$	0.276	0.258	**0.312**	**0.302**
BERT$_{cased}$	0.287	0.284	**0.294**	**0.290**
DistilBERT$_{uncased}$	**0.282**	**0.284**	0.277	0.258
DistilBERT$_{cased}$	0.271	0.241	**0.285**	**0.271**
BioRoBERTa	0.292	0.295	**0.329**	**0.318**
BioBERT$_{base}$	0.292	0.262	**0.333**	**0.315**
BioBERT$_{large}$	**0.279**	0.256	0.259	**0.277**

Table 7: MedQA Results using Different Readers and ColBERTv2.

6 Analysis

Looking back to the results shown in Table 7 , it has been observed that accuracy has improved in almost all the models for above 3% over the baselines using Elasticsearch with BM25, as the retriever presented during MedQA for development and testing evaluations. Comparing the Reader models such as BERT, DistilBERT and its variations (cased vs. uncased), it is observed that vanilla BERT fine-tuned on MedQA data performs better than DistilBERT fine-tuned over the same data.

Overall we observed that using ColBERTv2 get much better results for any of the reader models than using the Elasticsearch baselines except for DistilBERT uncased.

Doing some qualitative analysis with the retrievers we could observed that Elasticsearch performs retrieval using exact match strategy meanwhile ColBERT presumably returns more contextualized passages that could be helping our Readers to perform better. For example when asked question like “What is the nervous system”, we observed the retrieved passages by Elasticsearch were really short and towards term matching, meanwhile the passages retrieved from ColBERTv2 where longer and make more sense in terms of relevant passages to the question, this examples can be observed in more detail over Appendix B.

In terms of error analysis, we took some incorrect predictions from one of our experiments, specifically with BERT baseline, and performed analysis of some passages retrieved by ColBERT along with the question and the ground truth vs. predicted choice. To make it more visible, let us look at one example over Table 8 where it seems the model learnt to receive signals from the question+choice of answer being passed for prediction along with the context. The correct label option: A is highlighted in Bold, meanwhile the predicted option is C, this suggests that the model might be doing some sort of memorization and pattern matching instead of doing reasoning over the context and question being passed.

6.1 Ablation Studies

Taking some examples where the answer was different than the ground truth we could observe that there is not enough information in the passages that could contain the answer, and we decided to take an step further and do some ablation studies with removal of ColBERT passages and pass only the question-answer pairs to the reader only in order to evaluate the performance.

The difference of this small ablation study is shown over Table 9 where across the second column named *“with ColBERTv2”* one can observed the accuracy is around 6% to 7% on the test set for the models evaluated BERT (base-uncased) and BioMed-RoBERTa (base), giving us further indications that our hypothesis is headed to the right direction.

Question	A 23-year-old woman comes to the physician because she is embarrassed about the appearance of her nails . She has no history of serious illness and takes no medications. She appears well. A photograph of the nails is shown. Which of the following additional findings is most likely in this patient?
Choices	**A: Silvery plaques on extensor surfaces (GROUND TRUTH)** B: Flesh-colored papules in the lumbosacral C: Erosions of the dental enamel (PREDICTED) D: Holosystolic murmur at the left lower sternal border
Context	- Onychomycosis is more common in older adults and in persons with vascular disease,diabetes mellitus, and trauma to the nails . . . - Dermatophyte infections of the nails respond only to prolonged administration. - Finger nails may respond to 6 months of therapy, whereas toe nails are recalcitrant. . . - Chronic mucocutaneous candidiasis is a heterogeneous infection of the hair, nails , skin. - The condition may be mild and limited to a specific area of the skin or nails ,or it may take a... - A fine lanugo covers the face, body, and limbs. The skin is thin and dry, without its normal elasticity, and the nails are brittle. The dental enamel is eroded .

Table 8: Error Analysis Example of ColBERT with BERT as reader.

	without ColBERTv2		with ColBERTv2	
Model	Dev	Test	Dev	Test
BERT$_u$	0.276	0.258	**0.312**	**0.302**
BioRoBERTa	0.292	0.295	**0.329**	**0.318**

Table 9: MedQA Results using Different Readers and ColBERTv2.

