# Retrieval-Augmented Generation Facilitated Medical Question Answering

Yun-Chung Liu

## Abstract

Accurate machine medical question answering can provide actionable insights for caregivers and reduce their workload. Recent advances in large language modeling have improved the performance in various language tasks. The concept of retrieval-augmented generation (RAG) inspired subsequent endeavors on machine question answering, especially in domains where knowledge is updated rapidly, such as biomedical research. This study aims to investigate the effectiveness of applying RAG on medical question answering. MedQuAD, a medical question answer dataset containing more than 200 thousand question-answer pairs, was used for the study. Accuracy and F-1 score were applied to evaluate the generated yes-no answers. ROUGE-L and BLEU scores were used to assess the quality of generated long answers. With correctly paired question-document as input, the test accuracy and F-1 score achieved 74.49 and 81.41, higher than accuracy and achieved using questions alone (63.71 and 77.01 respectively). The ROUGE-L and BLEU scores using the correct question-document pair as input were 12.66 and 13.72, higher than using questions with retrieved documents and questions only as input. The results show the potential of RAG to improve the performance of machine medical question answering. Further research is needed to increase accuracy for clinical use.

## Introduction

Medical question answering can benefit caregivers by reducing workload and providing actionable insights [1]. The challenge of medical question answering is to generate accurate responses using natural human languages. Recent advances in large language models (LLMs) have dramatically enhanced machine capacity to generate natural language [2, 3]. However, these language models can generate counterfactual responses, for example, a recommended procedure to a medical condition that is *perfect in language usage* but *does not help address* the issue faced by the patient. With new findings being published on a regular basis, the problem becomes even more challenging. A document stating the newest findings might be proved wrong or irrelevant very soon. LLMs are trained on a large amount of documents to generate the most likely (or probable) response. However, the most probable response might not be the most accurate. This is especially true in the medical domain, where new knowledge, medicine, and procedures are being discovered on a daily basis. Generating the *most probable* answers to medical questions might be just wrong and lead to fatal consequences.

One approach to generating accurate answers to questions is the retrieval-augmented generation (RAG) [4] algorithm, which *retrieves* the most relevant documents to a question first and generates responses based on relevant documents. One approach to retrieve the most relevant document is to represent both the questions (or prompt) and documents using embedding (e.g. word2vec [5], Glove [6], or BERT [7]). Documents with the most similar embeddings (e.g. measured with Euclidean distance) would be retrieved for answer generation. Numerous attempts have been made to apply RAG for question answering and text generation. However, the application of medical question answering has not yet reached satisfactory results for real world deployment.

The aim of this project is to explore the possibility of leveraging RAG to increase the performance of medical question answering. Specifically, different inputs (question only, question concatenated documents associated with the question) for answer generation will be tested. The associated performances on medical question answering will be compared. The result of this project can be generalized to any other domain that requires up-to-date information for accurate question answering.

## Methods

### Dataset

The dataset used for this study is PubMedQA , a question-answering dataset containing three subsets, PQA-L, PQA-U, and PQA-A, with 1, 61.2, and 211.3 thousand question-answer pairs respectively [8]. The questions and answers were obtained from PubMed, one of the most popular biomedical research paper databases. Research papers with structured abstracts (i.e. with sections such as background, methods, results, and conclusion) were selected in this dataset. For papers with a yes-no question as the title, the titles were selected directly as *questions* in PQA-L and PQA-U. In the PQA-L (labeled) subset, human experts manually annotated the yes-no answer (yes: 55.2%, no: 33.8%, maybe: 11.0%). In the PQA-U (unlabeled) subset, yes-no answers were *not* provided. For the other papers, the titles, which are statements, were *converted* into questions, which formed the questions in the PQA-A (artificial) subset. For example, the title of the paper ”Spontaneous electrocardiogram alterations *predict* ventricular fibrillation in Brugada syndrome.” was transformed as “*Do* spontaneous electrocardiogram alterations *predict* ventricular fibrillation in Brugada syndrome?” and the yes-no answers were produced automatically (yes: 92.8 %, no: 7.2) according to the nature of the statement (positive or negative). For questions from all three datasets, the *long answers* to the question were the *conclusion* section of the abstract.

