SMART AGENTIC MEDIRAG

Mr. Mayank Chauhan

Deepika S, Harish P, Kousik K Y, Kowshik S

BACHELOR OF TECHNOLOGY – 4th YEAR

DEPARTMENT OF ARITIFICIAL INTELLIGENCE AND DATA SCIENCE

SRI SHAKTHI OF ENGINEERING AND TECHNOLOGY(AUTONOMOUS)

COIMBATORE-641062

ABSTRACT

Medical diagnosis and healthcare assistance often rely on human expertise, which can be time-consuming, error-prone, and inaccessible in remote areas. This paper presents **SMART AGENTIC MEDIRAG**, an intelligent AI-powered healthcare assistant that combines Retrieval-Augmented Generation (RAG) with agentic workflows to provide accurate, context-aware medical insights. The system leverages Large Language Models (LLMs), medical knowledge bases, and multi-agent coordination to analyze symptoms, retrieve relevant medical information, and generate reliable responses. By integrating semantic search, vector databases, and real-time reasoning agents, the system enhances diagnostic support, patient interaction, and healthcare accessibility. Experimental outcomes indicate that the agentic RAG architecture significantly improves response accuracy, contextual relevance, and decision support compared to traditional chatbot-based systems.

Keywords: Agentic AI, RAG, Healthcare AI, Medical Chatbot, Semantic Search, LLM, Vector Database, Clinical Decision Support.

INTRODUCTION

Healthcare systems worldwide face challenges such as limited access to medical professionals, delayed diagnosis, and increasing patient loads. Traditional digital healthcare tools often rely on static databases or rule-based systems that fail to understand complex patient queries or provide context-aware responses.

Recent advancements in Artificial Intelligence, particularly Large Language Models (LLMs), have enabled intelligent conversational systems capable of understanding and generating human-like responses. However, standalone LLMs may produce hallucinated or inaccurate medical information without access to reliable external knowledge sources.

To address these limitations, this paper proposes **SMART AGENTIC MEDIRAG**, an advanced healthcare assistant that integrates Retrieval-Augmented Generation (RAG) with agent-based architectures. The system dynamically retrieves verified medical information from structured knowledge bases and processes it using multiple intelligent agents responsible for reasoning, validation, and response generation. This approach ensures accurate, explainable, and context-aware medical assistance, making it suitable for real-world healthcare applications.

LITERATURE REVIEW

The integration of AI in healthcare has gained significant attention due to its potential to improve diagnostic accuracy and patient care.
Smith et al. (2024) explored AI-driven medical chatbots and emphasized the importance of contextual understanding in patient interaction. Their work highlighted the limitations of rule-based systems in handling complex medical queries.
Brown and Lee (2023) investigated Retrieval-Augmented Generation (RAG) models, demonstrating their effectiveness in improving factual accuracy by combining external knowledge retrieval with generative AI.
Johnson et al. (2025) introduced agentic AI systems that utilize multiple intelligent agents for task coordination, enabling better reasoning and decision-making in complex environments such as healthcare.
Patel and Kumar (2024) focused on semantic search and vector databases in medical applications, showing how embeddings can improve retrieval accuracy for clinical data.
These studies collectively establish the foundation for developing an agentic RAG-based healthcare system that ensures accurate, reliable, and scalable medical assistance.

METHODOLOGY

1. Data Collection and Preprocessing

- Collect medical datasets including symptom descriptions, disease information, and treatment guidelines.
- Use trusted sources such as clinical datasets, WHO guidelines, and medical literature.
- Preprocess data by removing noise, normalizing text, and structuring medical information.
- Store processed data in a structured format for efficient retrieval.

2. Knowledge Base and Vector Storage

- Convert medical data into embeddings using transformer-based models.
- Store embeddings in vector databases such as FAISS or Pinecone.
- Enable fast similarity search for retrieving relevant medical context.

3. Agentic Architecture Design

- Implement multiple AI agents with specific roles:
  - Query Agent: Understands user symptoms
  - Retrieval Agent: Fetches relevant medical data
  - Reasoning Agent: Analyzes retrieved information
  - Validation Agent: Ensures response reliability
  - Response Agent: Generates final output

4. Retrieval-Augmented Generation (RAG)

- Retrieve top-k relevant documents using semantic similarity.
- Pass retrieved context to the LLM.
- Generate accurate and context-aware medical responses.

5. Response Generation and Explanation

- Generate human-readable responses with explanations.
- Provide possible conditions, precautions, and recommendations.
- Ensure ethical constraints and avoid critical medical misguidance.

6. Performance Evaluation

- Measure accuracy, relevance, and response consistency.
- Evaluate system using real-world medical queries.
- Perform error analysis for continuous improvement.

SYSTEM DESIGN

The SMART AGENTIC MEDIRAG system follows a modular architecture combining AI agents and RAG pipelines. The workflow begins with user input, where symptoms or queries are processed by the Query Agent. The Retrieval Agent then searches the vector database to fetch relevant medical information. The Reasoning Agent interprets the retrieved data and collaborates with the Validation Agent to ensure correctness. Finally, the Response Agent generates a structured, user-friendly output.

The frontend interface enables users to interact with the system, while the backend integrates APIs, LLMs, and vector databases to provide real-time responses. This layered architecture ensures scalability, accuracy, and robustness in healthcare assistanc

IMPLEMENTATION

Step 1: Collect and preprocess medical datasets.

Step 2: Generate embeddings using transformer models.

Step 3: Store embeddings in FAISS/Pinecone vector database.

Step 4: Develop agent-based architecture using frameworks like LangChain or LangGraph.

