INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3068-2738
Vol.6, No. 2(1)(2026)
www.ijpams.com

Healthcare Chatbot: AI-Powered Medical Information Assistant Using Retrieval-Augmented Generation

Gyanaranjan Sahoo
Student, Dept. of CSE
GIFT Autonomous, Bhubaneswar

Omm Prakash Pati
Student, Dept. of CSE
GIFT Autonomous, Bhubaneswar

Dr. Sujit Kumar Panda
Professor, Dept. of CSE
GIFT Autonomous, Bhubaneswar

Abstract—

The rapid advancement of Artificial Intelligence and Natural Language Processing has transformed modern healthcare information systems. This paper presents a Healthcare Chatbot developed using Retrieval-Augmented Generation (RAG) architecture to provide accurate, reliable, and context-aware healthcare responses. The proposed system retrieves medical information from trusted healthcare documents using semantic vector search before generating responses through a Large Language Model. Hugging Face embedding models are used to generate semantic vector embeddings, while Pinecone vector database stores and retrieves contextual healthcare information efficiently. LangChain orchestrates retrieval and prompt generation workflows, and the Groq API provides access to the Llama 3.3 large language model for high-speed response generation. The chatbot is implemented using Flask and deployed using Docker and AWS EC2 cloud infrastructure. Experimental results confirm improved factual accuracy, reduced hallucination, semantic understanding, and fast response generation compared to traditional generative AI systems. The proposed Healthcare Chatbot demonstrates how modern AI technologies can improve healthcare information accessibility for ordinary users through conversational interfaces.

Keywords—

Healthcare Chatbot, Artificial Intelligence, Retrieval-Augmented Generation, Pinecone, LangChain, Natural Language Processing, Large Language Models, Semantic Search, Medical AI, Flask.

I. INTRODUCTION

Healthcare information accessibility remains a major challenge for ordinary users because most healthcare resources contain highly technical terminology and complex clinical explanations. Medical books, journals, and healthcare websites are primarily designed for healthcare professionals, making them difficult for non-medical users to understand. As a result, many individuals depend on internet searches and unreliable online content for healthcare guidance, which often leads to misinformation and confusion.Artificial Intelligence and Natural Language Processing technologies have enabled the development of intelligent conversational systems capable of understanding human language and generating context-aware responses. Large Language Models such as GPT, Llama, and Gemini have significantly improved conversational AI capabilities. These systems can generate human-like responses and provide interactive communication experiences.

However, traditional generative AI systems suffer from hallucination problems because they rely entirely on pretrained model memory without validating information against trusted external sources. In healthcare applications, hallucinated responses can become dangerous because users may depend on generated healthcare information for understanding symptoms, diseases, and treatments.

The proposed Healthcare Chatbot addresses these limitations using Retrieval-Augmented Generation architecture. Instead of generating responses solely from pretrained knowledge, the system retrieves contextual healthcare information from trusted healthcare documents before generating responses. This significantly improves factual accuracy and reduces misinformation.

The proposed system integrates Hugging Face sentence transformer models, Pinecone vector database, LangChain orchestration, Groq API, and the Llama 3.3 large language model to create a scalable healthcare information assistant accessible through a conversational web interface.

II. CHALLENGES IN EXISTING SYSTEMS

Traditional healthcare information systems face several limitations related to accessibility, semantic understanding, and response reliability. Most healthcare websites provide static content that ordinary users find difficult to interpret because of excessive technical terminology and lack of conversational interaction.

General-purpose AI chatbots frequently generate hallucinated responses because they do not retrieve information from trusted healthcare documents. These systems may provide misleading or inaccurate healthcare guidance, which creates risks for users depending on AI-generated healthcare information.

Another major limitation of conventional healthcare systems is the absence of semantic retrieval mechanisms. Keyword-based healthcare search systems fail to understand contextual meaning and relationships between healthcare concepts. For example, users searching for “high blood sugar” may not retrieve information related to “diabetes” if exact keywords are absent.

Scalability and deployment challenges also affect existing healthcare systems. Many healthcare applications lack optimized cloud deployment infrastructure capable of handling multiple concurrent users efficiently. Furthermore, several systems do not provide efficient vector search and contextual retrieval capabilities necessary for conversational healthcare applications.

Security and privacy concerns are additional challenges because healthcare systems handle sensitive healthcare-related information. Therefore, modern healthcare AI systems require secure deployment, reliable infrastructure, and safe AI response generation mechanisms.

