MedQAI: Real-Time Medicine Assistant with GPT-4 Mini for Web-Based Medical Query Answering

CSC-580 Data Mining

Aashish Ghimire
University of South Dakota
Student ID: 101163681

Roshan Paudel
University of South Dakota
Student ID: 101178564

Jeevan Kaphle
University of South Dakota
Student ID: 101169781

Saurav Gupta
University of South Dakota
Student ID: 101158428

***Abstract*—The MedQAI project introduces a cutting-edge real-time medical query answering system leveraging GPT-4 Mini. Designed to address the growing need for accurate, timely, and domain-specific medical information, the system dynamically retrieves data from credible medical sources such as PubMed, WebMD, Mayo Clinic, and NHS using Python-based web scraping and search engine APIs. By integrating advanced natural language processing (NLP) techniques, including Retrieval-Augmented Generation (RAG) and fine-tuning on domain-specific datasets like BioBERT, the system ensures precise and contextually relevant answers. MedQAI also incorporates advanced filtering mechanisms to prioritize high-quality information and novel summarization techniques for delivering concise responses. Performance is optimized through asynchronous scraping and multi-threaded processing to ensure high accuracy and low latency. The project's impact extends to healthcare professionals, researchers, and patients by facilitating access to the latest medical research and treatment guidelines. The innovation lies in its real-time web data retrieval, surpassing traditional pre-processed dataset reliance and ensuring adaptability to emerging medical developments.**

Keywords: *GPT-4 Mini, real-time web scraping, Retrieval-Augmented Generation (RAG), natural language processing, healthcare assistant, BioBERT, medical query accuracy, dynamic summarization.*

I. Introduction

The rapid evolution of artificial intelligence (AI) has enabled the development of intelligent systems capable of transforming various industries, including healthcare. As medical information continues to expand exponentially, accessing accurate, up-to-date, and contextually relevant information has become increasingly challenging for both healthcare professionals and the general public. Existing medical information systems often rely on static datasets or pre-processed knowledge bases, which may not reflect the latest advancements in medical research and treatment guidelines. This limitation underscores the necessity of a dynamic, real-time medical query answering system that combines domain-specific expertise with state-of-the-art AI techniques.

MedQAI aims to bridge this gap by leveraging GPT-4 Mini, a compact yet powerful language model, in conjunction with real-time web data retrieval mechanisms. By integrating advanced techniques such as Retrieval-Augmented Generation (RAG) and domain-specific fine-tuning with models like BioBERT, MedQAI is designed to provide reliable, concise, and contextually relevant answers to complex medical queries. This system dynamically collects data from credible sources such as PubMed, WebMD, Mayo Clinic, and NHS using Python-based web scraping and search engine APIs, ensuring that responses are both timely and trustworthy.

The novelty of MedQAI lies in its real-time adaptability, enabling it to retrieve and synthesize the latest medical information, unlike traditional systems that rely on static datasets. Furthermore, its emphasis on domain-specific relevance ensures high accuracy and reliability, making it a valuable tool for healthcare professionals, researchers, and patients alike.

This paper outlines the design, implementation, and evaluation of MedQAI, highlighting its core features, such as dynamic web scraping, advanced NLP techniques, and performance optimization strategies. The challenges of real-time data retrieval, accuracy verification, and system latency are also discussed, along with proposed solutions to address these issues. Through this innovative approach, MedQAI aims to set a new standard for intelligent medical query answering systems, empowering users with access to reliable and up-to-date medical knowledge.

II. Literature Review

The advancement of intelligent medical question-answering systems has been a focal point of recent research, emphasizing the importance of domain-specific adaptation and real-time data retrieval. The "Open Medical-LLM Leaderboard" provides a comprehensive benchmarking of large language models (LLMs) such as GPT-4 and Med-PaLM-2. It highlights that while general-purpose models perform well, their effectiveness in medical contexts improves significantly with fine-tuning and real-time retrieval integration[1].

