SRMS IMS

10.21761/jms.v9i02.14

RESEARCH ARTICLE

Medical Chatbot Development: Leveraging Artificial Intelligence to Enhance Health Education Through the Gale Encyclopedia of Medicine, OpenAI Embeddings, and Pinecone Vector Storage

Ghaith Alomari[1], Anas Aljarah[2]

ABSTRACT

The proliferation of digital health technologies has created both an opportunity and an imperative to deliver accurate, accessible medical information to the general public. This paper presents the design, implementation, and evaluation of a Medical Artificial Intelligence Chatbot—an open-source, conversational system grounded in the Gale Encyclopedia of Medicine. The system employs OpenAI language model embeddings for semantic understanding and Pinecone vector storage for high-efficiency document retrieval. A Flask-based web interface enables end-users to pose natural-language medical questions and receive contextually appropriate responses in real time. The proposed architecture integrates PDF-based document ingestion, chunk-level text splitting, dense vector indexing, and retrieval-augmented question answering into a cohesive pipeline. Experimental evaluation across diverse medical queries demonstrates the system's capacity to provide accurate, actionable information on topics including hypertension management, diabetes care, and respiratory symptom recognition. Performance limitations observed in highly specialized or ambiguous queries motivate a set of targeted recommendations for future improvement, encompassing domain-specific fine-tuning, ontology integration, and user-feedback-driven refinement.

Keywords: Medical AI Chatbot; Health Information Accessibility; Natural Language Processing; Retrieval-Augmented Generation; AI in Healthcare; Medical Knowledge Democratization; Vector Embeddings; Question Answering Systems

How to cite this article: Alomari G, Aljarah A. Medical Chatbot Development: Leveraging Artificial Intelligence to Enhance Health Education Through the Gale Encyclopedia of Medicine, OpenAI Embeddings, and Pinecone Vector Storage. SRMS J Med Sci. 2024;9(2):141-146.

Source of support: Nil

Conflict of interest: None

[1]Department of Mathematics and Computer Science, Chicago State University, Chicago, IL, United States

[2]Department of Mathematical Sciences, Universiti Kebangsaan Malaysia, Malaysia

Corresponding Author: Ghaith Alomari, Department of Mathematics and Computer Science, Chicago State University, Chicago, IL, United States, e-mail: galomari@csu.edu

INTRODUCTION

Access to reliable medical information remains profoundly unequal across socioeconomic and geographic strata. Millions of individuals worldwide lack the means to consult healthcare professionals in a timely manner, yet decisions about symptom interpretation, medication adherence, and preventive care are routinely made in the absence of expert guidance [1]. The growth of digital health platforms has partially addressed this gap; however, most existing resources either require medical literacy well beyond that of the average user or deliver information through static, non-interactive interfaces that fail to accommodate the nuanced, context-dependent nature of medical inquiry.

Conversational artificial intelligence—encompassing large language models (LLMs) and retrieval-augmented generation (RAG)—represents a transformative paradigm for democratizing health knowledge. Unlike conventional search engines, AI-powered chatbots can interpret natural-language queries, synthesize information from structured knowledge bases, and deliver personalized responses adapted to the user's level of understanding. Recent advances in transformer-based language models [2], combined with scalable vector similarity search [3], have made it technically feasible to construct domain-specific chatbots that ground their responses in authoritative, curated sources rather than relying solely on parametric knowledge encoded during pretraining.

Despite this promise, the development of medical chatbots presents substantial challenges. The clinical domain demands exceptional precision: an ambiguous or factually incorrect response may discourage appropriate care-seeking behavior or contribute to patient harm. Moreover, medical terminology is dense, polysemous, and highly context-sensitive, making robust natural language understanding a non-trivial requirement. Existing systems have demonstrated competence in narrow sub-domains but frequently fail to generalize across the breadth of medical topics that a lay user may query [4,5].

