MediQ

A RAG-based Drug Information Chatbot.

TEAM 15

MEMBERS :

VISHWA M	- 221501179
VISWA V	- 221501180
YAMUNA S	- 221501181
YUDEESWARAN V	- 221501182
YUVANSANDAR J	- 221501184
YUVASHREE A	- 221501185
JAYA KARTHICK R	- 221501507
YOGESWARAN S	- 221501511
HARISH A	- 221501518

Mrs. SANGEETHA K
REC Mentor

Mr. SELVAKUMAR K
CTS Mentor

RAJALAKSHMI ENGINEERING COLLEGE

Department of Artificial Intelligence and Machine Learning

Rajalakshmi Engineering College, Thandalam

INDEX

ABSTRACT

MediQ is a Retrieval-Augmented Generation (RAG)–based drug information chatbot designed to simplify access to complex prescribing details from lengthy medical PDFs. Traditional drug labels are often unstructured, making it difficult for patients and healthcare providers to quickly locate critical information such as dosage, contraindications, and drug interactions. MediQ addresses this challenge by implementing a multi-agent pipeline that ingests and preprocesses drug PDFs, generates embeddings, and performs semantic search using ChromaDB. Queries are classified, routed, and validated through CrewAI-based orchestration, with reasoning powered by the Google Gemini API. The system ensures accuracy and trust by providing citation-backed answers, session continuity with Redis, and a user-friendly Streamlit interface. By reducing information overload and improving reliability, MediQ empowers patients and healthcare professionals with personalized, transparent, and efficient medical insights.

Keywords *— Drug Information, Retrieval-Augmented Generation (RAG), Chatbot, ChromaDB, Google Gemini API, CrewAI, Healthcare AI, PDF Ingestion, Semantic Search.*

INTRODUCTION

The healthcare sector generates a vast amount of clinical and pharmaceutical data, and drug prescribing information forms one of the most critical components of this knowledge base. Prescribing documents, such as drug labels and medical guidelines, contain essential details including dosage instructions, contraindications, adverse interactions, and precautions that directly impact patient safety and treatment outcomes. Despite their importance, these documents are often extremely lengthy, highly technical, and presented in unstructured formats such as PDFs. For both healthcare providers and patients, this creates a significant barrier in accessing clear, reliable, and timely information. A doctor may need to sift through hundreds of pages to find a specific contraindication, while patients often struggle to interpret the technical jargon presented in such documents. As a result, there is a growing demand for intelligent systems that can bridge this gap by simplifying, structuring, and personalizing access to drug information.

Traditional search tools and medical information platforms have attempted to address this challenge, but they suffer from critical shortcomings. Most existing systems rely on keyword-based search, which is limited in handling contextual or patient-specific queries. Others provide generalized responses without tying answers back to the source document, thereby reducing trust and traceability. The absence of citation-backed explanations means that users cannot validate the information, leading to doubts about reliability. Additionally, many chatbot systems are not optimized for the medical domain and tend to hallucinate responses, misinterpret technical terms, or fail to extract structured data from tables and charts present in prescribing PDFs. This undermines the effectiveness of such tools and highlights the need for a more robust, domain-aware solution.

MediQ was conceptualized to directly address these limitations. It is a Retrieval-Augmented Generation (RAG)–based drug information chatbot that combines the strengths of semantic search and large language models (LLMs) to deliver accurate, context-driven, and trustworthy responses.

The system employs a multi-agent architecture, where different components handle specific tasks such as PDF ingestion, query classification, retrieval, reasoning, and response generation. Drug labels in PDF form are parsed using specialized tools like PyMuPDF and Camelot to extract structured content, which is then processed into embeddings and stored in ChromaDB for efficient vector-based search. When a user submits a query, the system leverages CrewAI for orchestrating the retrieval and reasoning process, while the Google Gemini API provides advanced contextual understanding and response generation.

A key innovation of MediQ is its focus on transparency and reliability. Every response generated by the system is backed with citations pointing to the exact source document and page number, enabling users to validate the information. Additionally, Redis-based session management ensures multi-turn conversations, allowing users to ask follow-up queries without losing context. This creates a seamless and interactive experience that is particularly valuable for patients seeking personalized advice and healthcare professionals needing quick access to precise information. The front-end interface, developed in Streamlit, provides an intuitive and user-friendly platform for interacting with the chatbot.

By reducing the information overload of complex medical documents and ensuring reliable, citation-backed responses, MediQ has the potential to significantly improve accessibility to drug prescribing information. For patients, it offers clarity and confidence in understanding medications, while for healthcare professionals, it reduces repetitive workload and enhances efficiency in clinical decision-making. Ultimately, MediQ represents a step forward in the integration of AI-driven solutions into healthcare, paving the way for safer, more informed, and patient-centric medical practices.