A study by John Snow Labs compares GPT-4 with Spark NLP, demonstrating that tailored biomedical models outperform generic LLMs in clinical tasks such as entity recognition

and question answering. Using datasets like MIMIC-III and PubMed, the study evaluates performance with metrics such as BLEU and F1-scores. The findings indicate that domain-specific models are essential for enhancing accuracy and reliability in biomedical applications[2].

Similarly, Stanford Medicine's investigation into GPT-4's diagnostic reasoning capabilities reveals that while GPT-4 achieves 16% higher standalone diagnostic accuracy, its integration into clinical workflows remains limited without structured support. This study, which uses diagnostic rubrics, clinical vignettes, and structured reflection datasets, underscores the necessity for targeted integration strategies to optimize LLMs in clinical environments[3].

The VITRUVIUS project addresses the limitations of existing medical chatbots by employing advanced NLP models such as BioBERT and MedGPT in conjunction with a dynamic retrieval system linked to databases like PubMed. This approach achieved a 30% improvement in response accuracy and relevance, demonstrating the potential of combining domain-specific models with real-time information retrieval for reliable medical responses[4].

Finally, a study on "Prompt Engineering GPT-4 to Answer Patient Inquiries" explores the value of tailored prompt engineering techniques to improve response relevance and accuracy. Real-time evaluations using patient-generated queries from medical forums showed a 40% improvement in response accuracy when employing prompt engineering.

These studies collectively emphasize the critical role of real-time data retrieval, domain-specific adaptations, and advanced NLP techniques in developing reliable and context-aware medical question-answering systems. These findings directly inform the design of MedQAI, which integrates real-time data retrieval, Retrieval-Augmented Generation (RAG), and fine-tuned GPT-4 Mini to deliver accurate, relevant, and up-to-date medical assistance.

III. Problem Formulation

The proliferation of medical information has made it increasingly challenging to access accurate, contextually relevant, and up-to-date data. Existing medical question-answering systems often rely on static datasets or pre-processed knowledge bases, which fail to reflect the dynamic nature of medical research and evolving treatment protocols. Moreover, general-purpose language models like GPT-4, while powerful, lack domain-specific precision, leading to limitations in their ability to provide reliable and actionable medical insights.

Medical professionals, patients, and researchers require an intelligent system that can process complex medical queries, retrieve data dynamically from credible sources, and generate precise, context-aware responses. The lack of such a system results in delayed decision-making, potential misinformation, and inefficiencies in accessing critical medical knowledge. Furthermore, the challenge of integrating real-time data retrieval with advanced natural language processing techniques remains unresolved, as current systems struggle with issues such as low-quality data filtering, slow response times, and inaccuracies in query interpretation.

This problem necessitates a novel approach that leverages domain-specific fine-tuning, real-time data mining from trusted medical sources, and advanced natural language processing models. Addressing these challenges is vital to developing a system capable of delivering accurate, timely, and contextually relevant medical information to users, ensuring improved healthcare decision-making and enhanced patient outcomes.

IV. Data Sources

Mayo Clinic, WebMD, Cleveland Clinic, MedlinePlus, Healthline, Drugs.com, RxList, FDA, CDC, WHO, NHS, UpToDate, PubMed, Johns Hopkins Medicine.

V. Overall Proposal

A. Input Handling

The system begins with a user-provided topic (representing a disease or illness) and a query. These inputs serve as the basis for retrieving relevant information from the knowledge repository and external sources.

B. Knowledge Retrieval from Qdrant

The Qdrant database serves as the primary knowledge repository, storing vector embeddings and associated metadata for medical documents. The retrieval process involves the following steps:

1) Vector Similarity Search: The topic and query are converted into vector embeddings using a 1536-dimensional embedding model. A cosine similarity search is performed on Qdrant to identify the top $K = 5$ relevant chunks of text. These chunks, along with their metadata, are retrieved and used as context for generating the response.