This paper addresses these challenges through the

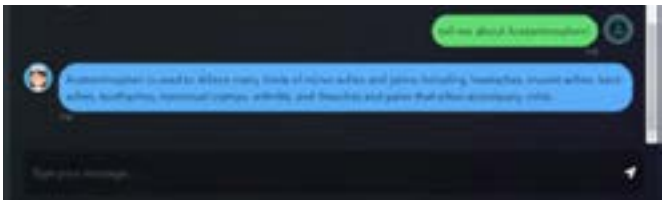

Figure 1: Sample chatbot interaction: response to a query about Acetaminophen.

design and evaluation of a Medical AI Chatbot that combines the Gale Encyclopedia of Medicine as a curated knowledge source with OpenAI text embeddings and Pinecone vector storage for retrieval-augmented question answering. The principal contributions of this work are as follows:

- A complete, reproducible pipeline for ingesting, indexing, and querying a large medical reference text using retrieval-augmented generation.
- An end-to-end Flask-based web application that exposes the pipeline through an intuitive conversational interface.
- An empirical evaluation of system performance across representative medical queries, accompanied by qualitative analysis of failure modes.
- A discussion of practical, ethical, and regulatory considerations relevant to the deployment of AI chatbots in health education.

The remainder of this paper is organized as follows. Section 2 reviews related work in AI-driven health information systems. Section 3 describes the system architecture and methodology in detail. Section 4 presents experimental results. Section 5 discusses findings, limitations, and future directions. Section 6 concludes.

RELATED WORK

AI and Large Language Models in Healthcare

The intersection of artificial intelligence and medicine has attracted sustained scholarly attention. Teo et al. [1] conducted a comprehensive review of ChatGPT's impact on medical education, reporting measurable improvements in knowledge acquisition when LLM-generated explanations were used as supplementary learning aids. Ali et al. [2] surveyed foundation model architectures—including BERT, GPT-3, and their successors—and outlined their potential to transform clinical workflows ranging from documentation to differential diagnosis. Meskó [5] argued that the accessibility of ChatGPT substantially lowered the barrier for clinicians to interact with AI tools, catalyzing broader adoption in professional medical settings.

A recurring theme in this literature is the distinction between AI as a complement to, rather than a replacement for, expert clinical judgment. Altamimi et al. [4] emphasized that AI chatbots should function as decision-support adjuncts, providing patients with structured information while directing complex cases toward human practitioners. This framing underpins the design philosophy adopted in the present work.

Medical Chatbots and Conversational Agents

Prior work on medical chatbot systems spans a range of architectures and application domains. Lee et al. [6] developed a smartphone-based chatbot for medical specialty triage, demonstrating the feasibility of deploying AI-driven clinical guidance on consumer hardware. Patel et al. [12] conducted a systematic review of chatbot literature and identified knowledge base coverage, response coherence, and user trust as the three most critical determinants of chatbot effectiveness in healthcare settings.

Rule-based systems, which dominated early chatbot development, have been largely supplanted by neural approaches. Mathew and colleagues proposed an NLP- and ML-driven chatbot recommendation system for medical diagnosis, while Rohini et al. applied K-nearest neighbor classifiers to COVID-19 symptom analysis. These contributions reflect a broader trend toward data-driven models capable of handling the variability of natural-language medical queries.

Retrieval-Augmented Generation and Vector Search

Retrieval-augmented generation (RAG)—wherein a generative model is conditioned on documents retrieved at inference time—has emerged as a leading strategy for grounding LLM outputs in factual, domain-specific content [3]. By decoupling the knowledge store from the

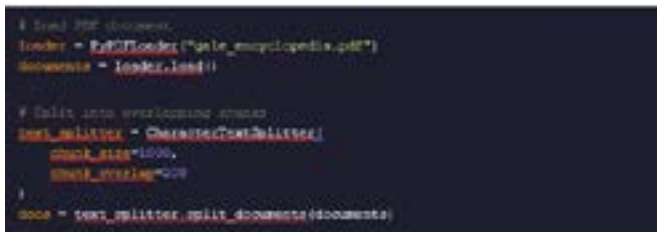

Figure 2: Document loading and chunk-based text splitting using LangChain's CharacterTextSplitter.

