Enhancing Clinical Diagnosis with Advanced NLP: A Comparative Study of Retrieval Augmented Generation Models

Dong Jian, Kirti Prakash

Abstract

After a series of experiments with Natural language processing(NLP) models that combine elements of both retrieval and generation methods to answer clinical diagnosis and medical questions, we demonstrate that a Retrieval Augmented Generation (RAG) model using a sentence transformer multi-qa-mpnet-base-dot-v1[1] as the encoder and GPT-3.5-turbo as the decoder achieves excellent results on multiple metrics such as correctness, faithfulness, and relevancy. Additionally, we introduce a second RAG model with a sentence transformer all-MiniLM-L6-v2[2] as the encoder and GPT-3.5-turbo as the decoder. This achieves better results: 20 percentage points improvement in the correctness from 58% to 78%. The implications of these enhancements offer significant promise for improving clinical decision-making tools, aiding healthcare providers with more accurate and reliable AI-assisted diagnostics.

1 Introduction

Healthcare and medicine are critical components in people's daily lives. Yet, many of today's AI models in these fields do not fully utilize the latest technologies to meet these needs, often failing in tasks such as question-answering with the desired accuracy. In the clinical field, high accuracy and fairness are required as people's lives are at stake. Considering these gaps, we introduce a series of experiments utilizing Retrieval Augmented Generation (RAG) models that improve the correctness score.

2 Background

Recent advancements in natural language processing (NLP) have significantly improved medical applications of language models. Google's Flan-PaLM, for example, achieved 67.6% accuracy[8] on USMLE-based MedQA questions, marking a 17% improvement over previous models and demonstrating its capability to handle complex medical inquiries. However, this paper introduces a more sophisticated Retrieval Augmented Generation (RAG) model, achieving 78% correctness[definition in this paper] in medical question answering. It's important to note that this

[1] It maps sentences & paragraphs to a 768-dimensional dense vector space and was designed for semantic search. It has been fine-tuned using 215 million (question, answer) pairs from diverse sources.
[2] It maps sentences & paragraphs to a 384-dimensional dense vector space and can be used for tasks like clustering or semantic search. It has been fine-tuned using more than 1 billion sentence pairs.

78% may not represent an apple-to-apple comparison with Flan-PaLM's results, as the evaluation method might differ.

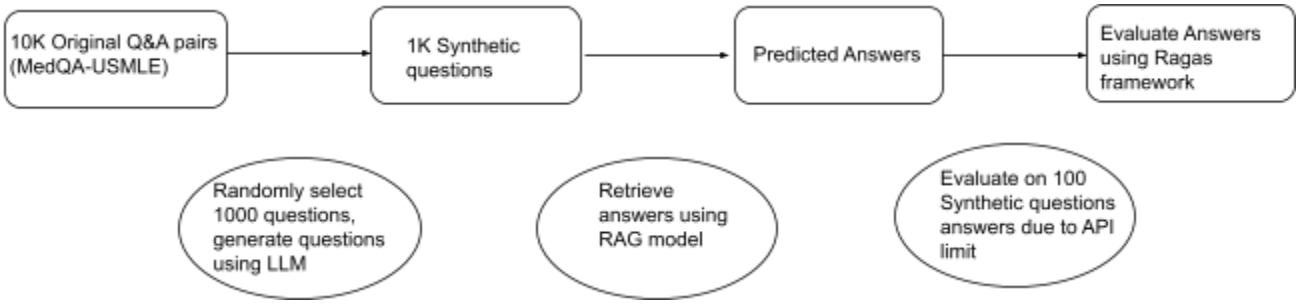
3 Model

Our study utilizes two configurations of the Retrieval Augmented Generation (RAG) model. The first configuration, the RAG1 model throughout this paper, incorporates a sentence transformer named multi-qa-mpnet-base-dot-v1 as the encoder and GPT-3.5-turbo as the decoder. The second model, the RAG2 model, updates the encoder to all-MiniLM-L6-v2 while keeping the rest of the model the same as RAG1. The Retrieval Augmented Generation (RAG) model plays a crucial role in combining retrieval and generative techniques to enhance the quality and relevance of responses in our described NLP models, particularly for complex tasks such as answering clinical diagnoses and medical questions.

