MediMate: An Agentic AI Framework Leveraging Multi-Agent Retrieval-Augmented Generation for Transformative Healthcare E-Commerce

**Nitin Gupta Mayur Dabade Mandar Kalse**

**Rohit Daund**

*Abstract*— The rapid evolution of digital healthcare demands systems that seamlessly integrate clinical precision with e-commerce efficiency. This paper introduces MediMate, an innovative Agentic AI-based healthcare chatbot designed to enhance the user experience for both healthcare professionals and patients. Leveraging multi-agent frameworks, advanced natural language processing, and vector search techniques, MediMate addresses the critical challenges of delivering accurate product information, enabling transparent price comparisons, and restricting order placement exclusively to authorized providers. The system’s architecture combines modular AI agents such as the router, validator, recommendation, general info, order, and summarization agents with robust back-end technologies (e.g., FastAPI, ChromaDB, and open-source LLMs). Extensive literature review and performance comparisons reveal significant improvements in accuracy, cost reduction, and efficiency. The proposed framework not only streamlines information retrieval and decision-making processes in healthcare e-commerce but also sets the stage for scalable integration with electronic health records (EHRs) and personalized clinical insights. This work outlines design decisions, system architecture, and future enhancements aimed at fostering reliable and secure digital healthcare interactions.

Keywords— Agentic AI, Healthcare Chatbot, E-Commerce, Clinical Decision Support, Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Natural Language Processing, Vector Search, Electronic Health Records (EHR), Data Privacy, Agile Development, Open-Source LLMs

# Introduction

The intersection of healthcare and e-commerce has grown exponentially, introducing unique challenges that demand precise and efficient solutions. Traditional healthcare e-commerce platforms often struggle to provide real-time, reliable product information while meeting stringent requirements for data privacy and regulatory compliance. In response to these challenges, this paper presents MediMate, an advanced agentic AI framework developed to revolutionize the healthcare shopping experience and clinical decision support.

MediMate is designed to bridge the gap between the rapid, consumer-oriented demands of e-commerce and the intricate, detail-sensitive needs of clinical environments. Healthcare professionals require immediate access to up-to-date, clinically validated product information, while patients seek user-friendly recommendations tailored to their individual health profiles. To address these diverging needs, MediMate employs a modular multi-agent architecture. This approach leverages specialized agents—each programmed for tasks such as routing queries, validating clinical data, generating recommendations, summarizing content, and processing secure orders. The system is built on a robust technological stack that includes advanced natural language processing algorithms, vector-based search using ChromaDB, and scalable backend services powered by FastAPI and Python.

Central to the innovation behind MediMate is its use of Retrieval-Augmented Generation (RAG) techniques, which combine information retrieval with generative AI to produce responses that are both contextually rich and precise. By integrating open-source large language models (LLMs) such as Llama-3 and employing Hugging Face embeddings, the system achieves enhanced efficiency and scalability. The framework also integrates cutting-edge technologies like LangChain for multi-agent coordination and PyPDF2 for document processing, ensuring seamless interoperability and real-time feedback mechanisms.

The development process is underpinned by agile methodologies, involving continuous iterations and regular mentor feedback. Early prototypes were developed using Streamlit for a rapid proof-of-concept, which informed subsequent enhancements focused on secure data handling and improved user-specific interactions. The project not only aims to empower healthcare providers with timely clinical insights but also ensures that product ordering remains a privilege exclusive to verified medical professionals, thus maintaining high standards of data security and privacy compliance (e.g., HIPAA).

Through comprehensive literature reviews and comparative analyses with other RAG approaches, the study positions MediMate as a state-of-the-art solution that demonstrates superior performance metrics in accuracy, cost reduction, and operational efficiency. This paper provides a detailed overview of the system architecture, including flow charts and multi-agent collaboration diagrams, while also discussing future prospects such as multilingual support, enhanced personalization, and further integration with EHR systems. The overarching goal is to set a benchmark for future research in agentic AI systems within the healthcare domain, ensuring that technology remains a catalyst for improved patient outcomes and streamlined clinical processes.