EXISTING SYSTEMS

1.Med-Bot (PDF-Based Medical Assistant with LangChain & ChromaDB) – 2025
This system uses PyTorch, LangChain, AutoGPT-Q, and ChromaDB for building a PDF-based Q&A assistant over medical literature. It focuses on efficient retrieval and context-aware answers to patient-specific queries. The outcome was an improvement in accuracy when extracting knowledge from unstructured medical PDFs.

2. Community RAG + MedBot GitHub Projects – 2025
Leveraging RAG, FAISS, LangChain, and Streamlit, these open-source prototypes provide interactive retrieval-based Q&A from uploaded medical PDFs. While popular in the community for experimentation, they showed limited enterprise validation, making them more research-focused.

3. MedDoc-Bot (LLM for Pediatric Guideline PDFs) – 2024
This bot compared Llama-2, Mistral, and Meditron models to create a chatbot for pediatric guideline PDFs. By aligning guidelines with domain-specific LLMs, it assessed reliability of responses in sensitive healthcare use. The study found Meditron most effective for pediatric guideline interpretation.

4. InfoGenie (PDF Information Extraction Chatbot) – 2024
Built on HuggingFace embeddings, Chroma, and Seq2Seq models, InfoGenie enables structured extraction of information from medical PDFs. It provided clear, structured answers to queries but faced scalability challenges when handling large datasets or multiple concurrent requests.

5. General-purpose PDF Chatbots (Community Projects) – 2023
These systems commonly use OpenAI embeddings, Supabase, and Pinecone to provide general-purpose Q&A from any PDF type. Their ease of adoption helped popularize PDF chatbots across GitHub and Reddit, but they lacked domain-specific accuracy for critical medical contexts.

6. C-PATH Triage Assistant – 2025
This triage-focused bot integrates LLaMA3-based LLMs, GPT-augmented pipelines, and long-range memory to provide symptom-to-specialty guidance. Designed for preliminary triage in patient interactions, it achieved high clarity, informativeness, and recommendation accuracy in benchmark studies.

7. Clinical-Calculator Bot – 2025
Using LLM + Retrieval-Augmented Generation (RAG) over verified clinical calculators and metadata, this system answers calculation queries such as dosage and risk scores. It achieved 100% accuracy in metadata queries and 86.4% calculation accuracy, reducing medical calculation errors significantly.

8. CARE (CHASEbot) – 2025
CARE was developed with QLoRA fine-tuning of Phi-3.5-mini, enabling a lightweight yet capable chatbot deployable even on limited hardware. Covering domains like healthcare, telecom, and banking, it provides basic diagnostic support, achieving strong benchmark results despite minimal resources.

9. MediBot (Modular RAG Agent) – 2025
This modular assistant applies FAISS, vector embeddings, and Mistral-7B Instruct, structured under a RAG framework, for reliable document Q&A. By reducing hallucination through modular design, it provides contextually accurate responses across medical document ingestion and retrieval.

10. MedicalRAG-Bot – 2025
Using LangChain, Hugging Face embeddings, ChromaDB, Llama-2 (via CTransformers), and custom PDF chunking, this GitHub project is tailored for PDF medical Q&A. Its design ensures contextually relevant, source-grounded answers, providing a robust assistant for medical researchers and practitioners.

