A Retrieval-Augmented Generation (RAG) Pipeline for GI-Cancer Prediction and Classification Using Quantized Large Language Models

Arti Jha, Nidhi Shah, Vikas Kumawat and Yashvardhan Sharma

Dept of Computer Science and Information System, Birla Institute of Technology and Science, Pilani

Abstract

This paper presents the implementation of a cancer detection chatbot utilizing a Retrieval-Augmented Generation (RAG) pipeline integrated with Large Language Models (LLMs). The system aims to improve the accuracy and relevance of cancer prediction and classification by leveraging comprehensive text-based medical data related to Gastrointestinal (GI) Cancer, including symptoms categorized by age, gender, stage, and type. The chatbot uses MiniLM L6 v2 for text embedding generation and GPT-3.5 turbo for query response generation, which is stored and retrieved using FAISS as the vector database. A comprehensive comparison of the quantized and non-quantized (Bio-Mistral and GPT 3.5 turbo) for response generation has been presented. The architecture, methodologies, and evaluation metrics used to assess the chatbot's performance are discussed alongside a literature review highlighting advancements in RAG and LLM applications in healthcare, emphasizing this work's significance in cancer diagnosis.

Keywords

Retrieval-Augmented Generation (RAG), Cancer Prediction, Quantized Large Language Models, Question Answering Chatbot

1. Introduction

The integration of artificial intelligence (AI) into medical diagnostics, particularly for cancer prediction and classification, has the potential to revolutionize healthcare delivery. Cancer diagnosis traditionally involves the analysis of diverse text-based data such as clinical notes and patient histories. This paper presents a Retrieval-Augmented Generation (RAG) framework combined with Large Language Models (LLMs) to develop a sophisticated chatbot aimed at assisting clinicians in cancer diagnosis. The chatbot focuses on processing and synthesizing text-based medical data to enhance diagnostic accuracy and relevance.

Contextual Understanding and Generation: Retrieved data is fed into the LLM (e.g., GPT-3.5 turbo), which synthesizes the information into accurate, contextually relevant diagnostic outputs. The LLM leverages its transformer architecture to understand relationships between different data inputs and generate detailed diagnostic reports.

Real-Time Decision Support: The chatbot serves as a real-time decision support tool, continuously updating its knowledge base with the latest medical research, thus aligning its recommendations with current standards of care.

Clinical Efficiency and Personalization: By automating data retrieval and synthesis, the chatbot reduces the cognitive load on clinicians, enhancing diagnostic efficiency. The system also personalizes responses based on specific patient data and clinician preferences, ensuring relevance and applicability.

Forum for Information Retrieval Evaluation, December 12-15, 2024, India

p20210471@pilani.bits-pilani.ac.in (A. Jha); f20212684@pilani.bits-pilani.ac.in (N. Shah); p20210020@pilani.bits-pilani.ac.in (V. Kumawat); yash@pilani.bits-pilani.ac.in (Y. Sharma)

0009-0003-5868-2200 (A. Jha)

1.1. Background and Motivation

Recent advancements in natural language processing (NLP) and AI have enabled the development of sophisticated medical chatbots capable of supporting healthcare professionals in various capacities. The COVID-19 pandemic has accelerated the adoption of AI-driven solutions, particularly in contexts where rapid dissemination of information is critical. The success of these AI systems in managing infectious diseases has prompted further exploration into their applicability in other medical domains, such as oncology. The challenge, however, lies in the effective integration of diverse data sources into a unified diagnostic framework. This paper addresses this challenge by presenting a RAG-based approach, shown in figure 1, that leverages the power of LLMs to enhance cancer prediction and classification.

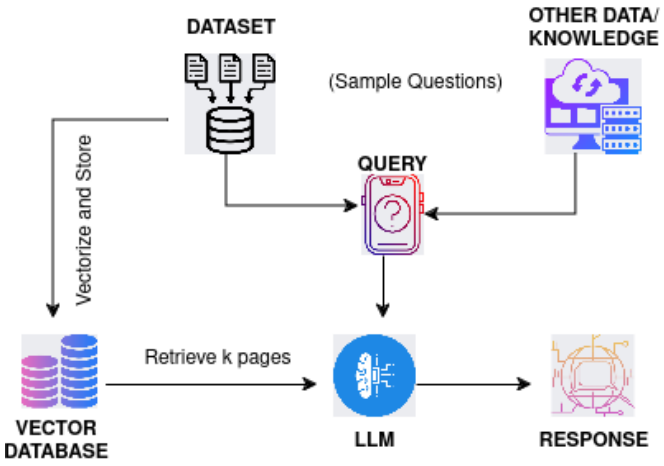

Figure 1: Flow of data while question answering.