### Preprocessing

The PQA-A subset was randomly split into training and validation dataset using a 9:1 ratio for the yes-no answer prediction task. The PQA-L, the human annotated dataset, was used as the test dataset for both yes-no questions and long-answer generation. All textual data were tokenized using the pretrained BERT model [7] using Hugging Face’s transformers library (version 4.40.1).

### Document Retrieval

All abstracts in the PQA-L subset were treated as candidate documents. All the questions and candidate documents were represented using the *[cls]* token of the pretrained BERT model . The question-document pair with the highest cosine similarity score was retrieved.

### Yes-No Answer Generation

The yes-no answer generation task was treated as a binary outcome prediction task. Input text was represented with the BERT model's [*cls*] token. The pretrained BERT model was applied, with a linear model on top of the attention layer. The final output passed through a *tanh* activation for the binary classification task. The model was fine-tuned on the PQA-A dataset after preprocessing with 3 epochs (learning rate: .000, .batch size: 8. The PQA-L dataset was used to test the performance of the model, using three kinds of input (question only, question plus the retrieved most relevant paper abstract, question plus the *actual* paper abstract). Questions with "maybe", instead of yes/no, as answers were excluded during the test phase. Overall accuracy and F1 score were used to assess the performance.

### Long Answer Generation

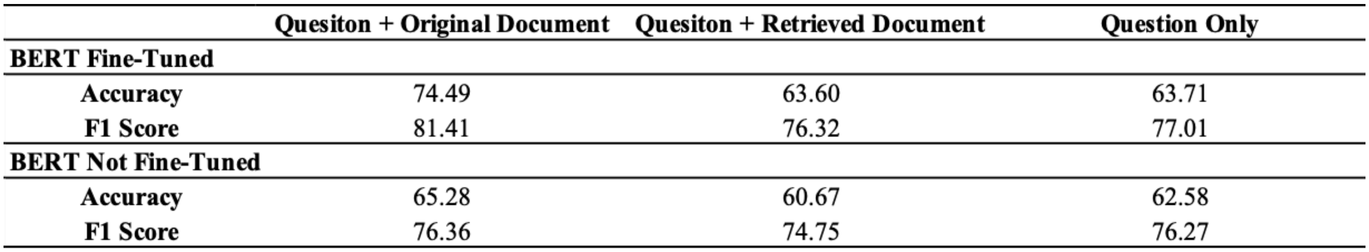
For the long answer generation task, the GPT-2 [9] model was used to generate answers to questions using different inputs (question only, question plus the retrieved most relevant paper abstract, question plus the *actual* paper abstract). The ROUGE (Recall-Oriented Understudy for Gisting Evaluation)-L and one-gram BLEU (BiLingual Evaluation Understudy) score [10] were used to assess the quality of the generated response. All the reported results were assessed using the held-out human annotated dataset (PQA-L).

## Results

### *Yes-No Answers*

Table 1 shows the results of yes-no question answering on the held-out dataset (PQA-L). Using questions concatenated with original documents (the source research papers) as model input, the accuracy and F1 score reached 74.49 and 81.41 if the parameters of the BERT was fine-tuned on the PQA-A dataset using the correct question-abstract. If only the weights of the final layer were fine-tuned (BERT Not Fine-Tuned), the test accuracy and F1 score were 65.28 and 76.36. The yes-no answers generated using questions concatenated with original documents as input outperformed answers generated with questions concatenated with the retrieved most probable document and question only.

**Table 1.** Test performance of yes-no answers

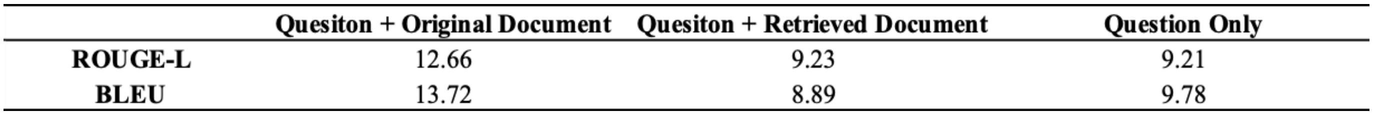


| | Quesiton + Original Document | Quesiton + Retrieved Document | Question Only |
|---|---|---|---|
| **BERT Fine-Tuned** | | | |
| **Accuracy** | 74.49 | 63.60 | 63.71 |
| **F1 Score** | 81.41 | 76.32 | 77.01 |
| **BERT Not Fine-Tuned** | | | |
| **Accuracy** | 65.28 | 60.67 | 62.58 |
| **F1 Score** | 76.36 | 74.75 | 76.27 |