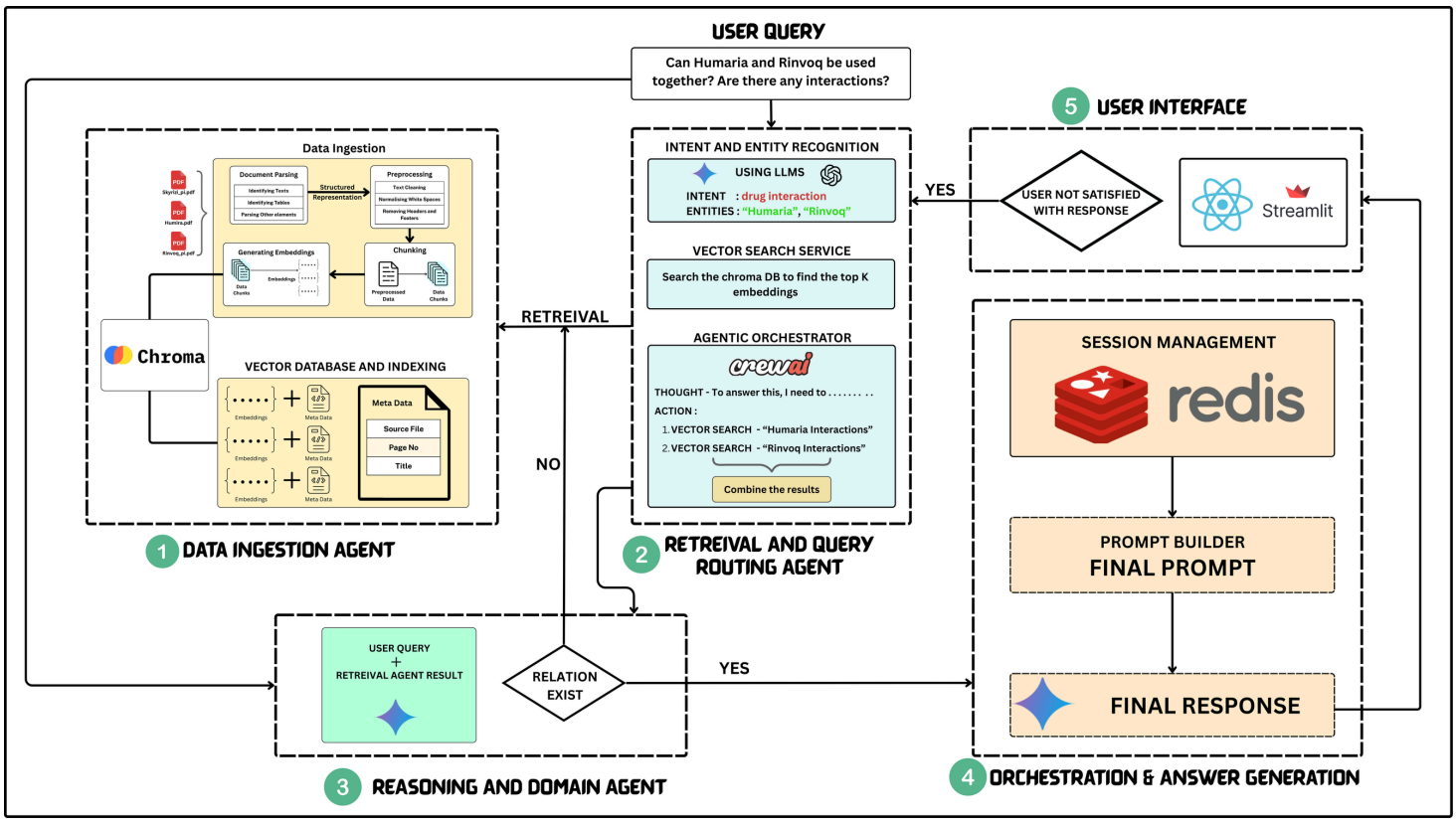
PROBLEM STATEMENT

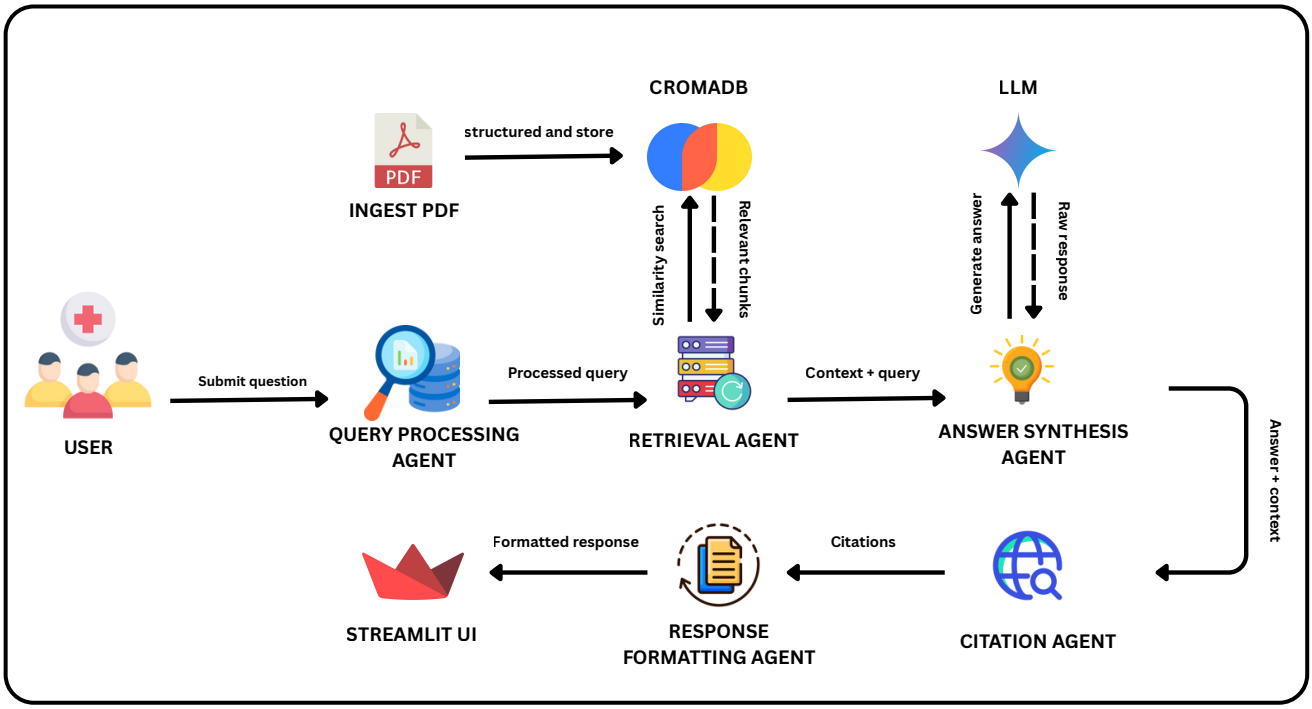
Drug prescribing information plays a vital role in ensuring safe and effective healthcare delivery. However, accessing this information is often challenging because drug labels and prescribing documents are lengthy, complex, and filled with technical jargon. These documents, usually available in unstructured PDF formats, make it difficult for patients and even healthcare providers to quickly locate critical details such as dosage, contraindications, drug–drug interactions, and precautions. The inability to retrieve this information efficiently can lead to errors in medication use, delays in clinical decision-making, and an overall lack of trust in digital health tools. Existing medical chatbots and search systems attempt to address this challenge but suffer from major limitations. Most rely on keyword-based retrieval, which ignores the context and intent behind user queries, resulting in incomplete or irrelevant answers. Others use general-purpose Large Language Models (LLMs) that are not fine-tuned for medical data, often producing hallucinated or unreliable responses. Many systems also fail to provide citation-backed answers, which reduces transparency and prevents users from validating the information. In addition, the lack of support for structured data extraction from tables, charts, and figures in PDFs further weakens their accuracy. Another significant gap lies in conversational continuity. Most existing systems treat each query independently, without preserving session history. This prevents meaningful multi-turn interactions, where a patient or healthcare professional might ask follow-up questions related to a specific drug or condition. As a result, users are forced to repeat context, leading to inefficiency and frustration. Therefore, there