Retrieval Augmented MedLM

Sharmila Devi
AI Consultant, AI Services
Google Cloud Consulting (GCC)
Google
Bengaluru, India
dsharmila@google.com

Gopala Dhar
AI Engineer, AI Services
Google Cloud Consulting (GCC)
Google
Mumbai, India
gopalad@google.com

Chaitanya Bharadwaj
Head of Clinical AI
Apollo247
Bengaluru, India
chaitanya.b@apollo247.org

Abdussamad M
Technical Lead
Apollo247
Bengaluru, India
abdussamad@apollo247.org

***Abstract*—This paper presents a novel approach to leverage large language models (LLMs) for medical question answering (QA) by integrating them with external knowledge sources. We utilize de-identified clinical discharge notes from MIMIC-IV and Apollo Hospitals as our data source. We propose a novel summarization technique that extracts and condenses the core medical information from the discharge notes, eliminating unnecessary verbosity. This results in concise "medical summaries" that effectively inform the LLM while reducing context overload. We evaluate our approach using RAGAS, a novel framework for label-free evaluation of Retrieval-Augmented Generation (RAG) pipelines. Clinician validation further confirms the effectiveness of our approach, highlighting its potential to enhance medical QA systems.**

Keywords—MedLM, RAG, MedPaLM, LLM, GenAI, PaLM2, PaLM, Embeddings, Discharge Notes, Medical QA systems

I. Introduction

In recent years, the development of artificial intelligence (AI) technologies, particularly large language models (LLM) has paved the way for innovative applications in the medical domain. Medical question answering (QA) systems, powered by advanced AI algorithms, have proven to be valuable tools in addressing a wide range of queries related to healthcare, diagnosis, treatment options, and general medical knowledge.

Clinicians and medical practitioners from large providers do not look for generic answers, but instead look for specific responses and citations from their hospital's discharge summaries, prescriptions and clinician notes. It makes sense to use this knowledge base as the grounding factor.

"Retrieval-Augmented Generation" (RAG) is a model architecture that combines elements of retrieval and generation in natural language processing tasks. The RAG model typically consists of two main components:

- Retrieval Component: This part of the model is responsible for retrieving relevant information from a large corpus of documents or knowledge sources.
- Generation Component: Once the relevant information has been retrieved, the generation component generates a response or output in natural language.

II. Experimentations

A. Data

The data that was used for our analysis was the Open Source dataset of MIMIC-IV-Notes: de-identified clinical notes and additionally we also had Apollo Hospital's de-identified clinical discharge notes along with several medical questions and their respective contexts. These notes contained unstructured textual data about the patient's admission history, present ailments, past ailments, allergies, family history etc. These notes were processed in JSONL format. Our dataset contained about 66,041 notes.

B. Approaches and challenges faced

In order to solve the typical Medical Q&A problem, MedLM[1] does a good enough job, however the intent was to enable an LLM to be able to use a data source as its source of information. In our context, it would imply using the MIMIC-IV de-identified discharge notes as the source of information.

In order to achieve this, we utilize the *Vertex textembedding-gecko* embedding model to create embeddings of each of the discharge notes, this embedding is a fixed size vector of length 768. After the embeddings were created, they were indexed and stored in a vector database, the vector database that we chose for this job was *Vertex AI Vector Search.* This enabled us to create an index that could allow swift embedding similarity using Approximate Nearest Neighbor (ANN) search based algorithms.

The initial approach that was undertaken, involved using the aforementioned index as the source of information, the rest of the pipeline consisted of a component responsible for taking input as a "Question" and a "Context". The "Question" and "Context" data was curated by Apollo's clinicians and that can be considered as our test set.

After the "Question" and "Context" data was served as input, the pipeline would run to find the indices of the closest neighbors of the input query in the index that was created, and then the next pipeline would perform a lookup to identify the discharge notes found at those indices. The metric to quantify the similarity was chosen to be cosine similarity, and for experimentation purposes we limited the index search to the top 3 most similar notes from the index.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00217

The information obtained after the lookup would be appended to the initial "Context" before being formatted as a query to be sent to LLM and the response of the LLM would be saved into a dataframe.

As a baseline experiment, the first LLM that we used was the text-bison model. Subsequently, with the same approach we utilized MedLM.

With this approach we ran into issues pertaining to the context length, i.e. the sum of the number of characters across all retrieved contexts was higher than the maximum permissible input token limit of the LLM. The other problem was that because of the large context size, the crux of the information present in the retrieved notes was being lost leading to sub-par results.