7 Conclusion

OpenQA for medical domain is continously being explored either with new state-of-the-art models or techniques made available in NLU or by contributions in other areas such as Information Retrieval, like ColBERTv2, where retrieval is done by leveraging the power of contextualized retrieval and getting good performance with out-of-domain data, as part of this we decide to step further with contributing towards the evolution of medical OpenQA systems by evaluating multiple datasets in the medical domain, obtaining results that outperform other IR mechanisms and reader models for answer extraction.

Even though we did not outperform results reported by MedQA, we think that further analysis and exploration should be done utilizing ColBERT indexing and retrieval as we belief and observed during experimentation that it could obtain a boost on building end-to-end OpenQA systems. Furthermore, we are encouraged to continue exploring how we can improve the results already obtaining perhaps by experimenting with autoregressive LMs to do passage compression and query re-writing before passing the context to the reader or evaluating reasoning by using autoregressive models instead of reader models with a head classifier on top of the CLS token classification output.

8 Acknowledgements

We would like to express our gratitude and thanks to Dr. Christopher Potts, Sebastian Hurubaru and Amelia Hardy for all the guidance during office hours, encouragement and feedback that allowed us to push further and be motivated to pursue and complete the project.

9 Authorship

All authors of this paper contributed equally to the work. Laura Uzcategui, was responsible for building and implementing in code the end-to-end OpenQA system in Python and initial rounds of fine-tuning. Young Don Ko, was responsible of running all fine-tuning and evaluation of the models for larger epochs. Research, analysis and paper writing was done equally by both authors.

References

Jafar A Alzubi, Rachna Jain, Anubhav Singh, Pritee Parwekar, and Meenu Gupta. 2021. Cobert: Covid-19 question answering system using bert. *Arabian journal for science and engineering*, pages 1–11.

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

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow, march 2021. *URL https://doi. org/10.5281/zenodo*, 5297715.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rakesh Chada and Pradeep Natarajan. 2021. FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6081–6090, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bijoyan Das and Sarit Chakraborty. 2018. An improved text sentiment classification model using TF-IDF and next word negation. *CoRR*, abs/1806.06407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Duy Dinh and Lynda Tamine. 2011. Irit at trec 2011: Evaluation of query expansion techniques for medical record retrieval. In *TREC*.

Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv*, abs/2109.10086.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11:6421.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019a. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, USA. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019b. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

O. Khattab and Matei A. Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *ArXiv*, abs/1907.11692.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer.

A MedQA Dataset format

Over section 3.2 we have described MedQA dataset, to be more descriptive, we show in Table 10 a concrete example of MedQA dataset to illustrate the format we have to work with for pre-processing of the data, each observation has a question that could be accompanied of a description as shown, the answer as a text and a list of options as a character with their respective possible choice of answer and the answer index to represent the position in which the correct option is. This index was used as label for the reader classifer on top of the CLS output token.

Field	Description	Example
Question	A sentence asking for particular knowledge or a paragraph with a description of a medical condition	A 17-year-old female is brought to the emergency room by her parents shortly after a suicide attempt by aspirin overdose. Which of the following acid/base changes will occur FIRST in this patient?
Answer	The answer represented as a short sentence.	Respiratory alkalosis
Options	A list of possible choices for the answer.	A: Non-anion gap metabolic acidosis B: Anion gap metabolic acidosis C: Metabolic alkalosis D: Respiratory alkalosis
Answer Idx	A letter indicating the Correct answer.	D

Table 10: MedQA Dataset format and examples.

B Qualitative Analysis Results from Passage Retrieval

As mentioned over section 6, we perform a qualitative analysis on passage retrieval, as we wanted to observe the quality of those comparing between Elasticsearch and ColBERTv2, Table 11 and 12 shows a clear example of how Elasticsearch is performing retrieval on what it seems to be based on term matching, meanwhile ColBERTv2 passages seems to be more meaningful and contextualized.