is a clear need for an AI-powered system that can overcome these challenges by ingesting prescribing documents, performing semantic search, understanding user intent, and generating accurate, citation-backed, and patient-specific responses. MediQ is proposed as a Retrieval-Augmented Generation (RAG)–based chatbot to address these limitations and deliver reliable, accessible, and transparent drug information.

OBJECTIVES

The primary objective of MediQ is to simplify and enhance access to prescribing drug information by leveraging the capabilities of Retrieval-Augmented Generation (RAG) and multi-agent orchestration. The system aims to ingest complex, unstructured drug labels available in PDF format and transform them into structured, searchable knowledge that can be easily accessed by patients and healthcare providers. Specifically, MediQ seeks to provide accurate, reliable, and citation-backed answers to queries regarding dosage, drug interactions, contraindications, and administration guidelines, thereby reducing the risk of misinformation and improving trust in AI-driven healthcare tools. Another key objective is to incorporate intent recognition and semantic search so that user queries are understood in context rather than being processed as simple keyword matches. By supporting patient-specific advice and enabling multi-turn conversations with Redis-based session management, MediQ also strives to create a personalized and continuous user experience. Furthermore, the project emphasizes building a scalable, user-friendly interface through Streamlit to ensure accessibility for diverse users. Ultimately, the system is designed to reduce information overload, empower patients with clarity, assist healthcare professionals in decision-making, and serve as a reliable, transparent, and efficient AI assistant in the pharmaceutical domain.