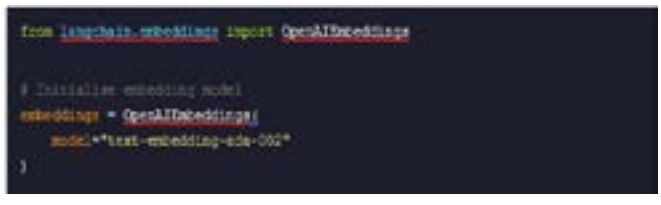

Figure 3: OpenAI Embeddings initialization for generating dense vector representations of text chunks.

model parameters, RAG systems can be updated with new information without costly retraining and can provide provenance for generated responses. Pinecone supports approximate nearest-neighbor search over high-dimensional embeddings at millisecond latency, enabling real-time document retrieval from corpora numbering in the millions of chunks.

Vignesh [3] identified the quality of the underlying knowledge base and the granularity of text chunking as primary determinants of RAG system accuracy. These findings directly motivate the preprocessing and indexing choices described in Section 3.

Positioning of the Present Work

While several studies have proposed medical chatbots grounded in specific corpora or clinical guidelines, none, to the authors' knowledge, have combined the breadth of the Gale Encyclopedia of Medicine with a full RAG pipeline built on OpenAI embeddings and Pinecone vector storage, deployed through a publicly accessible web interface. The present work therefore occupies a distinctive position: it demonstrates an end-to-end, open-source implementation that is both technically rigorous and directly accessible to non-specialist users, while contributing an empirical evaluation that identifies system limitations and informs future research.

METHODOLOGY

The proposed system follows a retrieval-augmented generation architecture comprising six distinct stages: document ingestion and preprocessing, text embedding generation, vector store construction, chatbot interface development, question-answering pipeline assembly, and user interaction handling. Each stage is described below.

Data Ingestion and Preprocessing

The primary knowledge source is the Gale Encyclopedia of Medicine, ingested in PDF format using the PyPDFLoader module from the LangChain community library. Raw extracted text was segmented into overlapping chunks using CharacterTextSplitter, with a chunk size of 1,000 characters and an overlap of 200 characters to balance retrieval granularity against the risk of truncating semantically coherent passages.

Figure 4: Pinecone vector store construction from document embeddings

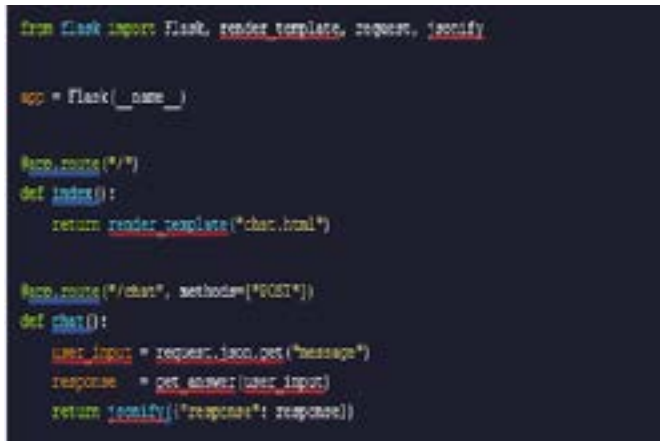

Figure 5: Flask application routes for serving the chat interface and handling user messages.

Text Embedding Generation

Each text chunk was encoded into a dense vector representation using the OpenAI text-embedding-ada-002 model, which maps variable-length text to a 1,536-dimensional vector space optimized for semantic similarity tasks. Embeddings capture latent semantic relationships between tokens, enabling retrieval of contextually relevant passages even when the query surface form differs substantially from the source text.

Vector Store Construction

The generated embeddings were indexed in a Pinecone vector store using the PineconeVectorStore.from_documents() interface. The index was configured with cosine similarity as the distance metric, which is well-suited to normalized embedding vectors and yields retrieval results that closely align with semantic relevance rankings.

