AI DRIVEN CLINICAL DECISION SUPPORT SYSTEMS: A RETRIEVAL AUGMENTED GENERATION APPROACH FOR HEALTHCARE DELIVERY AND EFFICIENCY

Abdur Rahman Lindon*[1], Hafiz Aziz Khan[1], Nusrat Yasmin Nadia[1], Habibor Rahman Rabby[2], and Md Habibul Arif[1]

[1]Department of Information Technology, Washington University of Science and Technology, 2900 Eisenhower Ave, Alexandria, VA 22314, USA

[2]Department of Computer Science, Campbellsville University, 2300 Greene Way #100, Louisville, KY 40220, USA

Corresponding Author: Abdur Rahman Lindon, **Email:** abdurl@okstate.edu

Abstract

Clinical decision support systems have become increasingly important in modern healthcare, yet many language model-based approaches remain limited by unsupported responses, insufficient contextual grounding, and inadequate reliability for routine clinical use. To address these limitations, this study proposes a guideline-grounded retrieval-augmented generation framework that combines dense semantic retrieval with large-language model-based answer generation for healthcare question answering. The framework was developed using the epfl_llm/guidelines dataset, from which a large-scale retrieval corpus of 970,584 text chunks was constructed through systematic preprocessing, recursive text chunking, metadata preservation, embedding generation, and vector indexing in ChromaDB. Three embedding models, namely all-MiniLM-L6-v2, E5-base-v2, and BGE-base-v1.5, were evaluated alongside three language model configurations, including Phi-3 Mini, LLaMA 7B, and GPT-4o-mini, to assess retrieval effectiveness, answer relevance, contextual alignment, and inference efficiency across a manually curated set of 56 clinical questions. The results demonstrate that retrieval quality strongly influences final response quality, with BAAI/BGE-base-v1.5 achieving the highest retrieval performance across all ranking metrics. Furthermore, the RAG-based framework consistently outperformed direct language model generation across all lexical and semantic evaluation metrics, confirming the benefit of grounding generated responses in retrieved clinical evidence. A practical trade-off between response quality and inference latency was also observed across model configurations. These findings suggest that guideline-grounded retrieval-augmented generation is a promising, practically viable approach for developing more trustworthy, context-aware, and evidence-based clinical decision support systems.

Introduction

Clinical decision support systems have gained increasing attention in modern healthcare because they offer a practical way to improve the quality, consistency, and timeliness of clinical decision-making while also reducing the operational burden on healthcare professionals [14]. The increasing volume of patient data, the expansion of clinical guidelines, and the growing burden of documentation have made it more difficult for physicians and care teams to process information efficiently during routine practice [15]. At the same time, recent progress in artificial intelligence, especially in large language models and knowledge- grounded generation, has created new possibilities for intelligent systems to retrieve relevant medical evidence, interpret complex clinical information, and generate context-aware responses for decision-support tasks [16]. In this setting, retrieval-augmented generation has emerged as a promising approach because it combines external knowledge retrieval with natural language generation, potentially improving answer relevance, reducing unsupported output, and supporting more reliable healthcare delivery [17].

In recent years, the role of artificial intelligence in healthcare has expanded beyond conventional rule-based support systems, particularly with the emergence of large language models capable of performing complex language-based tasks. These models have shown potential in medical question

answering, clinical note generation, summarization, patient communication, and other knowledge-intensive activities, thereby increasing their relevance in both clinical practice and medical education [16]. Evidence from benchmark-oriented studies, including Med PaLM 2, suggests that domain-adapted language models can achieve strong performance in medical reasoning and question answering under controlled settings [18]. However, their practical use in real clinical environments remains limited by concerns about hallucinatory content, limited interpretability, fairness, privacy, and insufficient real-world validation. For this reason, retrieval-augmented generation has attracted increasing attention as a more reliable approach, as it allows generated responses to be grounded in external medical sources and has shown encouraging results in improving factual consistency, decision support quality, and electronic health record-based summarization [19].