4 Datasets and Experimental Setup

We employed the MedQA_USMLE[0] (United States Medical License Exams) dataset, consisting of 10,178[3] question-and-answer pairs, due to its comprehensive coverage of clinically relevant topics and alignment with US medical licensing standards. The average token length of the questions is 174, with a median of 162 and a maximum of 911. See **Table 1 in the appendix** for data examples.

Below is the model pipeline chart:



For our experiments, we randomly selected 1,000 questions to generate synthetic questions and retrieved corresponding answers. Then evaluate 100 answers. This is to manage computational resources effectively and adhere to API usage constraints.

[3]MedQA data Data source

4.1 Top k and chunk size selection

The first parameter we experimented with is the top k of the RAG1 model. "top k" refers to the number of top retrieved documents or passages from a large corpus, based on their relevance to the input query. We tried top k set to 5 and 10 using the RAG1 model with the same chunk size 650. To get a quick turnaround to see the direction of the change, we tested on 5 same random question-and-answer pairs. Changing top k from 5 to 10 results in no significant change in correctness but lower faithfulness, context precision, and recall. This could be due to the noise introduced to the model when more documents retrieved with the k increased to 10. Secondly, we experimented with the chunk sizes 256, 650, and 800 using the RAG1 model and top k set to 5. The optimal chunk size is 650, with the highest correctness and across other metrics in general. Thus, we set our model with k = 5 and chunk size = 650.

4.2 Model Variations

RAG1 model is the RAG model with multi-qa-mpnet-base-dot-v1 as encoder and GPT-3.5-turbo as decoder. This is our baseline model. RAG2 is the RAG (Retrieval Augmented Generation) model with all-MiniLM-L6-v2 as encoder and GPT-3.5-turbo as decoder. The all-MiniLM-L6-v2 sentence transformer utilizes the pre-trained MiniLM-L6-H384-uncased model and is fine-tuned with a 1 billion sentence pairs dataset, including GOOAQ[1], SQuAD2.0[2], SearchQA[3], and PAQ[4]. The all-MiniLM-L6-v2 is based on the MiniLM architecture[5], known for its smaller, more efficient models that retain the performance characteristics of larger models. MiniLM models are designed to distill knowledge from larger models into a smaller form factor, often leading to efficient inference times without a significant drop in performance. It uses a deep self-attention mechanism that can capture more nuanced relationships in the data than other architectures. In contrast, the multi-qa-mpnet-base-dot-v1 transformer used a pre-trained "mpnet-base" model and fine-tuned with 215 million(question, answer) pairs, lower than all-MiniLM-L6-v2.

4.3 Evaluation

Prior works developing models for clinical questions and answers have focused on metrics such as classification accuracy, or natural language generation metrics such as BLEU[6] scores that fail to capture the clinical quality of consultations. Moreover, evaluating RAG architectures poses a significant challenge because of the different factors involved in evaluation namely: the capability of the retrieval system to find the correct and relevant context reference passages, the ability of the large language model (LLM) to present these passages in a grounded and faithful manner, and the overall quality of the generated output.

As we did not have specially trained resources for human evaluation, we sought to use the LLM-assisted evaluation. We choose to use the Ragas (RAG Assessment) framework[7] for evaluating ground truth answers. This is growing in popularity as evidenced in recent online posts referencing this framework at Amazon and Microsoft. With Ragas, we choose metrics: 1) for retrieval evaluation: context precision, context recall, context entity recall 2) for generation evaluation: faithfulness, answer relevancy, answer similarity 3) for overall quality: answer

correctness (see definitions below). These metrics were chosen to not only evaluate the accuracy of the answers provided but also how well the answers align with reliable source documents, the relevance of the answers to the questions posed, and the ability of the models to retrieve pertinent information from a comprehensive knowledge base. This approach provides a more nuanced understanding of the model's capabilities than traditional metrics. Using the LLM, we synthesized 1000 similar questions and retrieved corresponding answers. The initial intent is to have more predicted answers generated for evaluation, but the API limits such as TPM (token per min) and RPD (request per day) on GPT-3.5-turbo made us limit this to a random sample of 100 question-and-answer sets for each model's evaluation.