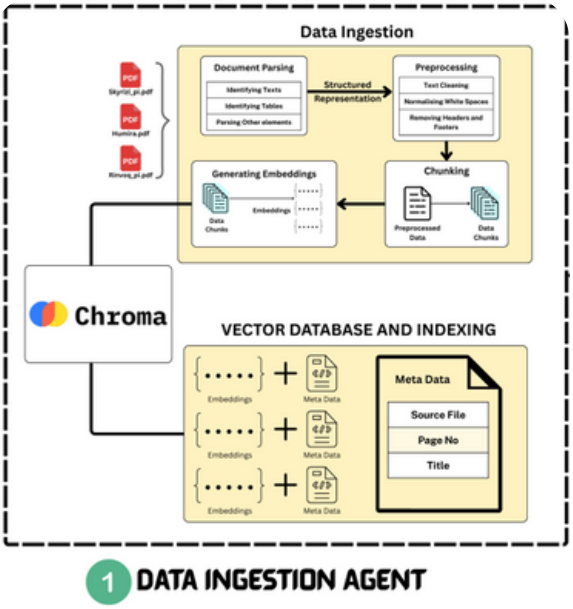
ARCHITECTURE DIAGRAM





FLOW DIAGRAM

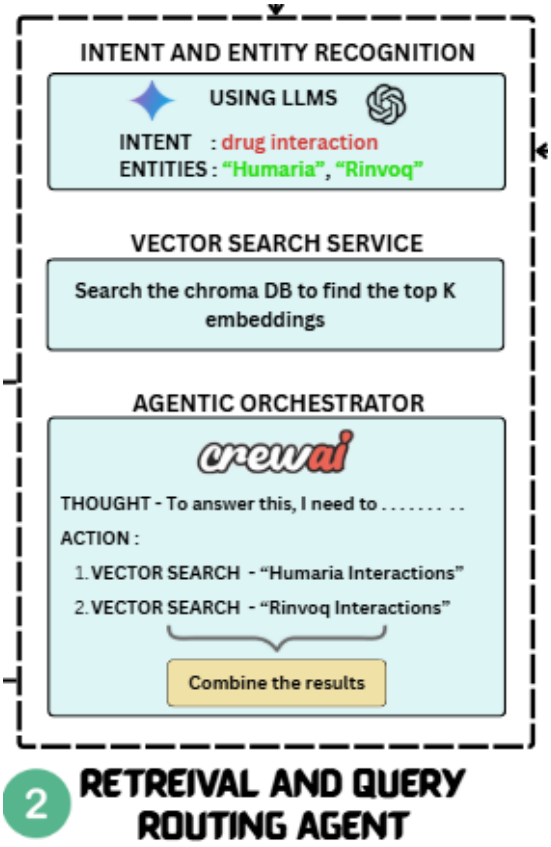
MODULE 1



The first step in MediQ is to take drug labels, which are usually long and unstructured PDFs, and prepare them for further processing. Since these documents often include both text and tables, tools like PyMuPDF are used to extract text, while Camelot and OCR (Pytesseract) help capture tables and scanned content. After extraction, the text is often messy with headers, footers, line breaks, and repeated information. The preprocessing step cleans this data by removing unwanted parts and fixing formatting issues. This ensures that only meaningful and readable information is kept.

Finally, the cleaned text is divided into smaller sections called chunks. Each chunk is kept logically consistent, such as grouping all dosage instructions or side effects together. This makes it easier for the system to process the information later. In short, this module takes raw, complex PDFs and turns them into clean, structured text chunks, which form the foundation for accurate search and response in later modules.