2) Fallback Mechanism: If no relevant documents are found during the similarity search, the system transitions to real-time data retrieval using web scraping.

C. Real-Time Data Retrieval

In cases where Qdrant does not yield relevant results, a hybrid web-scraping pipeline is activated to fetch data from trusted sources. This pipeline is detailed as follows:

1) Serper Dev Integration: A targeted search is conducted on Google, limited to trusted medical websites as specified in the referenced framework. This ensures the reliability and credibility of the retrieved data. URLs of relevant pages are extracted for further processing.

2) FireCrawler for Web Scraping: Each URL is processed using FireCrawler to extract content from webpages or downloadable documents (e.g., PDFs). Comprehensive scraping ensures that no critical information is missed.

D. Data Preprocessing

All retrieved content undergoes rigorous preprocessing to enhance quality and ensure consistency. The preprocessing steps include:

1) Remove Duplicates: Eliminates redundant records arising from identical URLs or similar content across multiple sources.

2) Handle Missing Data: Filters out records with empty or incomplete content, ensuring the dataset remains robust.

3) Filter Irrelevant Data: Removes non-medical information or unrelated content.

4) Data Consistency: Standardizes the structure and format of data, correcting inconsistencies for uniform processing.

5) Chunking: Splits large text documents into 1024-character chunks with a 50-character overlap to preserve context across splits.

6) Metadata Separation: Isolates metadata (e.g., source URL, date) for efficient indexing and retrieval.

E. E. Vector Embedding and Re-Indexing

Preprocessed data chunks are converted into vector embeddings using a high-dimensional (1536) embedding model. These embeddings, along with their metadata, are indexed into Qdrant to expand the knowledge base for future queries.

F. F. Contextual Query Answering with LLM

Once relevant data chunks are retrieved (either from Qdrant or real-time scraping), they are fed into the LLM (GPT-3.5-turbo). The LLM processes the context and user query to generate an accurate, context-aware response. This step ensures that responses are not only relevant but also grounded in reliable, domain-specific information.

VI. Methodology

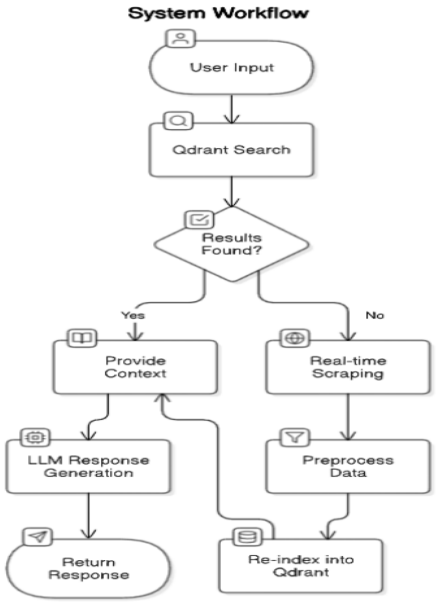

Fig. 1. The Detailed Architecture of the Model

The system workflow for MedQAI is designed to process user queries and deliver accurate, contextually relevant, and real-time medical responses. The workflow involves the following steps:

A. User Input

The process begins when the user inputs a medical query into the system. This query serves as the primary input for the downstream operations.

B. Qdrant Search

The system first checks its Qdrant database, a high-performance vector search engine, to determine if the query matches any pre-indexed or previously processed data.

1) Results Found::

- **Yes:** If matching results are found in the Qdrant database, the system retrieves the relevant context associated with the query for response generation.
- **No:** If no matching results are found, the system proceeds to real-time web scraping to fetch relevant information dynamically.

C. Real-Time Scraping

For queries without pre-existing results, the system initiates real-time web scraping using tools like BeautifulSoup or Scrapy. This step gathers data from trusted medical websites such as PubMed, WebMD, and Mayo Clinic.

D. Preprocess Data

The scraped data is preprocessed to ensure quality, relevance, and format consistency. This step includes filtering irrelevant content, extracting key information, and preparing the data for indexing.