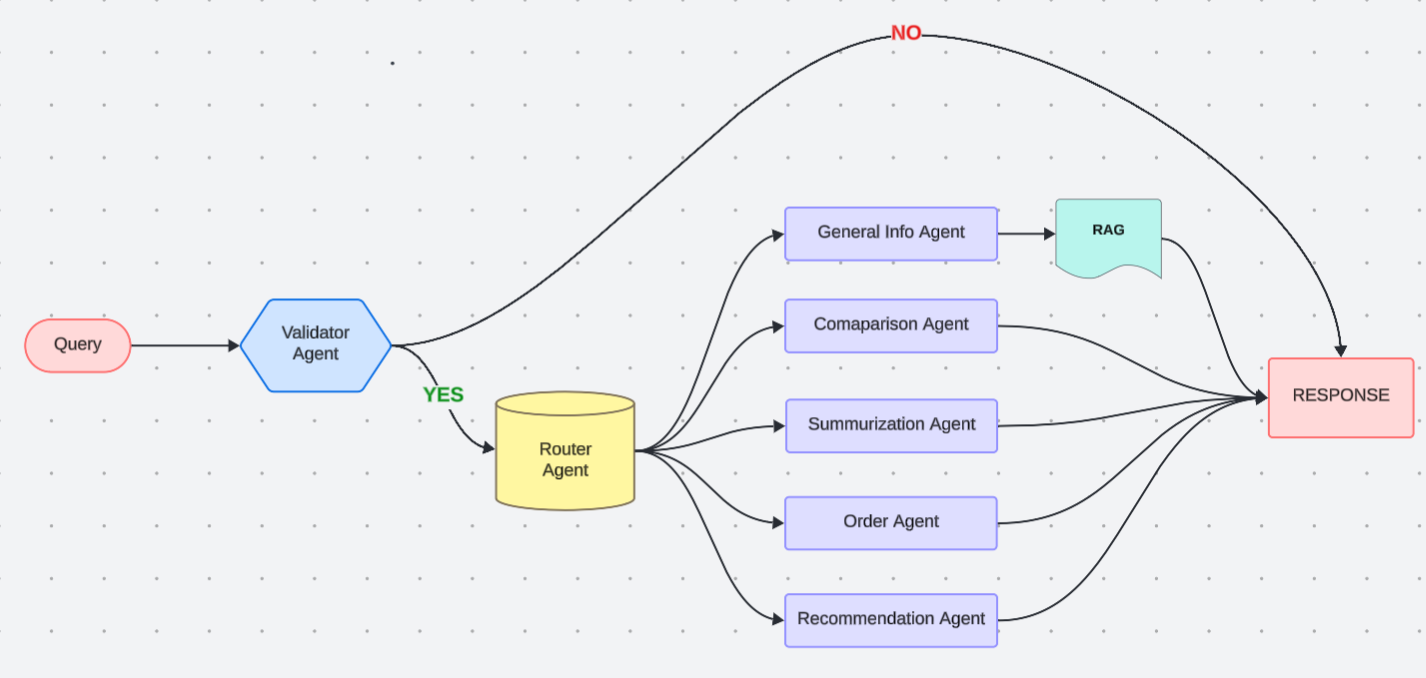
# Literature review

The integration of artificial intelligence (AI) into healthcare has spurred extensive research aimed at developing intelligent chatbots capable of assisting with medical diagnosis, patient interaction, and personalized treatment recommendations. The literature reviewed here spans a broad spectrum—from traditional natural language processing (NLP) methods to advanced retrieval-augmented generation (RAG) frameworks and multi-agent systems—each contributing unique insights toward creating robust healthcare solutions.  
This paper presents a healthcare chatbot that leverages traditional NLP methods—such as n-gram models, TF-IDF, and cosine similarity—to reduce the dependency on doctors for minor health issues. By extracting and ranking keywords and then comparing user queries against a robust SQL-based RDBMS knowledge base, the system delivers rapid and accurate automated responses. The approach highlights the practical benefits of combining statistical text models with structured databases to support efficient healthcare information retrieval and improve patient self-care [1].  
Focusing on rule-based conversational systems, this study develops a healthcare chatbot using AIML to create structured conversational rules and response templates, deployed on the Pandorabots platform. The integration of external APIs—for instance, from the CDC, PubMed, and Scopus—enables real-time retrieval of medical data, while summarization tools like SMMRY provide custom summaries for user queries. The architecture, built with a microservices approach and adhering to HL7 FHIR standards, ensures scalability, interoperability, and robust security measures such as data anonymization and encrypted communications [2].This work introduces a specialized chatbot that uses Retrieval-Augmented Generation (RAG) to support personalized treatment planning for Multiple Myeloma. By integrating domain-specific models like BioMed-RoBERTa-base and Mistral-7B with search and indexing solutions such as Amazon OpenSearch and LangChain, the system retrieves and synthesizes MM-specific literature and genomic data. The method simplifies complex genomic interpretations, delivering patient-specific recommendations that underscore the potential of RAG in precision medicine and clinical decision support [3].  
In an effort to enhance the reliability of medical large language models, this paper proposes MedGraphRAG—a novel graph-based framework. By employing techniques like triple graph construction and U-Retrieval alongside advanced models such as GPT-4 and LLaMA, and leveraging medical datasets like MIMIC-IV and PubHealth, the framework organizes complex medical information into structured graphs. This organization allows for evidence-backed responses and improved knowledge retrieval, thereby bolstering both accuracy and trustworthiness in clinical applications [4].This study presents a multi-agent workflow that employs GPT-40 in conjunction with a Reflexion framework to generate patient-friendly medical reports with minimal manual intervention. By integrating tools such as ICD-10 verification, Flesch-Kincaid readability scoring, and the AlfWorld module for multi-agent decision-making, the system achieves high accuracy and streamlines report generation. The approach significantly reduces the need for human editing, ensuring that the generated reports are both clinically precise and easily understandable for patients [5].This paper explores the concept of vertical AI agents—domain-specific modules that integrate dynamic API calls and RAG systems to automate complex workflows across industries, including healthcare. By focusing on context-aware and specialized decision-making capabilities, the study outlines how these agents can improve operational efficiency and decision quality. The discussion includes real-world scenarios where such vertical agents deliver measurable improvements in task automation, accuracy, and overall system performance, setting a blueprint for future implementations in various sectors [6].In this research, a simulation framework is developed that integrates Electronic Health Records (EHRs) with large language models through a multi-agent, RAG-based approach. Utilizing Neo4j to build a patient knowledge graph from the MIMIC-III database, the system creates realistic patient profiles. Specialized agents handle tasks ranging from