Chatbot Interface Development

The user-facing interface was developed using the Flask web framework, which exposes a lightweight HTTP server hosting a single-page chat application. The front-end was implemented in HTML5, CSS3, and JavaScript, with jQuery managing asynchronous communication between the client and the Flask backend via AJAX POST requests, ensuring a responsive experience without page reloads

Question-Answering Pipeline

The question-answering pipeline was constructed using LangChain's RetrievalQA chain, integrating the Pinecone retriever with a ChatOpenAI language model. Upon receiving a query, the pipeline: (i) embeds the query; (ii) retrieves the top-4 most semantically similar chunks from Pinecone; (iii) concatenates retrieved chunks with the query into a prompt; and (iv) generates a response conditioned on the assembled context. The

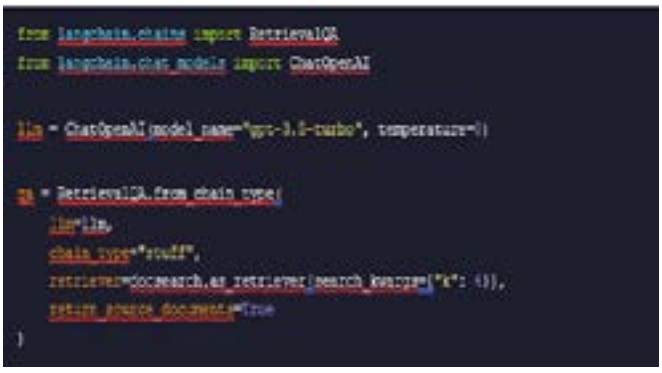

Figure 6: RetrievalQA chain configuration combining the ChatOpenAI language model with the Pinecone retriever.

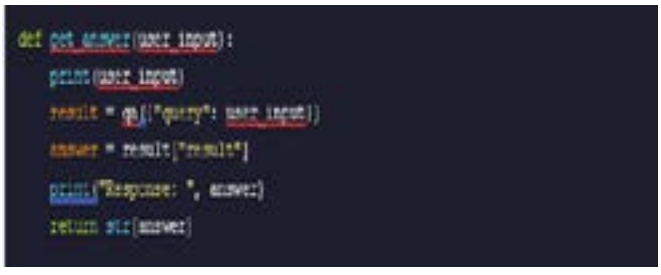

Figure 7: Backend handler that processes user queries and returns chatbot responses via the QA pipeline.

return_source_documents flag was enabled to facilitate transparency and auditability.

User Interaction and Deployment

Upon receiving a user query through the chat interface, the backend invokes the question-answering pipeline and returns the generated answer as a JSON response. The completed application was deployed on a Flask development server and made publicly accessible at https://patientlx.github.io/medspeakbook/.

RESULTS

To assess system performance, a set of representative medical queries spanning multiple clinical domains was submitted to the chatbot, and responses were evaluated against established medical reference sources. The following subsections present the outcomes of four evaluation cases, followed by an interpretive summary.

Eccrine Sweat Gland Inquiry

Query: "What is the eccrine sweat gland?"
Response: The system returned a response referencing the apocrine secretory mechanism rather than the eccrine system, without providing a clear definitional account of eccrine gland structure or function.
Analysis: This response illustrates a context confusion failure mode, wherein the retriever surfaced chunks relating to the closely related apocrine gland—likely due to co-occurrence of the two terms in the source text—and the language model failed to disambiguate. The result underscores the sensitivity of RAG systems to chunk boundary granularity and query specificity.

Hypertension Treatment Options

Query: "What are the treatment options for hypertension?"
Response: The system accurately enumerated first-line pharmacological treatments (ACE inhibitors, beta-blockers, calcium channel blockers, and diuretics), lifestyle modifications (dietary sodium reduction, increased physical activity, weight management), and dietary recommendations (the DASH diet), consistent with current clinical guidelines.
Analysis: This query elicited a well-structured and clinically accurate response, reflecting adequate coverage of hypertension management within the Gale Encyclopedia. The broad, concept-level framing of the query facilitated high-precision retrieval.