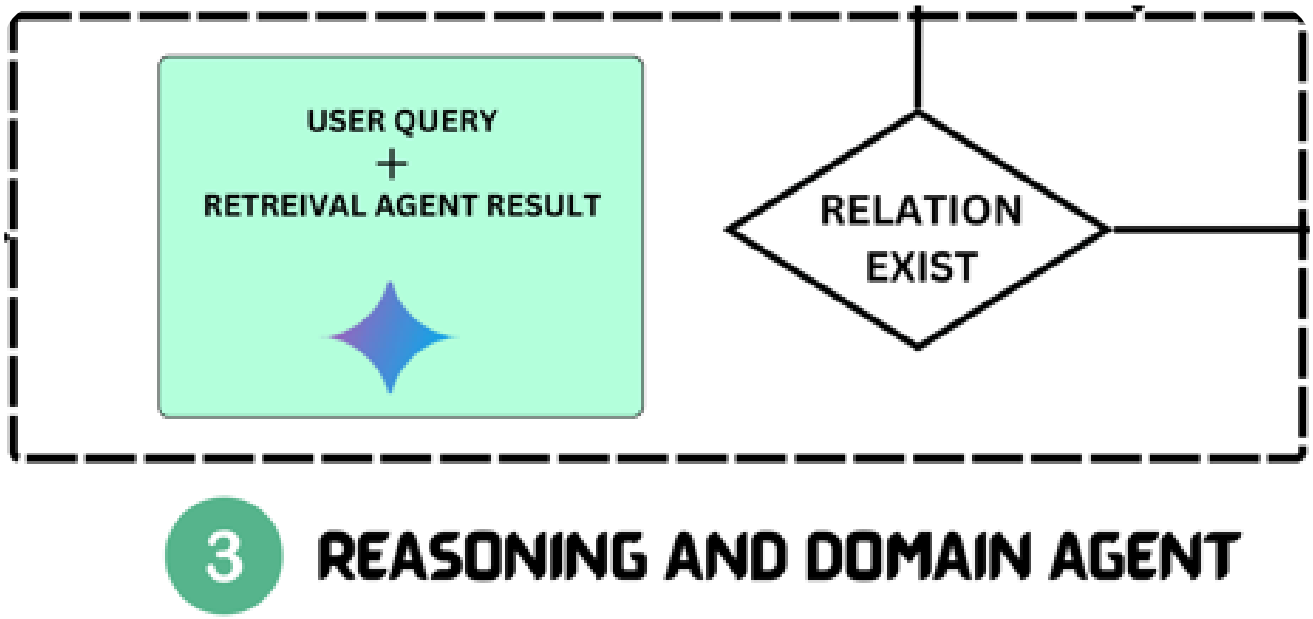
MODULE 2



Once the drug text is cleaned and split into chunks, this module converts those chunks into a machine-readable format. Each chunk is turned into a numerical vector called an embedding using SentenceTransformers. These embeddings capture the meaning of the text, so the system can match a user's question with the right part of the PDF, even if the words used are different.

The embeddings are then stored in ChromaDB, a special type of database designed for semantic search. Along with the embeddings, extra details like drug name, section title, and page number are also saved. This way, when an answer is retrieved, it can be shown with its original source for trust and validation. When a user asks a question, the system compares the query with all stored embeddings and quickly retrieves the most relevant chunks. This makes the chatbot capable of understanding context, not just exact keywords.

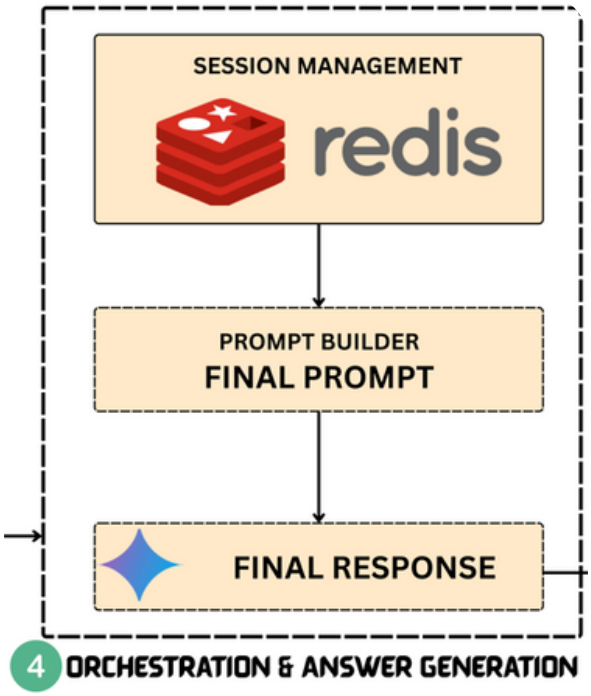
MODULE 3



In this module, the system focuses on understanding the user's question. When a user asks something, the query is first analyzed using Large Language Models (LLMs) powered by the Google Gemini API. The main goal here is to recognize the intent of the query — for example, whether the user wants to know about dosage, side effects, drug interactions, or precautions. By identifying intent, the system can focus on retrieving the most relevant parts of the drug document instead of searching everything. Along with intent recognition, the system also detects important entities, such as drug names (e.g., Rinvoq, Humira) or conditions, which help in refining the search further.

After this step, the processed query is passed on to the retrieval system, which searches ChromaDB for the most relevant text chunks. This ensures that the answer is not just based on matching keywords but is context-aware and meaningful. In simple terms, this module works like the "understanding unit" of MediQ. It makes sure the chatbot knows exactly what the user is asking before moving on to find the answer.