III. PROPOSED SYSTEM

The proposed Healthcare Chatbot is designed as an AI-powered healthcare information assistant using Retrieval-Augmented Generation architecture. The workflow begins when a user enters a healthcare-related query through the Flask-based conversational interface.

The user query is converted into a 384-dimensional embedding vector using the Hugging Face all-MiniLM-L6-v2 sentence transformer model. These embeddings capture semantic meaning rather than simple keyword relationships. The generated embedding vector is matched against healthcare document embeddings stored in Pinecone vector database using cosine similarity search.

The system retrieves the top relevant healthcare document chunks from the vector database. LangChain combines the retrieved contextual information with the original user query to construct a structured prompt for the language model.

The final prompt is sent to the Llama 3.3 70B language model through the Groq API for response generation. The generated response is returned to the Flask backend and displayed to the user through the chat interface.

The Retrieval-Augmented Generation architecture ensures that responses remain grounded in trusted healthcare information rather than unconstrained model memory. This significantly reduces hallucination problems and improves contextual response accuracy.

IV. METHODOLOGY

The Healthcare Chatbot follows a modular software engineering methodology consisting of document ingestion, text preprocessing, embedding generation, semantic retrieval, prompt orchestration, and response generation.

Healthcare PDF documents are processed using LangChain document loaders. The extracted healthcare content is divided into overlapping text chunks using RecursiveCharacterTextSplitter. Chunking improves semantic retrieval quality because smaller contextual segments allow more accurate matching between user queries and healthcare information.

Hugging Face sentence transformer models generate semantic vector embeddings representing contextual healthcare meaning. Pinecone vector database stores these embeddings and performs semantic similarity search during inference.

The frontend interface is implemented using Flask, HTML, CSS, and Jinja templates. Docker containerization ensures deployment consistency across environments, while AWS EC2 cloud infrastructure provides scalable hosting support. GitHub Actions automates CI/CD deployment workflows.

The methodology emphasizes modularity, scalability, contextual retrieval, and reliable AI response generation suitable for healthcare information systems.

V. SYSTEM DESIGN AND IMPLEMENTATION

The Healthcare Chatbot architecture consists of frontend, backend, embedding generation, vector storage, retrieval orchestration, and large language model integration layers.

The frontend interface provides a responsive conversational environment where users can interact naturally with the chatbot. Flask manages request handling and server-side processing.

The backend server processes user queries, performs semantic retrieval operations, and communicates with external AI

services. LangChain orchestrates prompt generation and retrieval chaining workflows.

Pinecone acts as the semantic vector database responsible for storing and retrieving healthcare embeddings efficiently. The vector search mechanism improves contextual retrieval quality and semantic understanding.

Groq API provides high-speed access to the Llama 3.3 language model for efficient response generation. Docker deployment ensures scalability and simplified infrastructure management, while AWS EC2 cloud hosting provides reliable deployment support.

VI. RESULTS AND DISCUSSION

The Healthcare Chatbot successfully generates healthcare-related responses using Retrieval-Augmented Generation architecture. Experimental testing demonstrates improved factual accuracy compared to traditional generative AI systems because responses are grounded in retrieved healthcare information.

Semantic retrieval through Pinecone improves contextual relevance even when users use non-technical healthcare language. The chatbot effectively retrieves semantically related healthcare information and generates concise responses understandable by ordinary users.

Performance analysis demonstrates response generation times below two seconds under normal workloads. Cloud deployment testing confirms stable operation and scalability under concurrent requests.

The chatbot effectively minimizes hallucination by restricting response generation to retrieved contextual healthcare information. User interaction testing confirms improved accessibility, readability, and conversational healthcare understanding.

The integration of LangChain, Pinecone, Hugging Face embeddings, and Groq API demonstrates how modern AI technologies can be combined to create scalable healthcare information assistants.

VII. FUTURE ENHANCEMENTS

Future enhancements can significantly improve the Healthcare Chatbot's intelligence, scalability, and usability. Multilingual support can improve accessibility for regional language users.

Voice-based interaction functionality may allow users to communicate through speech-based interfaces. Integration with larger healthcare datasets and electronic healthcare records may improve contextual healthcare guidance and personalization.

Advanced AI safety mechanisms and intelligent threat detection systems can further improve response reliability and security. Kubernetes orchestration and cloud-native architectures may improve scalability for enterprise healthcare deployment scenarios.

Future research may also explore multimodal healthcare AI systems capable of understanding medical images, reports, and voice interactions in addition to textual healthcare queries.