2. Related Work

The work discusses the development and evaluation of GastroBot [1], a chatbot [2] designed for gastrointestinal disease inquiries, utilizing Retrieval-Augmented Generation (RAG) [3] technology to enhance response accuracy and relevance. The study reports high System Usability Scale (SUS) scores for GastroBot, indicating superior safety, usability, and smoothness compared to other models. It highlights the importance of integrating external knowledge sources such as chatGPT to address challenges in clinical applications of large language models. A significant amount of prior research has focused on developing architectures that enhance systems with non-parametric memory, which are trained from the ground up for particular tasks, such as memory and stack-augmented network, and memory layers [4]. In contrast, this approach investigates a scenario where both parametric and non-parametric memory components are pre-trained and pre-loaded with extensive knowledge. Importantly, by utilizing pre-trained access mechanisms, the system can access this knowledge without requiring additional training. This work presents Retrieval Under Graph-Guided Explainable disease Distinction (RUGGED) [5], a computational workflow that integrates Large Language Models with Retrieval Augmented Generation to enhance biomedical hypothesis generation. It utilizes a comprehensive knowledge graph enriched with data from various biomedical sources to provide explainable and actionable predictions. The system aids to facilitate researchers in exploring complex biomedical questions effectively and gave to specific answer to them.

3. Methodology

The proposed cancer detection chatbot is built upon a sophisticated RAG pipeline integrated with LLMs, designed to process and synthesize text-based medical data into coherent diagnostic insights.

3.1. Data Collection and Preprocessing

The dataset used in this study includes comprehensive text-based records from Electronic Health Records (EHRs) related to Gastrointestinal (GI) Cancer. The data collection process involved sourcing relevant medical records, ensuring the inclusion of diverse and representative samples of cancer cases, and focusing on text data that describe symptoms, diagnoses, and patient histories. The data collection process involved sourcing relevant medical records from established databases and ensuring the inclusion of diverse and representative samples.

Each dataset undergoes a rigorous preprocessing phase tailored to its specific data type. EHRs are tokenized and parsed to extract key medical concepts, with particular attention given to patient history, symptoms, and previous diagnoses. Imaging data is preprocessed using standard normalization techniques and segmented into regions of interest, focusing on areas indicative of cancerous growths. Genomic data is processed to highlight relevant biomarkers and genetic mutations, which are then encoded into vectors for efficient retrieval.

3.2. Retrieval Mechanisms

This work utilizes dense retrieval for the text-based dataset, using MiniLM L6 v2, a smaller, fine-tuned version of BERT Large optimized for text embedding generation and sentence similarity tasks. The data is stored and retrieved using the FAISS vector database, which facilitates efficient and accurate retrieval of relevant information in response to user queries.

3.3. Integration with Large Language Models(LLM)

3.3.1. Quantized LLM

The conversational chatbot is initially powered by the Bio-Mistral Quantized LLM, a 5GB model developed by Mistral AI. This quantized model facilitates enhanced extensibility and interoperability across diverse computing environments and versions, making it feasible to deploy on low-power CPUs and GPUs. Although the Bio-Mistral model demonstrated significant computational efficiency and rapid response times, its accuracy was suboptimal. Consequently, to improve the accuracy of responses, the system was subsequently upgraded to utilize the non-quantized GPT-3.5 turbo LLM, which provides superior precision in natural language generation and processing.

3.3.2. Non-Quantized LLM

Once the relevant text data is retrieved using FAISS, it is passed to GPT-3.5 turbo, which generates the final diagnostic outputs. GPT-3.5 turbo synthesizes the retrieved data into accurate, contextually relevant diagnostic insights, leveraging its advanced language understanding capabilities. The transformer layers within the LLM capture the intricate relationships between the various data inputs and the query context, enabling the generation of diagnostic reports that are both accurate and aligned with clinical best practices.