Step 5: Integrate LLM (such as GPT-based models) for response generation.

Step 6: Implement RAG pipeline for context-aware responses.

Step 7: Design frontend interface using React.js.

Step 8: Deploy backend using FastAPI for real-time interaction.

Step 9: Test system with various medical queries.

Step 10: Optimize performance and improve accuracy through feedback.

OUTPUT

The system provides an interactive healthcare assistant interface where users can input symptoms or medical queries. The output includes:

- Possible medical conditions
- Context-aware explanations
- Suggested precautions
- Basic recommendations

The system ensures clarity, reliability, and user-friendly responses for better healthcare support.

improvement.

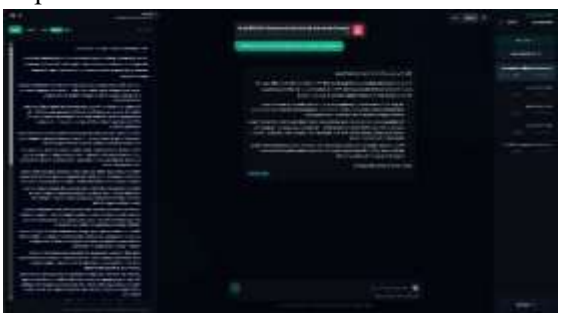

Fig 1: **PDF UPLOAD AND CANVAS EDITING VIEW**

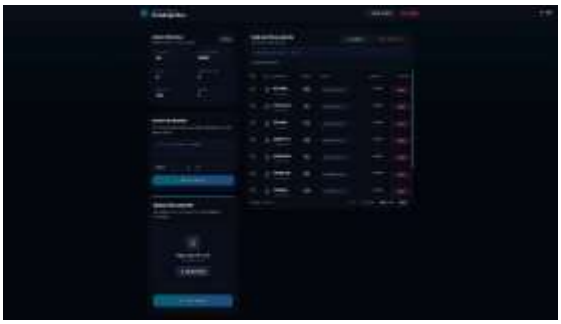

Fig 2: **KNOWLEDGE BASE INDEX (ADMIN CONSOLE)**

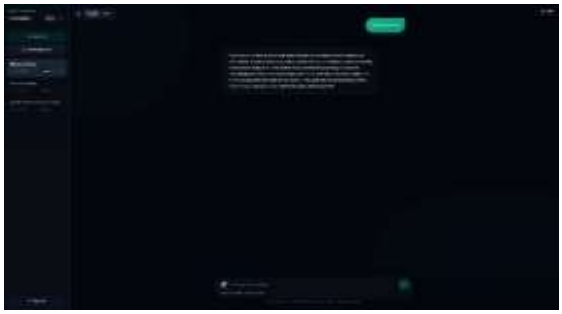

Fig 3: **QUERY RESPONSE (FAST MODE OUTPUT)**

FUTURE ENHANCEMENTS

Integration with wearable health devices for real-time monitoring

Multilingual support for wider accessibility

Voice-based interaction system

Integration with hospital management systems

Advanced predictive diagnostics using patient history

BENEFITS

1. Provides instant medical assistance
2. Improves healthcare accessibility in remote areas
3. Reduces dependency on manual consultation for basic queries
4. Enhances diagnostic support using AI
5. Scalable and efficient healthcare solution
6. Ensures accurate and context-aware responses

CONCLUSION

SMART AGENTIC MEDIRAG presents a powerful AI-driven healthcare assistant that combines Retrieval-Augmented Generation with agentic intelligence. By integrating semantic search, multi-agent reasoning, and large language models, the system delivers accurate, reliable, and context-aware medical insights.

The proposed system overcomes limitations of traditional chatbots by ensuring factual correctness and improved reasoning capabilities. Its scalable and modular design makes it suitable for real-world healthcare applications, contributing to improved patient care, accessibility, and decision-making.

REFERENCES

1. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.https://arxiv.org/abs/2312.10997

2. Gupta, S., et al. (2024). A Comprehensive Survey of Retrieval-Augmented Generation. arXiv preprint arXiv:2410.12837.https://arxiv.org/abs/2410.12837

3. Klesel, M., & Wittmann, H. F. (2025). Retrieval-Augmented Generation (RAG). Business & Information Systems Engineering, 67(4), 551–561. https://doi.org/10.1007/s12599-025-00945-3

4. Patnaik, R. (2025). Retrieval-Augmented Generation: Enhancing AI with Reliable Knowledge. International Journal of Science and Research. https://langchain-

ai.github.io/langgraph/

5. Gupta, S., et al. (2024). A Comprehensive Survey of Retrieval-Augmented Generation. arXiv preprint arXiv:2410.12837. https://arxiv.org/abs/2410.12837

6. Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv preprint arXiv:2501.09136. https://arxiv.org/abs/2501.09136.

7. Nguyen, T., et al. (2025). MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Reasoning. arXiv preprint arXiv:2505.20096. https://arxiv.org/abs/2505.20096.

8. Maragheh, R. Y., et al. (2025). ARAG: Agentic Retrieval-Augmented Generation for Personalized Recommendation. arXiv preprint arXiv:2506.21931. https://arxiv.org/abs/2506.21931

9. Liang, J., et al. (2025). Reasoning Agentic Retrieval-Augmented Generation: A Survey. arXiv preprint arXiv:2506.10408. https://arxiv.org/abs/2506.10408

10. Du, M., Xu, B., Zhu, C., Wang, S., Wang, P., Wang, X., & Mao, Z. (2026). A-RAG: Scaling Agentic Retrieval-Augmented Generation via Hierarchical Retrieval Interfaces.