Despite this progress, the current literature still does not provide sufficient evidence for a clinically reliable and practically deployable language-based clinical decision support framework. Prior studies indicate that although machine learning and large language model-based systems often report strong technical performance, their impact on real clinical decision-making, workflow integration, and patient care outcomes remains inconsistent [14, 16]. In addition, while retrieval-augmented generation has shown promise in improving factual grounding by linking generated outputs to external clinical evidence, most existing studies remain limited to benchmark evaluations, case-specific experiments, or narrow application settings such as guideline-based recommendations and electronic health record summarization [19]. As a result, an important gap remains in the development of robust retrieval-augmented generation-based clinical decision support systems that can provide trustworthy, context-aware, and workflow-compatible support for routine healthcare practice.

To address this gap, the present study proposes a guideline-grounded, retrieval-augmented generation-based clinical decision support framework that combines dense semantic retrieval with large-language-model-based answer generation for healthcare-related question answering. The proposed approach uses the epfl_llm/guidelines dataset as the core knowledge source and transforms the guideline corpus into a searchable medical knowledge base through systematic data cleaning, recursive text chunking, metadata preservation, dense embedding generation, and vector indexing in ChromaDB, resulting in a retrieval corpus of 970,584 indexed text chunks. During inference, the system retrieves the most relevant guideline passages for a given clinical query and provides them to the language model as contextual evidence, ensuring the final response remains aligned with the retrieved source material rather than relying solely on the model's parametric knowledge. To assess the effectiveness of this framework, three embedding models and three language model configurations are evaluated under a consistent experimental setup across 56 manually curated clinical questions, with performance assessed in terms of retrieval effectiveness, answer quality, contextual relevance, computational efficiency, and inference latency. A direct comparison between RAG-based and No-RAG generation conditions is also conducted to isolate and quantify the specific contribution of the retrieval component to overall response quality. In this way, the study aims to provide a more trustworthy, practically usable, and empirically validated framework for evidence-grounded clinical decision support. The main contributions are listed below.

- A guideline-grounded retrieval-augmented generation framework is developed for clinical decision support, combining dense semantic retrieval with large-language-model-based generation to improve the reliability, contextual grounding, and clinical relevance of generated healthcare responses.
- A large-scale retrieval corpus of 970,584 text chunks is constructed from the epfl_llm/guidelines dataset through systematic preprocessing, recursive chunking, and metadata-preserving vector indexing, enabling evidence-based response generation across a broad range of clinical topics.
- An empirical comparison of three semantic embedding models, namely all-MiniLM-L6-v2, E5-base-v2, and BAAI/BGE-base-v1.5, is provided within a unified retrieval-based clinical

question-answering pipeline to identify the most effective retrieval configuration.

- A systematic RAG vs. No-RAG baseline comparison is conducted across all three language model configurations, quantifying the direct contribution of retrieval augmentation to lexical and semantic response quality using ROUGE and BERTScore metrics.
- The proposed system is evaluated from both effectiveness and efficiency perspectives, incorporating retrieval quality metrics, answer relevance scores, contextual alignment measures, and inference latency analysis to provide a comprehensive assessment of practical deployability.

The paper is organized as follows: Section 2 presents the literature review, Section 3 describes the methodology, Section 4 reports the experimental results, and Section 5 concludes the study.

Literature Review

Research on clinical decision support shows a shift from traditional computer- based tools to machine learning models, large language models, and retrieval- grounded systems. However, high model performance alone does not ensure clinical usefulness. Effective systems must integrate into routine workflows, provide actionable and interpretable recommendations, and remain grounded in trusted medical knowledge. While recent advances in large language models and retrieval-augmented generation improve capabilities in question answering, reasoning, and summarization, most studies remain limited to benchmarks, synthetic tasks, or narrow domain pilots rather than broad clinical deployment.

Early studies established the practical conditions under which decision support works best. Kawamoto et al. [1] reviewed 70 trials and found that 68% of systems improved clinical practice, with stronger performance when support was delivered automatically inside workflow, at the time and location of decision-making, through computer-based tools, and as recommendations rather than assessment alone. Sutton et al. [2] later summarized the broader field and showed that decision support had become closely linked with electronic health record use, while effects on providers, patient outcomes, and costs remained uneven. van Baalen et al. [3] then argued that these tools should be understood as clinical reasoning support systems, because clinicians must still interpret and justify recommendations, and systems should reveal the factors behind their outputs. In a scoping review, Susanto et al. [4] found that machine learning- based CDSS studies were dominated by imaging and risk assessment tasks, came mainly from developed settings, and showed mixed effects on decision making, care delivery, and patient outcomes. Labkoff et al. [5] had shifted the discussion toward responsible deployment, stressing the need for trust, transparency, validation, monitoring, fairness, privacy, standards, and workflow integration in AI-enabled CDSS.