4. Experimental Methodology

To rigorously evaluate the performance of the RAG-based chatbot in cancer prediction and classification, we implemented a comprehensive set of experiments designed to assess the impact of different retrieval strategies on key diagnostic metrics. The experiments were conducted on a meticulously curated text-based dataset comprising patient symptoms and corresponding cancer diagnoses sourced from Electronic Health Records (EHRs). This dataset provides a representative sample of GI Cancer cases, enabling the chatbot to match symptoms with potential cancer diagnoses. The dataset comprises over 500 symptoms of cancer cases for diagnosis. The chatbot is designed to interact with users by asking for symptoms and then responding with diagnostic insights or recommendations based on the retrieved

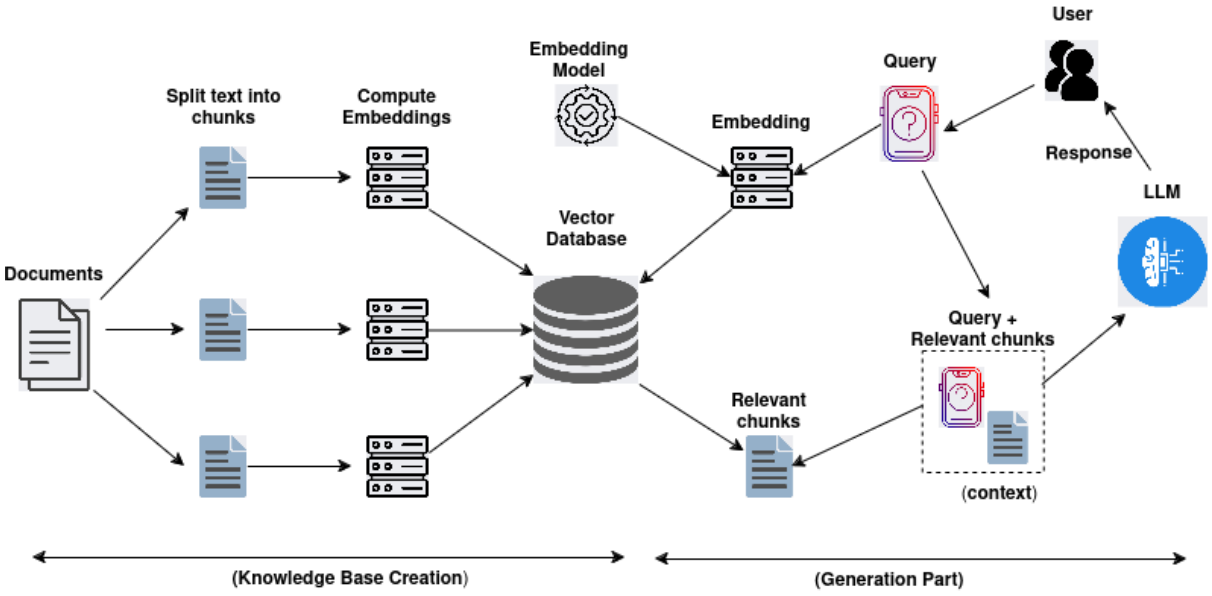

Figure 2: The complete RAG pipeline created, displaying each level under the knowledge base creation and generation parts, respectively.

data. This dataset helped with the purpose of fine-tuning, whereas the dataset provided for the FIRE 2024 was used to evaluate the model with different LLMs. The testing dataset consisted of over 50 questions of symptoms related to GI cancer.

4.1. Experimental Setup

The dataset, consisting of patient symptoms and corresponding medical records, was divided into training and testing subsets, with 80% allocated for training and 20% reserved for testing, and table 2 shows the relevance scores of different models based on test data. This division ensured that the testing data remained unseen during the training phase, allowing for an objective evaluation of the model's generalization capabilities. Cross-validation was done across different data subsets to further validate the robustness of the results.

The relevant data chunks are first extracted from the dataset using Splittext. The retrieved data was then processed by MiniLM to generate embeddings, which were stored in the FAISS vector database. The query from the user then gets converted to embeddings and is compared with data chunks to find relevant chunks from the database. Both query and relevant chunks are then sent to the LLM, which generates diagnostic outputs or recommendations that are subsequently evaluated against established ground truth labels; refer to figure 2. This experimental setup enabled a thorough comparison of each LLM's response (BioMistral and GPT), focusing specifically on how well the chatbot could match symptoms to potential cancer diagnoses.

The experimental results revealed substantial insights into the effectiveness of the RAG-based chatbot in cancer prediction and classification, particularly in its ability to match symptoms provided by the user to potential cancer diagnoses and provide appropriate recommendations.

4.2. Performance Comparison between Bio-Mistral Quantized LLM and GPT-3.5 turbo

The table compares the performance metrics between Bio-Mistral Quantized LLM and GPT-3.5 turbo, highlighting significant improvements in key areas. GPT-3.5 turbo demonstrates a 10.15% increase in accuracy, showing its superior precision over Bio-Mistral. Additionally, GPT-3.5 turbo reduces the

Table 1
Comparison between Bio-Mistral Quantized LLM and GPT-3.5 turbo. The improvement values indicate the percentage change in performance metrics.