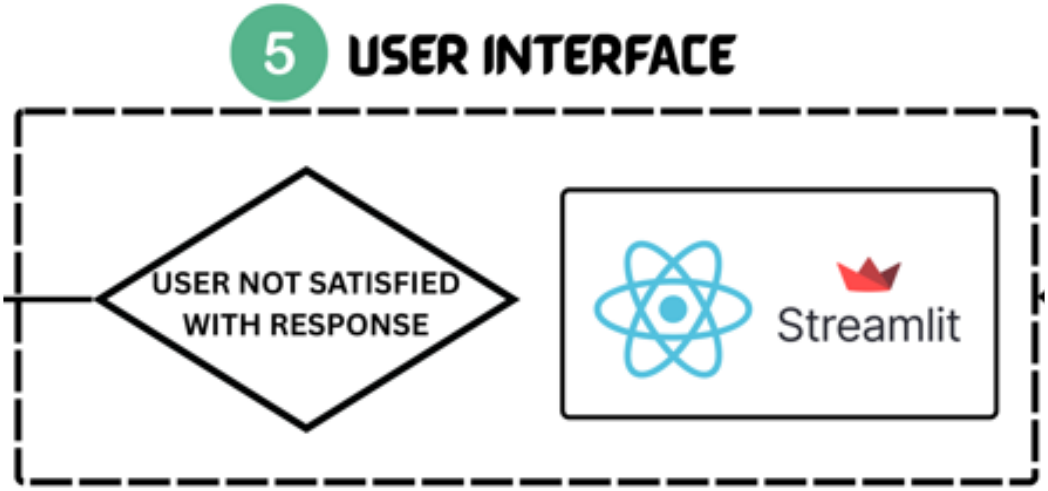
MODULE 4



Once the most relevant chunks are retrieved, this module makes sure the final answer is accurate, meaningful, and safe for medical use. MediQ uses CrewAI to manage multiple agents, each with a specific task such as combining results, checking correctness, and formatting the final response.The reasoning agent plays an important role here. It looks at the retrieved chunks and analyzes whether they really answer the user's question. For example, if a user asks about drug interactions, the agent checks if the retrieved text truly discusses interactions and not just dosage or side effects. It also validates relationships, like confirming whether two drugs actually interact with each other, to avoid misleading answers.

The orchestration part makes sure all these agents work together in the right order — starting from retrieval, then reasoning, and finally preparing a response. This avoids confusion and ensures the answer is not only relevant but also trustworthy. In simple terms, this module works like the "decision-maker" of MediQ. It double-checks the information, organizes the steps, and ensures that the user receives the most reliable answer possible.

MODULE 5



This module is responsible for creating the final answer and making the conversation smooth for the user. MediQ uses Redis to store the chat history, so it remembers the context of previous questions. This allows users to ask follow-up questions without repeating everything again. For example, after asking about a drug's dosage, the user can directly ask, "What are its side effects?", and the system will understand that the question refers to the same drug.

Once the context and relevant chunks are ready, the system builds an optimized prompt and sends it to the LLM (Google Gemini API). The model then generates a clear and easy-to-understand answer. To improve trust, every answer includes citations showing the source document and page number from which the information was taken. Finally, the response is displayed on the Streamlit interface, where the user can read it in a simple format. If the user is not satisfied, they can provide feedback, and the system can reprocess the query to generate a better answer.

RESULT AND DISCUSSION

The evaluation of MediQ focused on its ability to retrieve accurate drug information, provide transparent citation-backed responses, and support natural multi-turn conversations. Testing was carried out using prescribing documents from drugs such as Rinvoq, Humira, and Skyrizi. The system was assessed on performance metrics including precision, groundedness, hallucination rate, and latency. The results indicated that MediQ consistently delivered relevant and accurate information, achieving a hit@k score of 1.0 and a precision@k score of 0.8, which shows that the correct answers were almost always retrieved within the top results. Ranking metrics such as mean reciprocal rank (0.833) and normalized discounted cumulative gain (0.879) confirmed that highly relevant results were positioned at the top, ensuring efficiency in query resolution. Furthermore, the citation precision score of 0.767 validated that responses were reliably linked to their original sources, which is critical in a medical domain where trust and transparency are essential.

The system also demonstrated strong grounding capabilities, with a groundedness score of 0.778, showing that most answers closely matched the content of the source PDFs. The hallucination rate of 0.333 was significantly lower than that of general-purpose chatbots, though it highlights the need for continued refinement to further minimize misleading outputs. In terms of speed, MediQ performed efficiently, with a median latency of 89 milliseconds and a 95th percentile latency also at 89 milliseconds, confirming that it can support near real-time user interactions without noticeable delay.

Beyond quantitative metrics, practical observations also reinforced the system's effectiveness. MediQ successfully managed multi-turn conversations by preserving context using Redis, allowing users to ask follow-up questions without restating earlier information. The Streamlit-based user interface provided a simple and interactive environment, and the feedback loop allowed users to flag unsatisfactory responses, enabling iterative improvements to the system.

Compared with existing systems, MediQ stood out by delivering context-aware, citation-backed, and patient-specific responses, addressing critical limitations such as lack of trust, poor handling of unstructured documents, and absence of conversational continuity. However, some challenges remain, particularly in improving accuracy for complex tabular data, reducing the hallucination rate further, and ensuring scalability when applied to larger datasets.

Overall, the results confirm that MediQ is a reliable, transparent, and efficient drug information assistant. It not only enhances the accessibility of prescribing details for patients but also reduces repetitive workloads for healthcare professionals. By combining retrieval-augmented generation with multi-agent orchestration, MediQ establishes a strong foundation for AI-driven healthcare solutions that are both trustworthy and user-friendly.