data retrieval to reasoning and response generation, ensuring that interactions are personalized and contextually relevant. This innovative simulation aids in clinical decision support and medical training by providing a dynamic, data-driven patient model [7].This paper offers a comprehensive analysis of multi-agent AI systems tailored for healthcare, combining large language models with reinforcement learning, ensemble methods, and IoT sensor networks. It examines how explainable AI tools such as SHAP and LIME enhance transparency and quality control in clinical decision-making. The system’s integration with EHRs through secure APIs and standardized protocols demonstrates its potential to improve diagnostic accuracy, resource allocation, and real-time patient monitoring while maintaining robust data security [8].Leveraging the BERT model for natural language understanding, this study develops a medical chatbot that demonstrates high accuracy and precision in processing user queries. By capitalizing on BERT’s ability to capture contextual nuances and complex medical terminology, the system provides reliable and coherent responses. The work underscores the efficacy of transformer-based architectures in creating more intuitive and effective communication tools for healthcare settings [9].This comprehensive study explores the transformative potential of generative AI models—including Med-PaLM, BioGPT, ChatGPT, and DALL-E—across various healthcare applications such as medical imaging, drug discovery, personalized treatment, and clinical trials. Through an extensive review of case studies and emerging models, the paper discusses both the capabilities and limitations of these technologies, emphasizing their potential to revolutionize healthcare diagnostics, research, and patient engagement while also addressing challenges like data quality and ethical concerns [10].In this work, a contextual chatbot is developed using deep learning methodologies that combine tokenization, stemming, and bag-of-words representations with a feed-forward neural network implemented in TensorFlow. The system processes user inputs stored in JSON formats to accurately determine intent and generate context-sensitive responses. By integrating auxiliary tools such as Pickle and Random to enhance performance, the study demonstrates how deep learning can be harnessed to provide nuanced and reliable healthcare support in real time [11].This paper outlines an AI-based healthcare chatbot that employs TF-IDF and cosine similarity algorithms to effectively classify and manage common health conditions. By differentiating between minor and major ailments, the system offers appropriate guidance—ranging from medicine suggestions for minor issues to scheduling appointments for serious conditions. The study provides evidence that even relatively simple NLP techniques can substantially improve patient triage, reduce healthcare costs, and ease the burden on medical professionals [12]. Offering a comprehensive review, this paper examines a variety of techniques used in the development of medical chatbots, including natural language processing, machine learning, knowledge graphs, ensemble learning, and bi-directional attention models. It discusses how the integration of these methods leads to improved diagnostic accuracy and more realistic simulations of doctor-patient interactions. The review emphasizes that a hybrid approach, combining multiple advanced techniques, is essential for overcoming the limitations of single-method systems and achieving robust performance in healthcare applications [13].This study introduces EREBOTS, a platform designed to support personalized health-assistant chatbots while ensuring stringent privacy compliance. Utilizing technologies such as MongoDB for non-personal data, Pyrv for managing sensitive information, and the SPADE framework for multi-agent orchestration, the system provides secure, personalized interactions. The integration of communication protocols like Prosody further enables both group and direct messaging, making EREBOTS a versatile solution for multi-scenario healthcare applications that require high levels of data security and user privacy [14].This paper evaluates the performance of several large language model–based systems—including ChatGPT-4, Claude 3 Opus, Gemini Pro 1.5, OpenEvidence, and ChatRWD—in answering real-world clinical questions. Through a