Question: What is the nervous system
Elasticsearch Passages retrieved
- "What are the main divisions of the central nervous system" - "What are the major cell types of the central and peripheral nervous systems" - " What is important for the neurologist is the fact that the nervous system is involved in most of infants with major malformations" - "What role does plasticity play in neural development and in response to damage of the nervous system" - "Functionally the nervous system is organized into a somatic nervous system and visceral nervous system" - "What information about the state of the muscle is sensed by the muscle spindles and what afferent fibers convey this information to the central nervous system CNS" - "Nervous System" - "CNS central nervous system ENS enteric nervous system" - "Activity in the enteric nervous system is modulated by the sympathetic nervous system"
ColBERTv2 Passages retrieved
- Anatomically, the nervous system is divided into the following The central nervous system CNS consists of the brain and the spinal cord, which are located in the cranial cavity and spinal canal, respectively. The peripheral nervous system PNS consists of cranial, spinal, and peripheral nerves - Flint Beal cause of both frontotemporal dementia and amyotrophic lateral scle. The human nervous system is the organ of consciousness, cognition, ethics, and behavior as such, it is the most intricate structure known to exist. - The somatic nervous system consists of nerves that carry conscious sensation from peripheral regions back to the CNS and nerves that exit the CNS to innervate voluntary skeletal muscles. In contrast, the visceral nervous system consists of nerves that carry sensory information into and motor autonomic innervation out of the CNS to regulate homeostatic functions - When studied through imaging or in histopathology, the nervous system is observed in sections cut from one of three different planes a coronal plane, which divides the nervous system into anterior and posterior parts the sagittal plane, which is oriented at a right angle to the coronal plane and divides the nervous system into left and right parts. - The central nervous system consists of the brain located in the cranial cavity and the spinal cord located in the vertebral canal. The CNS is protected by the skull and vertebrae and is surrounded by three connective tissue membranes called meninges. The brain and spinal cord essentially float in the cerebrospinal fluid that occupies the space between the two inner meningeal layers.

Table 11: Question 1: What is the nervous system - Comparison between passage retrieved

Question: What is PCOS
Elasticsearch Passages retrieved
- "PCOS" - "PCOS PCOS with metabolic PCOS with metabolic syndrome syndrome and Type 2 diabetes mellitus overt vascular and or renal disease" - "Treatment of PCOS in adolescence" - "Treatment of Hyperandrogenism and PCOS" - "polycystic ovary syndrome PCOS" - "PCOS polycystic ovarian syndrome" - "PCOS Polycystic ovary syndrome" - "Genetic studies of PCOS reported allele sharing in large PCOS patient populations and linkage studies focused on candidate genes most likely to be involved in the pathogenesis of PCOS" - "The most prevalent are those characterized by androgen excess often with insulin resistance including what is arguably the most common endocrinopathy in women polycystic ovary syndrome PCOS " - "PCOS Irregular menses slow onset hirsutism obesity infertility hypertension a family history of PCOS or DM"
ColBERTv2 Passages retrieved
- The 2003 Rotterdam Consensus Workshop concluded that PCOS is a syndrome of ovarian dysfunction along with the cardinal features hyperandrogenism and polycystic ovary PCO morphology Table 31. 4. It is recognized that women with regular cycles, hyperandrogenism, and PCO morphology may be part of the syndrome - PCOS is arguably one of the most common endocrine disorders in women of reproductive age, affecting 5 to 10 of women worldwide. This familial disorder appears to be inherited as a complex genetic trait - Polycystic ovary syndrome PCOS is a common disorder that affects premenopausal women and is characterized by chronic anovulation and hyperandrogenism Chap. 412. Insulin resistance is seen in a significant subset of women with PCOS, and the disorder substantially increases the risk for type 2 DM - Because the syndrome is heterogeneous and poorly defined, clinical difficulties result in diagnosis and management 90. For the sake of simplicity, PCOS may be defined as LH dependent hyperandrogenism - Although the complete list of potential causes is long, as noted below, the most common causes of amenorrhea in women with normal secondary sexual characteristics and normal pelvic are pregnancy, polycystic ovarian syndrome, hyperprolactinemia, primary ovarian insufficiency also known as premature ovarian failure

Table 12: Question 2: What is PCOS - Comparison between passage retrieved