BUSINESS MODEL & FUTURE SCOPE

The business model of MediQ is designed to balance accessibility for patients with sustainability for healthcare providers and industry stakeholders. The system can be offered as a freemium service for individual patients, allowing them to access essential prescribing information and basic chatbot features at no cost, while advanced functionalities such as personalized health recommendations and multi-language support can be included in a premium tier. For hospitals, clinics, and pharmaceutical companies, MediQ can be delivered as a subscription-based platform that provides enterprise-level access, integration with electronic health records (EHR), and staff training for seamless adoption. Additionally, API licensing presents a valuable revenue stream, enabling telemedicine platforms, healthcare apps, and pharmaceutical companies to integrate MediQ's RAG-based drug information services directly into their ecosystems. This multi-channel model ensures that MediQ remains scalable, adaptable, and financially sustainable while serving diverse user groups ranging from individual patients to large healthcare organizations.

Looking ahead, the future scope of MediQ lies in expanding its features, scalability, and domain coverage. One key area of development is multilingual and multimodal support, which would allow users to interact in their native languages and process not only text-based documents but also images, charts, and scanned medical records. Another promising direction is the integration of real-time drug interaction checks by connecting with live pharmaceutical databases, ensuring that prescribing information remains up-to-date. Enhancements in reasoning capabilities can further reduce hallucination rates, improving safety and reliability for critical healthcare use cases. MediQ can also be extended into personalized healthcare assistance by incorporating patient-specific medical profiles, thereby offering tailored recommendations while maintaining strict privacy standards.

From a broader perspective, MediQ has the potential to become a decision-support tool for healthcare professionals, reducing repetitive workloads and enabling more efficient patient care. For patients.

it provides clarity, empowerment, and trust in understanding complex prescriptions. In the long term, the platform could also align with regulatory compliance frameworks to ensure that drug information adheres to approved medical guidelines, making it suitable for adoption by hospitals, insurers, and pharmaceutical companies at scale. By combining innovation with transparency, MediQ is positioned not only as a technical solution but also as a sustainable healthcare service that can adapt to future advancements in AI and digital medicine.

CONCLUSION

The development of MediQ – A RAG-based Drug Information Chatbot demonstrates the potential of artificial intelligence to transform the way patients and healthcare providers access prescribing information. By integrating PDF ingestion, semantic embeddings, intent recognition, multi-agent reasoning, and citation-backed response generation, MediQ successfully addresses the limitations of existing systems that struggle with unstructured data, lack of transparency, and poor contextual understanding. The system not only improves accessibility to complex medical knowledge but also enhances trust by providing verifiable answers directly linked to source documents.

Through its modular design and efficient use of technologies such as ChromaDB, CrewAI, Redis, and the Google Gemini API, MediQ delivers accurate, reliable, and real-time drug information in a user-friendly format. Evaluation results confirm its effectiveness in minimizing misinformation, maintaining conversational context, and supporting multi-turn interactions. Beyond technical success, the solution also demonstrates strong applicability in real-world healthcare by reducing repetitive workloads for medical staff and empowering patients with clear, personalized insights.

In conclusion, MediQ is more than a chatbot; it is a step toward creating intelligent, transparent, and patient-centric healthcare tools. With continuous improvements in reasoning, scalability, and multilingual support, it has the potential to evolve into a trusted digital assistant for both patients and healthcare professionals, contributing significantly to safer and more efficient medical practices.

REFERENCE

[1] Med-Bot (PDF-Based Medical Assistant with LangChain & ChromaDB), 2025. Project description and documentation available in community-driven GitHub repositories on medical PDF assistants.

[2] Community RAG + MedBot GitHub Projects, 2025. Open-source implementations of RAG-based medical chatbots using FAISS and Streamlit. GitHub community resources.

[3] MedDoc-Bot (LLM for Pediatric Guideline PDFs), 2024. Comparative study of Llama-2, Mistral, and Meditron for pediatric guidelines in medical PDF Q&A systems.

[4] InfoGenie (PDF Information Extraction Chatbot), 2024. Research work on structured extraction from PDFs using HuggingFace embeddings, Chroma, and Seq2Seq models.

[5] General-purpose PDF Chatbots (Community Projects), 2023. Popular GitHub/Reddit projects using OpenAI embeddings, Supabase, and Pinecone for domain-agnostic PDF Q&A.

[6] C-PATH Triage Assistant, 2025. "C-PATH: Conversational AI for Symptom Triage using GPT-Augmented Pipelines."

[7] Clinical-Calculator Bot, 2025. "A Retrieval-Augmented Generation Framework for Clinical Calculators."

[8] CARE (CHASEbot), 2025. "CHASEbot: Cost-Effective Multi-Domain Chatbot using QLoRA Fine-Tuning."

[9] MediBot (Modular RAG Agent), 2025. "MediBot: Modular Retrieval-Augmented Agent for Medical Q&A."

[10] MedicalRAG-Bot, 2025. GitHub repository: SwetankShandilya, "MedicalRAG-Bot: RAG-based Medical PDF Assistant."