rigorous assessment involving 50 clinical queries and evaluations by a panel of physicians, the study reveals significant performance variability among the models. The findings highlight the need for domain-specific training and refinement to enhance the clinical relevance and accuracy of LLMs, underscoring the potential of specialized AI systems in clinical decision support [15].Although focused on the restaurant industry, this paper presents a novel RAG framework that integrates a T5 base model with a Neo4j knowledge graph and TF-IDF embeddings to enhance the quality of automated responses. Evaluated using the BLEU metric, the study demonstrates improved response precision and conversational diversity. The methods outlined offer transferable insights for healthcare applications, particularly in terms of enhancing context-aware retrieval and generating high-quality, domain-specific responses in conversational agents [16].This research investigates the deployment of multi-agent systems using frameworks such as LangGraph and CrewAI, combined with Langsmith and ollama embedding models, and supported by Faiss for vector search. The study details how task delegation, real-time feedback, and graph-structured control can be orchestrated in a multi-agent environment to improve efficiency and scalability. The approach is particularly relevant for complex healthcare applications where collaborative agents must process and integrate vast amounts of data in real time [17].This paper outlines an agentic framework that incorporates a search agent, web-scraping agent, and comparison agent to autonomously analyze and compare products across multiple brands online. By leveraging dynamic data aggregation and advanced decision-making algorithms, the system provides comprehensive comparative analyses. While the focus is on e-commerce, the methodology offers valuable insights for healthcare product recommendation systems, demonstrating how agentic AI can facilitate informed decision-making in competitive markets [18].Focusing on the application of agentic AI systems in data analytics, this paper compares the performance of vertical AI agents across multiple industries—including retail, financial services, manufacturing, supply chain, and trading operations. By presenting detailed performance metrics in terms of accuracy, cost reduction, and efficiency improvements, the study demonstrates how autonomous AI agents can transform decision-making processes. The insights gained from these cross-industry comparisons provide a robust framework for applying similar methodologies to healthcare analytics [19].This survey paper offers an in-depth comparative analysis of various Retrieval-Augmented Generation (RAG) frameworks, including Agentic RAG, Simple RAG, Advanced RAG, Modular RAG, and Graph RAG. Detailed workflow diagrams and performance evaluations illustrate how each method retrieves and generates content in diverse applications such as healthcare, finance, and education. The survey emphasizes the strengths of Agentic RAG—particularly its scalability, accuracy, and efficiency—positioning it as a promising approach for future intelligent retrieval systems [20].Introducing iMedBot, this paper details a comprehensive web-based intelligent agent that integrates a Flask backend with a React.js frontend, supported by deep learning models such as T5 for text summarization and CNNs for image diagnosis. Hosted on AWS with CI/CD pipelines and HIPAA-compliant security measures, iMedBot delivers accurate predictions and efficient text summaries, seamlessly integrating with electronic health record systems. The study demonstrates the system’s potential to enhance diagnostic workflows and streamline healthcare operations through scalable, secure technology [21].This research presents AgentClinic as a benchmark platform for evaluating AI performance in simulated clinical settings. Utilizing advanced large language models—such as GPT-4, Claude-3.5, and Llama-3—along with multimodal and multilingual capabilities, the framework simulates various clinical roles (doctor, patient, measurement, moderator) to assess diagnostic accuracy and interaction quality. The incorporation of persistent notebooks and adaptive learning tools provides a rigorous environment to test AI-driven clinical decision-making, highlighting both current challenges and opportunities for improvement in AI-assisted healthcare [22].