### *Long Answers*

Table 2 presents the performance of generated long answers to biomedical questions in the test dataset. Using the question plus the relevant document, the resulting ROUGE-L and BLEU score reached 12.66 and 13.71, higher than long answers generated using questions with retrieved documents and questions-only as input.

**Table 2.** Test Performance of generated long answers



| | Quesiton + Original Document | Quesiton + Retrieved Document | Question Only |
|---|---|---|---|
| **ROUGE-L** | 12.66 | 9.23 | 9.21 |
| **BLEU** | 13.72 | 8.89 | 9.78 |

## Discussion

This study explored the effectiveness of retrieval augmented answer generation for biomedical questions. The results showed that accurate document retrieval can largely improve the performance on the medical question answering (both yes-no and long answers) using the PubMedQA dataset. With the correct question-document pairs as input, the accuracy and F1 score for yes-no questions reached 74.49 and 81.41, the ROUGE-L and BLEU score for generated long answers were 12.66 and 13.72, outperforming answers generated with questions only or questions concatenated with the retrieved most relevant document.

The reason for the better performance achieved with correctly paired documents could be that the document provides the *context* needed for generating answers to the questions posed. The strength of large language models, such as BERT and GPT-2 used for yes-no answer prediction and long answer generation, are their capacity of representing documents holistically. With the correct context provided, the models are able to optimize the weights to produce better predictions on the task. On the other hand, inaccurate question-document pairing can hinder answer generation performance.

Although current results suggest that medical question answering performance could be improved with accurate documents retrieved, the proper algorithms for retrieving the most relevant document remains yet to be explored. One potential enhancement is to use the BERT model trained on biomedical literature (e.g. BioBERT [11] ) for the document representation. The model used for document retrieval tasks can also be fine-tuned on a larger set of question-context pair data.

The metrics used to evaluate the quality of long answers to medical questions are imperfect. Common metrics, such as BLEU, ROUGE, and METEOR scores, were reported to be positively associated with human judgment in general question answering tasks. However, the relationship between yes-no answers and long answers generated is *unclear*. A generated response with a high BLEU, ROUGE, or METEOR score can simply be *counterfactual*. Further research is needed to address this issue.

Several other limitations exist for the current study. First of all, due to the constraint of available computational resources, the amount of documents used in the project is limited. This harms the generalizability of the current result because in the real world settings, there would be even more candidate documents to choose from, making the retrieval task even more challenging. Also, more thoroughly model fine-tuning for both the yes-no answer prediction and the long answer generation task could further improve model performances. The performance achieved by current models are still largely inferior to human performance [7]. Addressing these limitations can be the next steps of this present exploration.

## Conclusion

This study demonstrates how identifying the correct document associated with biomedical questions can improve performance on medical questions answering both on factual (yes-no) and long answer (conclusion) generation. With large language models such as BERT and GPT-2, inputting questions with the correct question-document pair resulted in testing accuracy and F1 score of 74.49 and 81.41 for yes-no question, and ROUGE-L and BLEU score of 12.66 and 13.72 for generated long answers, The algorithms to retrieve the correct document and the metric to evaluate quality remains to be improved, which could be the next steps for following research endeavors.

## References

1. Ben Abacha, Asma, and Dina Demner-Fushman. "A question-entailment approach to question answering." *BMC bioinformatics* 20 (2019): 1-23.
2. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
3. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
4. Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.
5. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).
6. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. (2014).
7. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
8. Jin, Qiao, et al. "Pubmedqa: A dataset for biomedical research question answering." *arXiv preprint arXiv:1909.06146* (2019).
9. Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.
10. Chen, Anthony, et al. "Evaluating question answering evaluation." *Proceedings of the 2nd workshop on machine reading for question answering*. (2019).
11. Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.