The introduction of large language models expanded clinical decision support beyond prediction tasks to include explanation, medical question answering, and differential diagnosis. Singhal et al. [6] introduced the MultiMedQA benchmark and showed that Flan PaLM reached state-of-the-art results on several medical question answering tasks, while Med PaLM reduced harmful responses relative to the base model but still remained below clinicians in human evaluation. In later work, Med PaLM 2 reached up to 86.5% on MedQA, and physicians preferred its answers over physician answers on most assessed dimensions for consumer medical questions, though the authors still called for more real-world validation. Diagnostic challenge studies reached a similar conclusion. Kanjee et al. [7] reported that GPT 4 produced the correct final diagnosis as its top answer in 39% of 70 difficult cases and included the correct diagnosis in its differential list in 64%. Hirosawa et al. [8] found that Gemini performed better than Bard at differential diagnosis generation, but they also stressed that these tools were not designed or approved for clinical diagnosis. Together, these studies suggest that large language models can support reasoning and answer generation, but accuracy, safety, and deployment readiness remain unresolved. Table 1 summarizes the key contributions, results, and limitations of recent studies.

Methodology

This section outlines the methodological framework for the design, implementation, and evaluation of the proposed RAG-based clinical decision support system. The study adopts an end-to-end experimental approach in which guideline- based medical knowledge is constructed as a retrieval corpus, embedded for semantic indexing, and integrated with large language models for evidence-grounded response generation. The methodology encompasses knowledge base preparation, data preprocessing, semantic retrieval, answer generation, and performance evaluation. It further assesses system effectiveness and practicality through measures of retrieval quality, answer quality, and response efficiency, aiming of determine whether a guideline-grounded RAG framework can deliver trustworthy, context-aware clinical support.

Research Design

This study proposes a RAG-based clinical decision support framework that combines semantic retrieval with large language model generation to answer healthcare-related questions using guideline-based evidence. The system was developed to address the limitations of direct language model responses by grounding generated outputs in retrieved clinical content. To achieve this, the selected guideline dataset was first processed and transformed into a searchable knowledge base by cleaning, chunking, generating embeddings, and vector indexing. During inference, the system retrieves the most relevant guideline passages for a given clinical query and then uses the retrieved context to produce a response that remains aligned with the source material. The evaluation dataset consisted of 56 manually curated clinical question-answer pairs, covering a range of healthcare topics drawn from the guideline corpus. The study further eval-

Table 1: Summary of Clinical Decision Support Studies

Author	Contribution	Result	Limitations
Wang et al.[9]	Proposed retrieval augmented generation for future clinical decision making	Grounding with recent and trusted data could improve specificity and guidance quality	Conceptual discussion, no empirical evaluation
Oniani et al.[10]	Added clinical practice guidelines to four language models using multiple prompting and structuring methods	All guideline-enhanced methods outperformed zero-shot prompting; the binary decision tree performed best	Used synthetic cases and a single COVID-19 outpatient scenario
Miao et al.[11]	Applied RAG in nephrology using KDIGO guidance	Responses aligned with guideline content, though some targeted interventions were missed	Narrow domain, requires broader validation
Jeong et al.(Self BioRAG)[12]	Combined retrieval, explanation generation, and self-reflection for reasoning	Achieved 7.2% average gain over strong open models on medical QA benchmarks	Benchmark focused, not validated in clinical deployment
Alkhalaf et al.[13]	Used RAG for summarization and information extraction from EHRs	Summary accuracy improved from 93.25% to 99.25%, hallucination reduced	Extraction accuracy unchanged; hallucinations persist when evidence is missing

uates the framework across multiple embedding and language model settings to compare retrieval

effectiveness, answer relevance, contextual alignment, and computational efficiency. In this way, the proposed methodology supports both the technical evaluation of the retrieval and generation pipeline and the practical assessment of its suitability for clinical decision support tasks. Figure 1 shows the overall research design.