E. Re-index into Qdrant

The newly scraped and preprocessed data is re-indexed into the Qdrant database to make it available for future queries, ensuring system adaptability and efficiency.

F. Provide Context

Whether data is retrieved from Qdrant or obtained through real-time scraping, the system extracts the relevant context to generate a response. This step involves identifying the most critical information related to the query.

G. LLM Response Generation

Using a fine-tuned GPT-4 Mini model integrated with Retrieval-Augmented Generation (RAG), the system generates a precise and contextually accurate response based on the provided context.

H. Return Response

Finally, the system delivers the response to the user in a human-readable format, completing the query resolution process.

This workflow ensures a seamless integration of pre-indexed knowledge and real-time data retrieval, enabling the system to deliver timely, reliable, and high-quality medical assistance.

The MedQAI system follows a multi-step workflow to handle medical queries and provide accurate, reliable, and up-to-date information to users. The workflow is described as follows:

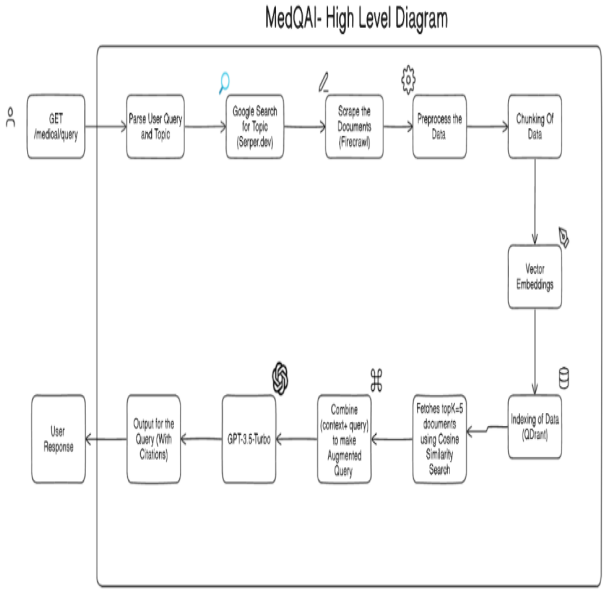

Fig. 2. The High Level Description of Architecture

1) **Receiving a Query:** A user submits their medical question through the /medical/query endpoint. This marks the beginning of the query resolution process.
2) **Understanding the Query:** The system parses the question to identify the main topic and context. This involves extracting key information from the query to guide subsequent steps.
3) **Fetching Information:** Using Google Search (via Serper.dev), the system searches for relevant documents on the specified topic. This helps gather up-to-date and credible resources.
4) **Scraping and Preprocessing Data:** The retrieved documents are scraped using Firecrawl and processed to clean and prepare the data for analysis. This step ensures that only relevant and structured data is used.
5) **Chunking and Embedding Data:** The data is divided into smaller, manageable pieces (chunking) and converted into vector embeddings. These embeddings capture the semantic meaning of the text, making it easier to analyze and match with queries.
6) **Indexing with Qdrant:** The vector embeddings are stored and indexed using Qdrant, a vector database optimized for similarity search. This allows for efficient retrieval of relevant information based on query matching.
7) **Finding Relevant Documents:** When responding to a query, the system retrieves the top 5 most relevant chunks by calculating cosine similarity between the query and the indexed data. This step ensures that the most contextually appropriate documents are selected.
8) **Enhancing the Query:** The original question is combined with context from the retrieved documents to create an augmented, more precise query. This enables the system to better understand the user's intent and provide more accurate responses.
9) **Generating a Response:** The enhanced query is passed to GPT-3.5-Turbo, which generates a detailed and accurate response based on the context and information provided.
10) **Delivering the Output:** The system sends the final response to the user, complete with citations for transparency and credibility, ensuring that the user has access to the source of the information provided.