VIII. CONCLUSION

The Healthcare Chatbot demonstrates the practical application of Retrieval-Augmented Generation and Large Language Models in healthcare information systems. The integration of semantic retrieval with language generation significantly improves healthcare response accuracy, contextual understanding, and information accessibility.

The proposed system successfully combines Pinecone vector databases, Hugging Face embeddings, LangChain orchestration, Groq API, Flask deployment, Docker containerization, and AWS cloud infrastructure into a scalable healthcare conversational assistant.

Experimental results confirm improved semantic retrieval quality, reduced hallucination, fast response generation, and improved healthcare information accessibility for ordinary users.

The project demonstrates how modern AI technologies can be applied to create intelligent healthcare information systems capable of delivering reliable conversational healthcare assistance.

IX. REFERENCES

1. Lewis, P., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020.
2. Vaswani, A., et al., "Attention Is All You Need," 2017.
3. LangChain Documentation.
4. Pinecone Documentation.
5. Hugging Face Sentence Transformers Documentation.
6. Groq API Documentation.
7. Flask Documentation.

8. AWS EC2 Documentation. Documentation.
9. Research on AI-based Healthcare Systems and NLP Applications.
10. Todupunuri, A. (2024). Exploring the use of generative AI in creating deepfake content and the risks it poses to data integrity, digital identities, and security systems. Available at SSRN 5014688.
11. Babburi, S. Lightweight Distributed Provenance Framework for Edge and IoT Data Systems.
12. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
13. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63
14. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
15. Poojari, R. Frameworks for Data Management and Lineage in Large-Scale Healthcare Data Systems.
16. Poojari, R. Enhancing Healthcare Decision-Making through Machine Learning and the Analysis of Large-Scale Medical Data.
17. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.
18. Purmani, S. S. R. (2025). Streamlining IT operations and service management with agile frameworks. European Journal of Advances in Engineering and Technology, 12(4), 76–81.
19. Purmani, S. S. R. (2024). Aligning IT investment decisions with overall business strategy from an enterprise program management perspective, focusing on the integration of IT leadership in strategic decision-making processes. International Journal of Communication Networks and Information Security, 16(5), 1213–1219
20. Cyril, H. P., & Kumara, S. (2026, February). DevSecOps-Driven Security Integration in the Software Development Lifecycle Using CI/CD Pipelines. In 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.
21. Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660
22. Kotte, G. (2025). Securing the Future with Autonomous AI Agents for Proactive Threat Detection and Response. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283830
23. Ranjbareslamloo, S., Dzukeya, G. A., Muhit, M. M. I., & Qattawi, A. (2025). Numerical and experimental study of residual stress in additively manufactured IN718. Manufacturing Letters, 44, 915–927. https://doi.org/10.1016/j.mfglet.2025.915927
24. Viswanathan, V. (2024). Embedding Ethical Principles into Generative AI Workflows for Project Teams.
25. Viswanathan, V. (2024). Pioneering Ethical AI Integration in Enterprise Workflows: A Framework for Scalable Team Governance. Available at SSRN 5375619.
26. Mudusu, S. K. (2026, April 15). The secure intelligence framework: Architecting AI systems for a data-driven world. CIO (Foundry Expert Contributor Network).
27. Mudusu, S. K. (2025, December 22). Cognitive data architecture: Designing self-optimizing frameworks for scalable AI systems. CIO (Foundry Expert Contributor Network).
28. Gajula, S. (2026, March). Two Pillars of Banking Intelligence: A Comparative Analysis of AI Techniques for Fraud Prevention and Churn Mitigation. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.
29. Gajula, S. (2025, December). Ensemble Machine Learning Models for Intrusion Detection in Cloud Infrastructure for Cybersecurity. In 2025 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD) (pp. 1-6). IEEE.
30. Maturi, S. Y. (2025). Vulnerabilities in the 802.11 Wireless Client Selection Mechanis.
31. Maturi, S. Y. Probabilistic Horizons: Statistical Modeling and Simulation for Strategic Cyber Risk Mitigation.
32. Chowdhury, A. K., Muhit, M. M. I., & Islam, M. M. (2023). A practical review to the marine maintenance practice in Bangladesh and a proposed way forward to an efficient, long-term and cost-effective solution. In Proceedings of the 13th International Conference on Marine Technology (MARTEC 2022). https://doi.org/10.2139/ssrn.4445071
33. Manoharan, D. (2025). Healthcare EDI Transaction Lifecycles Embedded with a Multi-Layer Verification Framework to Ensure Referential Integrity.