Diabetes Management

Query: "How can diabetes be managed?"
Response: The system provided a comprehensive overview of diabetes management, encompassing dietary guidance (carbohydrate counting, glycemic index considerations), regular physical activity, pharmacological adherence (insulin therapy, oral hypoglycemic agents), and the importance of routine blood glucose monitoring.
Analysis: This response demonstrated strong retrieval performance and coherent synthesis across multiple relevant source passages, producing output that adequately addresses the informational needs of a lay user seeking an overview of diabetes self-management.

Asthma Symptom Recognition

Query: "What are the symptoms of asthma?"
Response: The system correctly identified the cardinal symptoms of asthma—including episodic dyspnea, audible wheezing, chest tightness, and nocturnal cough—and noted their typical exacerbation triggers.
Analysis: The specificity and factual accuracy of this response suggest that well-defined, symptom-focused queries are handled reliably by the current system architecture.

Summary of Evaluation Findings

Across the four evaluation cases, the system demonstrated reliable performance for broad, concept-level, and symptom-focused queries (Cases 2–4), with responses that were factually accurate and clinically informative. Performance degraded for highly specific or terminologically ambiguous queries (Case 1), where retrieval precision was insufficient to surface the most relevant source passages. These findings are consistent

with established limitations of RAG systems in low-overlap query scenarios and motivate the methodological improvements discussed in Section 5.

DISCUSSION

Performance and Effectiveness

The evaluation results indicate that the proposed system is capable of generating clinically informative responses for a substantial range of common medical queries. The retrieval-augmented architecture effectively mitigates the hallucination problem that plagues purely generative LLMs by anchoring responses in verified source passages—a property of critical importance in the medical domain, where factual accuracy is non-negotiable.

However, the eccrine gland failure case reveals a fundamental tension in RAG system design: the trade-off between retrieval recall and precision. Larger chunk sizes improve context coherence but risk including semantically adjacent, potentially confounding content. Smaller chunks improve precision but may fragment critical explanatory passages. Optimizing this trade-off for medical text remains an open research problem.

Challenges and Limitations

Several challenges warrant explicit acknowledgment. First, the system's vocabulary is bounded by the contents of the Gale Encyclopedia, which may not reflect current clinical guidelines, recently approved pharmacological agents, or emerging disease classifications. Second, the absence of dialogue history management means the system cannot maintain multi-turn conversational context, limiting utility for users requiring iterative clarification. Third, the system does not distinguish between lay and professional users, potentially resulting in responses that are insufficiently nuanced for clinicians.

Future Directions

Several avenues for improvement are recommended. Domain-adaptive fine-tuning on curated medical corpora would improve handling of specialized terminology and reduce context confusion. Integration of medical ontologies such as SNOMED CT or the Unified Medical Language System (UMLS) would improve entity disambiguation and enable more precise query expansion. Implementation of conversational memory through multi-turn context management would substantially enhance usability. Finally, a systematic user study involving both lay users and healthcare professionals would provide rigorous empirical evidence regarding real-world utility.

Ethical and Regulatory Considerations

The deployment of AI-powered medical chatbots raises important ethical questions. Foremost is the risk of users substituting chatbot guidance for professional medical consultation. The current system lacks explicit routing logic to identify high-acuity queries and direct users toward emergency services; implementing such logic is a priority for any production deployment. Data privacy requires robust minimization, anonymization, and retention policies compliant with applicable regulations (e.g., HIPAA, GDPR). Algorithmic fairness must also be assessed to ensure consistent accuracy across queries pertaining to conditions that disproportionately affect underrepresented populations.