- Answer correctness: Gauges the accuracy of the generated answer when compared to the ground truth, looking at two key aspects of semantic similarity and factual similarity.
- Faithfulness: Aka. Groundedness, indicates how well the answers generated by the model align with the retrieved context from source documents.
- Answer relevancy: Assesses how related the generated answers are to the questions asked. For example, the score is 0 if the answer is "I don't have the specific information to answer your question about the pathogenic mechanism in the scenario you provided."
- Context precision: Evaluates how precise the model is in selecting top k ranked document chunks that are relevant to the given question.
- Context recall: Measures the model's ability to retrieve all relevant document chunks related to the given question.
- Context entity recall: Measures how well the retrieval component of the system is able to fetch relevant entities within the context of the query or the task at hand. An "entity" in this sense typically refers to specific pieces of information that are relevant to the query, such as names, places, facts, or other identifiable data points. The metric assesses the model's ability to recall (retrieve) these entities accurately from the knowledge base or data set it has access to. High scores in this metric indicate that the system is effective at identifying and retrieving the specific entities relevant to the task or query.
- Answer similarity: Assesses the semantic resemblance between the generated answer and the ground truth.

The RAG1 model's correctness evaluated on 100 random question-and-answer pairs seems to have been dragged down by low context recall. See **Table 2 in the appendix** for evaluation results. Here is an example with low correctness and context recall when the predicted answer does not answer the question's intent: The **question** is "What is the most appropriate pharmacological **treatment** for the underlying condition of a 13-day-old male presenting with eye redness, ocular discharge, cough, nasal discharge, and lung findings of hyperinflation with bilateral infiltrates, despite a negative fluorescein test, following limited prenatal care and administration of silver nitrate drops and vitamin K post-delivery?" The **ground truth answer** is **"Oral erythromycin"** as the question asks for treatment, but the **predicted answer** is "The most likely **diagnosis** for the 13-day-old male with the described symptoms is **Oligoarticular juvenile idiopathic arthritis**." Here the predicted answer is a diagnosis instead of a treatment. The answer correctness is only 0.19, the faithfulness score is 0.5, context recall is 0.33, and context entity recall is 0, while context precision is 0.99. Although the top context retrieved is

also a 13-day-old male with similar symptoms(high context precision) it did not retrieve the treatment “Oral erythromycin”(low context recall and context entity recall is 0).

Thus, we proposed the second RAG model RAG2 with context embedding updated to all-MiniLM-L6-v2 and kept all other parameters the same. After applying the RAG2 model, we evaluated again with the same framework. The new model produced a 20 percentage point higher correctness and 11 percentage point higher context recall with other metrics all slightly higher(3-7 percentage point). Taking the same 13-day-old male example above, the RAG2 model **predicted the answer** as “The most appropriate **pharmacotherapy** for the underlying condition of a 13-day-old male with the described symptoms is **oral erythromycin**.” This captured the **ground truth answer “oral erythromycin”**, and **improved the context recall to 1**. Also, correctness and faithfulness also improved to 0.97 and 1 respectively. The encoder is responsible for converting both the question and the documents (or document chunks) into embeddings in a shared space. Fine-tuned with 1 billion sentence pairs, all-MiniLM-L6-v2 generates better embeddings that capture the essence of the texts more effectively. Consequently, the model's retrieval component can more accurately match the question embeddings with the relevant document embeddings. When the encoder is upgraded to a more advanced model like all-MiniLM-L6-v2, it boosts the foundational capabilities of the RAG model. This includes improved language comprehension and a more effective representation of both questions and context in the embedding space. These enhancements directly contribute to higher correctness in answers and better retrieval of relevant document chunks, explaining the observed improvements in the correctness and context recall metrics.