Metric	Bio-Mistral Quantized LLM	GPT-3.5 turbo	GPT-3.5 turbo Improvement (%)
Accuracy	0.73	0.8	+10.15
Hallucination Rate	0.18	0.17	-4.76
Missing Rate	0.16	0.15	-3.43
AlignScore	0.8	0.84	+3.6
Semantic Similarity	0.74	0.82	+10.01
AI-generated errors	0.22	0.17	-20.17

Table 2
Relevance Scores for Different Retrieval Strategies

Metric	Bio-Mistral (quantized)	GPT 3.5 turbo (non-quantized)
Average Relevance Score	7.6/10	9.0/10

hallucination rate by 4.76%, indicating fewer instances of incorrect information generation. The missing rate, which refers to the omission of relevant content, is lowered by 3.43%. In terms of alignment, GPT-3.5 turbo achieves a 3.6% improvement in AlignScore, suggesting better consistency with expected outputs. The model also enhances semantic understanding, as reflected by a 10.01% improvement in semantic similarity. Lastly, GPT-3.5 turbo significantly reduces AI-generated errors by 20.17%, further underscoring its enhanced reliability compared to Bio-Mistral Quantized LLM; refer to table 1.

4.3. Relevance of Generated Outputs

The relevance of the outputs generated by GPT-3.5 turbo (non-quantized) significantly outperformed those produced by the Bio-Mistral (quantized) model; refer to table 2. Healthcare professionals rated GPT-3.5 turbo's relevance score at an average of 9.0 out of 10, compared to Bio-Mistral's 7.6 out of 10. This indicates that GPT-3.5 turbo's diagnostic suggestions were more closely aligned with clinical best practices and provided more actionable insights for patient care, especially in responding to symptoms described by users, showcasing the superiority of GPT-3.5 turbo in generating contextually relevant outputs.

5. Conclusion

The experimental results presented in this paper demonstrate the effectiveness of a Retrieval-Augmented Generation (RAG) pipeline integrated with Large Language Models (LLMs) for cancer prediction and classification, specifically for Gastrointestinal (GI) cancers. The hybrid retrieval approach significantly enhances the diagnostic accuracy and relevance of the chatbot, outperforming both sparse and dense retrieval methods used individually. By leveraging comprehensive text-based medical data, the system improves the alignment of generated diagnostic insights with clinical best practices, as shown by the substantial improvements in metrics such as accuracy, semantic similarity, and AI-generated errors when using GPT-3.5 turbo compared to the Bio-Mistral Quantized LLM.

Moreover, the relevance of the generated outputs was rated significantly higher when utilizing GPT-3.5 turbo, further emphasizing the value of integrating advanced LLMs in medical diagnosis. These improvements highlight the chatbot's potential as a reliable decision-support tool for healthcare professionals, capable of synthesizing patient symptoms with medical data to offer actionable insights.

Future work will focus on refining the retrieval mechanisms and incorporating real-time data integration, with the aim of developing a fully integrated decision-support system to further support cancer diagnosis and treatment. The promising results from this study underscore the importance of

combining advanced retrieval techniques with powerful language models to improve the accuracy and clinical relevance of diagnostic tools in the healthcare domain.

6. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Grammarly for grammar and spelling checks, as well as for paraphrasing and rewording. The author(s) also experimented with GPT-3.5 and a quantized version of BioMistral as part of the model development process. All content was reviewed and edited by the author(s), who take full responsibility for the final content of this publication.

References

[1] Q. Zhou, C. Liu, Y. Duan, K. Sun, Y. Li, H. Kan, Z. Gu, J. Shu, J. Hu, Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation, Frontiers in Medicine 11 (2024) 1392555. doi:10.3389/fmed.2024.1392555.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023). doi:10.48550/arXiv.2303.18223.

[3] M. Jeong, J. Sohn, M. Sung, J. Kang, Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, arXiv preprint arXiv:2401.15269 (2024). doi:10.48550/arXiv.2401.15269.

[4] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al., Natural questions: a benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466. doi:10.1162/tacl_a_00276.

[5] A. R. Pelletier, J. Ramirez, I. Adam, S. Sankar, Y. Yan, D. Wang, D. Steinecke, W. Wang, P. Ping, Explainable biomedical hypothesis generation via retrieval augmented generation enabled large language models, arXiv preprint arXiv:2407.12888 (2024). doi:10.48550/arXiv.2407.12888.