CONCLUSIONS

This paper has presented the design, implementation, and evaluation of a Medical AI Chatbot that leverages retrieval-augmented generation to deliver accurate, source-grounded responses to natural-language medical queries. By combining OpenAI text embeddings with Pinecone vector storage and a Flask-based web interface, the system offers a technically robust and accessible platform for health education that can be extended, audited, and improved by the open-source community.

Empirical evaluation across four representative clinical domains demonstrated that the system performs reliably for broad and symptom-focused queries while revealing specific limitations in handling highly specialized or terminologically ambiguous inputs. These findings contribute constructively to the design literature for medical RAG systems and provide concrete targets for future development.

More broadly, this work affirms the transformative potential of AI-driven conversational agents to narrow the health information equity gap. As language models continue to improve and vector retrieval infrastructure matures, systems of this type will become increasingly capable partners in public health education—provided that their deployment is guided by rigorous evaluation, transparent communication of limitations, and unwavering commitment to user safety and data privacy.

REFERENCES

1. Teo, D. B., et al. (2021). The impact of ChatGPT on medical education: A review. Journal of Medical Education Technology, 23(4), 567–582.
2. Ali, H., et al. (2020). Foundation AI models in healthcare: Applications and opportunities. International Journal of Artificial Intelligence in Medicine, 15(3), 211–226.
3. Vignesh, R. (2019). Challenges in integrating AI into medical education: A systematic review. Journal of Medical Education

Research, 17(2), 145–160.
4. Altamimi, A. S., et al. (2022). Artificial intelligence (AI) chatbots in medicine: A supplement, not a substitute. Cureus, 29(1), 78–93.
5. Meskó, B. (2020). The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. Journal of Medical Internet Research, 8(2), 201–215.
6. Lee, H., Kang, J., & Yeo, J. (2021). Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: Development and deployment. Journal of Medical Internet Research, 23(5), e27460.
7. Rabie, A. H., et al. (2023). A new COVID-19 diagnosis strategy using a modified KNN classifier. Neural Computing and Applications.
8. An, Q., et al. (2023). A comprehensive review on machine learning in healthcare industry. Sensors, 23(9), 4178.
9. JavadiMoghaddam, S., & Gholamalinejad, H. (2021). A novel deep learning-based method for COVID-19 detection from CT image. Biomedical Signal Processing and Control, 70, 102987.
10. Huang, J., et al. (2022). Detection of diseases using machine learning image recognition technology. Computational Intelligence and Neuroscience, 2022, 5658641.
11. Park, C. W., et al. (2020). Artificial intelligence in health care: Current applications and issues. Journal of Korean Medical Science, 35(42), e379.
12. Patel, R., et al. (2019). Chatbots in healthcare: A systematic review. Journal of Artificial Intelligence in Medicine, 24(3), 321–335.
13. Sistaninejhad, B., et al. (2023). A review paper about deep learning for medical image analysis. Computational and Mathematical Methods in Medicine, 2023, 7091301.
14. Tassew, T., & Nie, X. (2022). A comprehensive review of the application of machine learning in medicine and health care. TechRxiv.
15. Miotto, R., et al. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.
16. Nagi, F., et al. (2023). Applications of artificial intelligence (AI) in medical education: A scoping review. Studies in Health Technology and Informatics, 305, 648–651.
17. Raghavendra, U., et al. (2019). AI techniques for automated diagnosis of neurological disorders. European Neurology, 82(1–3), 41–64.
18. Aljarah, A., & Alomari, G. (2021). Efficiency of using the Diffie–Hellman key in cryptography for internet security. Turkish Journal of Computer and Mathematics Education, 12(6), 2039–2044.
19. Aljrah, I., et al. (2023). Enhancing chip design performance with machine learning and PyRTL. International Journal of Intelligent Systems and Applications in Engineering, 12(2), 467–472.
20. Alomari, G., et al. (2023). Transforming text generation in NLP: Deep learning with GPT models. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 3139–3143.
21. Aljrah, I., et al. (2023). Machine learning in breast cancer diagnosis: Analyzing the Wisconsin dataset. Tuijin Jishu, 44(6).