# Methodology

The project uses a multi-agent system designed to process user questions through a structured pipeline that intelligently determines the most appropriate response strategy. Its designed for healthcare E-commerce system. The system integrates rule-based decision-making with retrieval-augmented generation (RAG) techniques to handle a wide variety of user requests efficiently. The methodology consists of interconnected components that work sequentially to analyze, route, and respond to user inputs, ensuring accuracy and relevance in every interaction.

The system architecture follows a linear workflow with conditional branching based on query analysis. The process begins when a user submits a query, which is first evaluated by the Query Validator Agent. This agent performs several checks, including syntax validation. If the query fails any of these checks, the system responds the use about the problem and ask them to submit query again. Valid queries proceed to the Router Agent, which classifies them into one of five categories: general information requests, complex questions requiring external knowledge, comparison queries, summarization requests and ordering requests as shown in Fig 1. The routing decision is made through a combination of keyword analysis, sentence structure patterns, predefined intent classifiers, and contextual embeddings. There are certain rules defined for each agent so according to that router routes that query.



*Fig1. Architectural Flow of Agent-Based Query Processing*

For straightforward factual queries related to healthcare products, the General Info Agent retrieves answers from a curated knowledge base of common facts, using similarity search with of embedded query vector and other content embedded vectors stored inside the vector database. For knowledge base we have embedded all our medical products data into vector database. So, whenever user asks query about the medical product then RAG agent will convert that query into vector embedding. Converted query will check if there are any similar vectors in the vector database and then retrieve the required chunks. All the chunks along with query is provided to the Large Language Model to produce a grounded and well-informed response.

Comparison queries are handled by the Comparison Agent, which identifies the entities or concepts to analyze, extracts relevant features, and constructs a structured comparison matrix. The results are presented in both tabular and narrative formats, ensuring clarity and ease of understanding.

For summarization requests, the Summarization Agent determines the scope and length requirements, processes source material through extractive and abstractive techniques, and delivers condensed yet informative summaries in bullet-point or paragraph form based on user preference.

Order related queries, such as order placements, order inquiry are processed by the Order Agent, which extracts product specifications, validates availability, and generates personalized suggestions based on user history, popular items, and complementary products. When query enables the order agent then system open new page with order related tabs, where user can place orders, view previous orders. In previous orders tab, it shows all the information related to the previously ordered products along with their metadata. We are storing all the orders related data into the MongoDB atlas. Even system also provides the images for that particular medical product. Whenever user confirms order then all the order confirmation details will be sent to his/her registered mail. Also, the system supports both text and visual presentation of recommendations, enhancing user experience.

To improve the adaptability and accuracy of the system over time, a feedback loop mechanism is integrated into the architecture. After delivering a response, users are able to provide feedback through a thumbs up and thumbs down buttons. If user likes the response, then it will be stored in the MongoDB database in feedback collection. If the user dislikes the response, then system will pop-up one feedback section at the top of the application where we can write why we don’t like the response and all. These responses are logged and stored securely inside MongoDB database, forming a valuable dataset for monitoring agent performance. Over time, this feedback data can be used to retrain specific agents by changing their prompts to make it better. This continuous learning approach helps the system to evolve based on real user interactions, ultimately improving relevance, personalization, and overall user satisfaction.