5 Limitations and Future Improvements

This study faces several limitations, including reliance on limited pre-existing datasets and lack of real-world testing, which may impact the generalizability of our results. Additionally, the computational demands of such advanced models may restrict their deployment in real-time clinical applications. To enhance model performance, we propose developing a third variant using BioLinkBERT-base, an encoder pre-trained on PubMed, if additional time and resources become available. Furthermore, involving medical professionals and licensed bodies in the evaluation process is crucial to ensure the models' efficacy and appropriateness for clinical use provided by the necessary medical reviews and feedback.

Also, the current models focus on generating answers in a conventional question-and-answer format. However, real-world clinical interactions often resemble a dialogue more than a simple Q&A. Future iterations of our models could aim to produce answers in a dialogue fashion. This approach would not only make the interactions more natural and user-friendly but could also handle follow-up questions, providing a more comprehensive and practical tool for clinical settings.

6 Conclusion

This study has demonstrated significant advancements in applying Retrieval Augmented Generation (RAG) models for clinical diagnosis and medical question answering. By upgrading the encoder to the all-MiniLM-L6-v2 model, we have observed a remarkable improvement in model performance across various metrics including correctness, context recall, and faithfulness. The updated model not only enhances the accuracy of responses but also ensures a deeper contextual understanding, which is critical in medical applications where precision is paramount.

These findings have profound implications. They suggest that even minor enhancements in the encoding mechanisms of AI models can lead to substantial improvements in the reliability and utility of AI-driven clinical question-and-answer models. This is especially relevant in the medical field, where the accuracy of information can significantly affect patient outcomes.

7 Appendix

Table 1: An example of a question-and-answer pair used from the dataset.

Question	Ground Truth Answer
A 13-day-old male is brought in by his mother for eye redness and ocular discharge. Additionally, the mother reports that the patient has developed a cough and nasal discharge. Pregnancy and delivery were uncomplicated, but during the third trimester, the mother had limited prenatal care. Immediately after delivery, the baby was given silver nitrate drops and vitamin K. Upon visual examination of the eyes, mucoid ocular discharge and eyelid swelling are noted. A fluorescein test is negative. On lung exam, scattered crackles are appreciated. A chest radiograph is performed that shows hyperinflation with bilateral infiltrates. Which of the following is the best pharmacotherapy for this patient's underlying condition?	Oral erythromycin

Table 2- Evaluation metrics: the 2 highlighted rows are the RAG1 and RAG2 differences. The first row shows no material difference when dropping the sample size from 950 to 100. The bottom 4 rows show the top k and chunk size experiments in 4.1.

Model(top k, chunk size)(score range 0 to 1)	# of random Q&A pairs evaluated	Accuracy (Correctness)	Faithfulness	Answer relevancy	Context precision	Context recall	Context entity recall	Answer Similarity
RAG1(k=5, Chunk size=650)	**950**	0.61	0.86	0.82	0.78	0.60	0.21	0.84
RAG1(k=5, Chunk size=650)	**100**	0.58	0.83	0.81	0.76	0.63	0.24	0.84
RAG2(k=5, Chunk size=650)	**100**	0.78	0.89	0.79	0.83	0.74	0.27	0.88
RAG1(k=5, Chunk size=650)	**5**	0.66	0.72	0.82	1	0.87	0.20	0.82
RAG1(k=10, Chunk size=650)	**5**	0.67	0.66	0.83	0.80	0.80	0	0.86
RAG1(k=5, Chunk size=256)	**5**	0.25	0.95	0.67	0.60	0.55	0	0.77
RAG1(k=5, Chunk size=800)	**5**	0.51	0.75	0.85	1	0.80	0.20	0.85

8 Reference

0. Jin et al. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams.arXiv:2009.13081v1 (2020)

1. Khashabi et al. GOOAQ: Open Question Answering with Diverse Answer Types. arXiv:2104.08727v2 (2021)

2. Rajpurkar et al. 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. (2018)

3. Dunn et al. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. arXiv:1704.05179(2017)

4. Lewis et al. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. arXiv:2102.07033(2021)

5. Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957(2020)

6. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation in Proceedings of the 40th annual meeting of the Association for Computational Linguistics (2002), 311–318.

7. Es et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217v1 (2023)

8. Singhal et al. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138(2022)

Table of Content