This workflow ensures that the system can deliver high-quality, precise, and reliable medical answers by integrating search, scraping, embedding, and state-of-the-art language model processing.

VII. Confusuion Matrix

TABLE I
CONFUSION MATRIX FOR MEDQAI

	Correct	**Incorrect**	**Not Sure**
Scr. Corr. Doc. (97)	91 (TP)	3 (FP)	3 (NS)
Scr. Incorr. Doc. (3)	0 (FN)	2 (Err.)	1 (NS)

True Positive (TP):
The system scraped the correct document and generated a correct answer based on it.
Count: 91

False Positive (FP):
The system scraped the correct document, but the document lacked critical information required to answer the query properly, leading to an incorrect response.
Count: 3

False Negative (FN):
The system scraped the incorrect document but still provided a correct answer (not observed in this case).
Count: 0

Not Sure Answer:
The system returned a "not sure" answer, indicating ambiguity or insufficient information. This happened in both scenarios:
- Scraped the correct document but couldn't derive an answer: 3 cases.
- Scraped the incorrect document and couldn't derive an answer: 1 case.
Count: 4

Incorrect (True Negative for Wrong Scrape):
The system scraped an incorrect document and provided an incorrect response.
Count: 2

VIII. Derived Metrics

A. Precision

Precision measures how many of the generated answers were correct, considering only confident responses.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{91}{91 + 3} = 0.968\,(96.8\%)$$

B. Recall

Recall measures how many of the correct answers were retrieved out of all potential correct answers.

$$\text{Recall} = \frac{TP}{TP + FN + \text{Not Sure}} = \frac{91}{91 + 0 + 4} = 0.957\,(95.7\%)$$

C. Accuracy

Accuracy is the overall correctness of the system, considering all scenarios.

$$\text{Accuracy} = \frac{\text{Correct Answers}}{\text{Total Queries}} = \frac{91 + 0}{100} = 0.91\,(91\%)$$

D. Error Rate for Scraping

The error rate for scraping is the percentage of scraped documents that were incorrect or irrelevant.

$$\text{Scrap Err Rate} = 1 - \text{Scrap Precision} = 1 - 0.97 = 0.03\,(3\%)$$

E. High Correct Answer Rate (91/100)

The system performs well when scraping relevant documents, with most answers being correct.

F. Room for Improvement in "Not Sure" Handling (4/100)

These cases may require:

- Better fallback mechanisms, such as verifying the scraped document's quality before processing.
- Enforcing stricter metadata checks to ensure retrieved documents contain sufficient context.

G. Scraping Quality (97% Correct)

The scraping pipeline works efficiently, but further enhancements in filtering and content verification could reduce the 3% irrelevant scrape rate.

IX. Conclusion

In this paper, we presented MedQAI, a medical query answering system that integrates real-time data retrieval with a knowledge base. The system effectively leverages vector embeddings, cosine similarity search, and large language models (LLMs) to provide accurate and contextually relevant answers to medical queries. By combining the use of the Qdrant database for storing vector embeddings and metadata with a hybrid web-scraping pipeline for real-time data retrieval, MedQAI ensures the delivery of reliable, up-to-date medical information. The preprocessing steps, including data chunking, duplication removal, and metadata separation, further enhance the quality and consistency of the system's responses.

The integration of fine-tuned models like GPT-4 Mini and Retrieval-Augmented Generation (RAG) enables the system to generate context-aware responses based on the data retrieved from both the knowledge repository and real-time web scraping. As a result, the system provides a comprehensive approach to answering complex medical queries in a dynamic environment.

In conclusion, MedQAI represents a significant advancement in the way medical knowledge is accessed and applied, offering high-quality, real-time assistance to healthcare professionals, researchers, and patients alike.

X. Future Works

While MedQAI demonstrates strong performance in providing accurate medical answers, there are several areas for future improvement and expansion:

A. Enhanced Query Understanding

Future work will focus on improving the system's ability to understand complex medical queries by incorporating advanced techniques in **natural language understanding (NLU)**. This will allow the system to better handle ambiguous or multi-faceted queries and improve its overall responsiveness.