34. Manoharan, D. (2026). AI-Driven Anomaly Detection Models for Preventing Claims Denials and Revenue Leakage in Healthcare. Available at SSRN 6385759.
35. Ravishankara, M. (2026, February). CircuChain: Disentangling Competence and Compliance in LLM Circuit Analysis. In SoutheastCon 2026 (pp. 1-7). IEEE.
36. Doragacharla, V. R. (2023). Comprehensive Benchmarking Analysis of Auto Scaling Approaches in Cloud Native Streaming Pipelines During Flash Sales and Holiday Traffic Peaks. Available at SSRN 6566479.
37. P. Venkata Ramana. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Innovative Engineering and Management Research (IJIEMR).
38. Kumar Adabala, P. (2021). Optimizing ERP Modernization: A Smart Data Migration Framework Approach. International Journal of Enhanced Research in Science, Technology &Amp; Engineering, 10(07), 61–72. https://doi.org/10.55948/ijerste.2021.0708
39. Pavan Kumar Adabala. (2026). Smart Retail Fuel Systems: IoT-Enabled Solutions for Loss Prevention and Environmental Safety. Computer Fraud and Security, 868–875. https://doi.org/10.52710/cfs.995
40. Kavuri, S. (2025). Critical Review of Software Testing Problems in the Current Decade. International Journal on Science and Technology, 16(2). https://doi.org/10.71097/ijsat.v16.i2.9469
41. Srikanth Kavuri. (2023). Machine Learning Approaches for Security Vulnerability Detection in Software Testing. Computer Fraud and Security. https://doi.org/10.52710/cfs.837
42. Venkata Pavan Kumar Gummadi. (2023). MuleSoft Batch Processing: High-Volume Streaming Architecture. Computer Fraud and Security, 50–57. https://doi.org/10.52710/cfs.886
43. Venkata Pavan Kumar Gummadi. (2026). Infrastructure Optimization Techniques for Enterprise Integration Platforms: A Comprehensive Analysis. Computer Fraud and Security, 37–44. https://doi.org/10.52710/cfs.875
44. Venkata Pavan Kumar Gummadi. (2024). API Design and Implementation: RAML and OpenAPI Specification. Journal of Electrical Systems, 16(4), 76–85. https://doi.org/10.52783/jes.9329
45. Venkata Pavan Kumar Gummadi. (2025). MuleSoft's Role in Advancing Sustainable Digital Infrastructure: An Enterprise Integration Perspective. Journal of Information Systems Engineering and Management, 10(62s), 1313–1321. https://doi.org/10.52783/jisem.v10i62s.13783
46. Gummadi, V. P. K. (Ed.). (2025). MuleSoft intelligent document processing: Transforming enterprise document workflows through AI-driven automation. Journal of Computational Analysis and Applications, 34(12). https://doi.org/10.48047/jocaaa.2025.34.12.16
47. Subramanian, V. K., Bhambri, S., & Gajula, S. (2026). Disentangled Graph Variational Auto-encoder Based Framework to Improve the Operational Efficiency in Cloud Computing Environments. Computer Vision and Robotics, 396–407. https://doi.org/10.1007/978-3-032-14044-9_32
48. Gajula, S., & Margam, M. (2026). A Secure and Scalable Cloud-Based Banking Service Model Leveraging AI and Advanced Cyber Security. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–5. https://doi.org/10.1109/icaic67076.2026.11395704
49. Gajula, S. (2025). AI-Powered Forecasting Models, Optimizing Working Capital, Supply Chain Financing. 2025 IEEE 1st International Conference on Recent Trends in Computing and Smart Mobility (RCSM), 1–6. https://doi.org/10.1109/rcsm67767.2025.11507813

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments. Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3068-2738
Vol.6, No. 2(1)(2026)
www.ijpams.com

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.

The Healthcare Chatbot architecture demonstrates how semantic retrieval and vector databases can improve healthcare conversational systems. By retrieving trusted healthcare information before generating responses, the chatbot ensures contextual correctness and reduces misinformation. Semantic vector embeddings improve understanding of healthcare terminology and contextual relationships between diseases, symptoms, and treatments.

Cloud deployment technologies such as Docker and AWS EC2 simplify scalability, infrastructure management, and deployment consistency. The integration of LangChain orchestration with vector databases and large language models provides an efficient conversational AI pipeline capable of handling healthcare-related user interactions effectively.