In order to solve this problem, we devised a novel approach wherein the discharge notes were not directly used to create an index. It was observed that often a lot of information contained in the notes is English verbiage that adds little to no contextual benefit. Hence, we summarized the notes in such a way that only the medical information should be preserved, while reducing the excess verbiage. In this way, we were able to transform the discharge note into a medical summary document. The architecture is captured in Fig. 1.

While keeping the rest of the process exactly the same, we created a new index with the summarized medical documents and ran a new set of experiments on it.

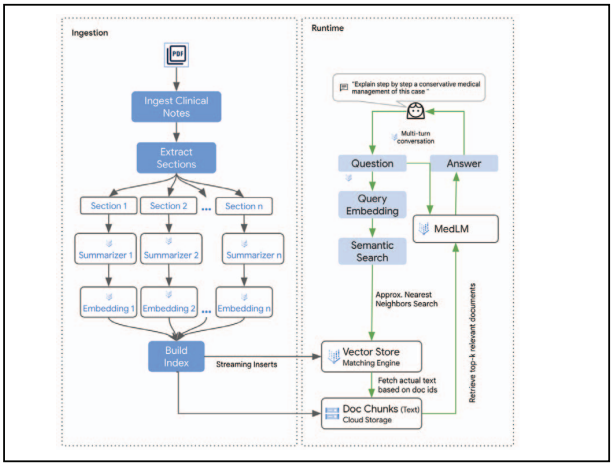

Fig. 1. **Architecture of Retrieval Augmented MedLM**

C. *Evaluations and Results*

In order to quantify the metrics of the RAG(Retrieval Augmented Generation) pipelines we built, we used a novel approach of RAGAS (Automated Evaluation of Retrieval Augmented Generation)[2]. In this framework, we used 4 label-free eval metrics, viz. Faithfulness, Answer Relevancy, Context Precision and Harmfulness.

As mentioned above, we first ran experiments on the baseline LLM, i.e.Vertex PaLM2 text-bison and then on the specialized LLM, MedLM. Both these experiments used the same vector index of discharge notes.

The next set of experiments were run on the summarized discharge notes that utilized the vector index built on them. While evaluating for this approach, we calculate the RAGAS metrics using 2 different contexts, the first one being the context of the parent discharge note, on which the summary had been generated. This was done to validate how much of the "useful" information is retained after the summarization. The second context was the summarized discharge note itself.

As expected the answer_relevancy is higher when the metrics are calculated on the summaries and likewise the context_precision is higher when the metrics are calculated on the parent discharge notes. One thing to note here is that in either of these cases, the vector index used is of the summarized discharge notes.

Apart from using the quantifiable metrics, for certain "Questions" the responses were also vetted by Apollo's clinicians and the observations made by the experts matched the order of our demonstrated results.

TABLE I. EXPERIMENTATION RESULTS

	Faithfulness	**Answer Relevancy**	**Context Precision**	**Harmfulness**
Experiment 1	0.4472	0.42802	0.6	0.4
Experiment 2	1	0.69666	0.4	0.2
Experiment 3	0.7863	0.76214	0.06666	0
Experiment 4	1	0.74096	0.66666	0

- Experiment 1: Text-Bison + Discharge Notes
- Experiment 2: MedLM + Discharge Notes
- Experiment 3: MedLM + Summarized Discharge Notes (metrics calculated on summary)
- Experiment 4: MedLM + Summarized Discharge Notes (metrics calculated on parent notes)

III. CONCLUSION

In this paper, we explored the potential of leveraging large language models (LLMs) for medical question answering (QA) by integrating them with external knowledge sources. We demonstrated that directly using de-identified discharge notes as the information reservoir, while seemingly straightforward, posed challenges related to context overload and subpar LLM performance due to excessive verbiage.

To overcome these limitations, we introduced a novel approach of summarizing discharge notes, extracting and condensing the core medical information while eliminating unnecessary verbiage. This resulted in concise "medical summaries" that effectively informed the LLM while reducing context overload and improving answer relevance.

Our future endeavors will involve exploring advanced summarization techniques, investigating the impact of different LLM architectures, and potentially developing interactive interfaces that allow users to refine and adjust the retrieved summaries for optimal LLM performance.

ACKNOWLEDGMENT

[1] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi et al. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138 [cs.CL], 26 Dec 2022

[2] Shahul Es, Jithin James, Luis Espinosa-Anke, Steven Schockaert et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL], 26 September 2023