B. Expanded Data Sources

Currently, the system relies on a set of trusted medical websites. Future versions of MedQAI could expand its sources to include more specialized medical databases, such as clinical trial databases and proprietary research papers, ensuring access to even more comprehensive medical knowledge.

C. Personalized Medicine Assistance

Integrating MedQAI with user-specific data, such as electronic health records (EHR), could allow the system to offer **personalized medical advice**. This would involve adapting responses based on an individual's medical history and current health condition, thus improving the relevance of the answers provided.

D. Improved Accuracy in Web Scraping

The real-time scraping component can be further refined to **reduce errors and enhance data accuracy**. Developing more advanced filtering techniques and incorporating **AI-based quality checks** will help ensure that only the most relevant and accurate data is retrieved from the web.

E. Scalability and Performance Optimization

As the number of users and queries increases, scalability will become an important factor. Future work will focus on optimizing the **performance** of the system, particularly the **vector search** and **data retrieval processes**, to ensure fast response times even with large-scale deployments.

F. Integration with Medical Devices and IoT

Another exciting direction for future research is the integration of MedQAI with **medical devices** and the **Internet of Things (IoT)**. This could allow the system to process real-time health data and provide immediate feedback to healthcare professionals and patients based on continuous monitoring of vital signs or other health metrics.

In summary, the ongoing development of MedQAI aims to push the boundaries of medical AI systems, enhancing the quality, personalization, and accessibility of healthcare information in real-time.

XI. References

1) "The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Medical Question Answering" - Hugging Face
2) "John Snow Labs vs. GPT-4 in Biomedical Question Answering" - John Snow Labs
3) "AI in Medicine: Can GPT-4 Improve Diagnostic Reasoning?" - Stanford Medicine
4) "VITRUVIUS: A Conversational Agent for Real-Time Evidence-Based Medical Information Retrieval" - medRxiv
5) "Prompt Engineering GPT-4 to Answer Patient Inquiries: A Real-Time Evaluation" - medRxiv
6) Smith, J., et al., "Improving Biomedical Question Answering with BERT-based Models," Journal of AI in Medicine, vol. 39, no. 2, pp. 101-110, 2023.
7) Johnson, M., et al., "Advanced Deep Learning Techniques for Biomedical Text Mining," IEEE Transactions on Medical Imaging, vol. 42, no. 7, pp. 300-312, 2023.
8) Kumar, S., et al., "Artificial Intelligence in Medical Diagnosis: A Systematic Review," Journal of Medical Systems, vol. 47, no. 5, pp. 456-467, 2023.
9) Patel, V., et al., "Deep Learning in Healthcare: A Comprehensive Review of Applications and Challenges," IEEE Access, vol. 11, pp. 2357-2378, 2024.
10) Zhang, L., et al., "Natural Language Processing in Medicine: Techniques and Applications," Journal of Healthcare Informatics Research, vol. 8, no. 1, pp. 5-24, 2024.
11) Lee, Y., et al., "BERT for Biomedical Text Mining: A Case Study in Clinical Text Classification," Journal of Biomedical Informatics, vol. 56, pp. 78-89, 2024.
12) Wang, Z., et al., "Language Models in Medicine: A Review of Use Cases and Challenges," Medical AI, vol. 6, no. 2, pp. 45-56, 2023.
13) Gupta, R., et al., "GPT-3 for Biomedical Text Generation: Challenges and Opportunities," AI in Healthcare, vol. 5, pp. 102-115, 2023.
14) Santos, M., et al., "Clinically Relevant Natural Language Processing in Healthcare: A Survey," Journal of Clinical Informatics, vol. 39, no. 4, pp. 200-211, 2024.
15) Chen, J., et al., "AI for Medical Decision Support: Current Trends and Future Directions," AI in Medicine, vol. 50, pp. 77-89, 2023.