Evaluation

Database and security factors

The implementation incorporates multiple knowledge sources, including a local FAQ database, commercial product catalogs, industry-specific references, and curated web content. All text generation components share common optimization features such as temperature-controlled randomness, length adjustments, repetition prevention, and style matching to maintain consistency and clarity.

To ensure continuous improvement, the system undergoes rigorous evaluation through automated correctness scoring, human rating panels, A/B testing, and latency monitoring. Comprehensive error handling mechanisms are in place, including fallback responses for low-confidence scenarios, clarification prompts for ambiguous queries, and graceful degradation in case of component failures.

This methodology provides a robust and scalable framework for handling diverse user queries through specialized agents while maintaining coherence across the entire system. The modular design allows for independent enhancements of each component, ensuring adaptability and long-term efficiency.

# Result and discussion

# Conclusion and future scope

# refrences

1. Kavitha, B. R., and Chethana R. Murthy. "Chatbot for healthcare system using Artificial Intelligence." *Int J Adv Res Ideas Innov Technol* 5 (2019): 1304-1307..
2. AIML-Based Healthcare Chatbot on the Pandorabots Platform
3. Quidwai, Mujahid Ali, and Alessandro Lagana. "A rag chatbot for precision medicine of multiple myeloma. medRxiv." (2024): 2024-03
4. Wu, Junde, et al. "Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation." *arXiv preprint arXiv:2408.04187* (2024).
5. Sudarshan, Malavikha, et al. "Agentic LLM Workflows for Generating Patient-Friendly Medical Reports." *arXiv preprint arXiv:2408.01112* (2024).
6. Bousetouane, Fouad. "Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents." *arXiv preprint arXiv:2501.00881* (2025).
7. Yu, Huizi, et al. "AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow." *arXiv preprint arXiv:2409.18924* (2024).
8. Borkowski, Andrew, and Alon Ben-Ari. "Muli-Agent AI Systems in Healthcare: Technical and Clinical Analysis." (2024).
9. Babu, Arun, and Sekhar Babu Boddu. "Bert-based medical chatbot: Enhancing healthcare communication through natural language understanding." *Exploratory research in clinical and social pharmacy* 13 (2024): 100419.
10. Chamola, Vinay. "Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations." (2024).
11. Kandpal, Prathamesh, et al. "Contextual chatbot for healthcare purposes (using deep learning)." *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2020.
12. AI-Based Healthcare Chatbot Using TF-IDF and Cosine Similarity
13. Reyner, Andrew, et al. "Medical Chatbot Techniques: A Review." *Medical Chatbot Techniques: A Review* (2020).
14. Calvaresi, Davide, et al. "EREBOTS: Privacy-compliant agent-based platform for multi-scenario personalized health-assistant chatbots." *Electronics* 10.6 (2021): 666.
15. Low, Yen Sia, et al. "Answering real-world clinical questions using large language model based systems." *arXiv preprint arXiv:2407.00541* (2024).
16. Bhat, Vani, et al. "Retrieval augmented generation (rag) based restaurant chatbot with ai testability." *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*. IEEE, 2024.
17. Duan, Zhihua, and Jialin Wang. "Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+ CrewAI." *arXiv preprint arXiv:2411.18241* (2024).
18. Agentic Frameworks for Product Comparison Among Brands
19. Agentic AI in Data Analytics: Transforming Autonomous Insights
20. Singh, Aditi, et al. "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG." *arXiv preprint arXiv:2501.09136* (2025).
21. Xu, Chuhan, and Xia Jiang. "iMedBot: A Web-based Intelligent Agent for Healthcare Related Prediction and Deep Learning." *arXiv preprint arXiv:2210.05671* (2022).
22. Schmidgall, Samuel, et al. "AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments." *arXiv preprintarXiv:2405.07960* (2024).