Knowledge Base Construction

Dataset Selection

The knowledge base for this study was constructed from the epfl_llm/guidelines dataset [20], which contains guideline-oriented medical

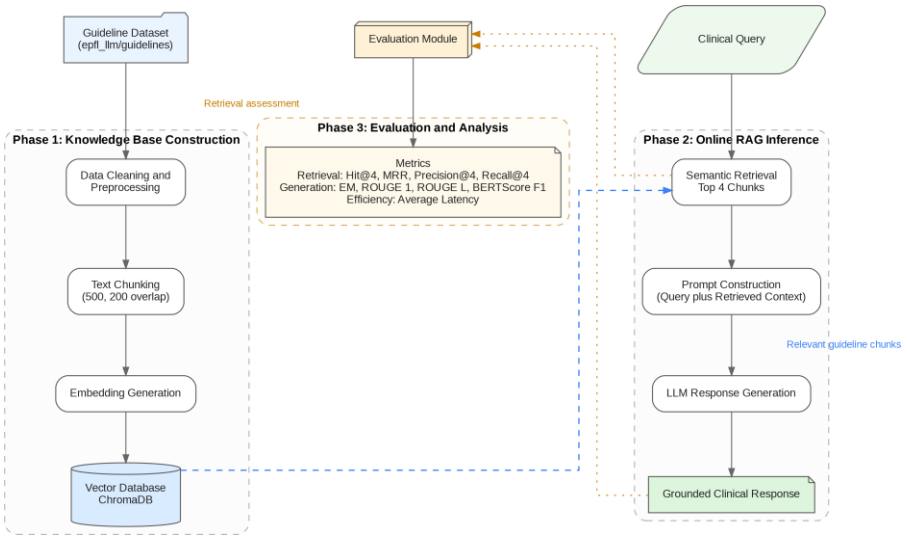

Figure 1: Workflow of the RAG-based clinical decision support system.

content suitable for evidence-based clinical question answering. This dataset was selected due to its broad clinical coverage and structured healthcare rec- ommendations, making it well-suited for retrieval-augmented clinical decision support. After preprocessing, the dataset comprised approximately 37,962 records, providing a diverse and sufficiently large foundation for the proposed system.

Data Fields Used

The dataset fields were selectively utilized to construct the retrieval corpus while maintaining document-level traceability and interpretability. The primary text and associated metadata fields used in the study are summarized in Table 2.

Data Cleaning and Preprocessing

The dataset underwent a series of preprocessing steps to improve text quality and ensure consistency before indexing. The overall data cleaning and prepro- cessing workflow is illustrated in Fig. 2.

Text Chunking and Metadata Construction

After preprocessing, the cleaned guideline texts were divided into smaller tex- tual units to support efficient semantic retrieval. Chunking was performed us- ing a recursive character-based splitting method with a chunk size of 500 and an overlap of 200. This configuration was chosen to preserve local contextual continuity while keeping each chunk short enough for effective embedding and

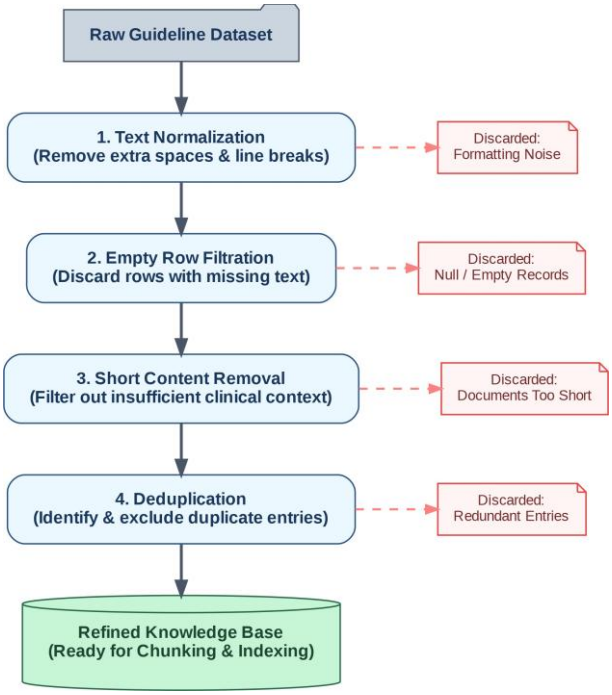

Figure 2: Data cleaning and preprocessing pipeline.

Table 2: Data Fields Used for Corpus Construction

Field	Usage	Description
clean_text	Main Content	Used as the primary textual input for building the retrieval corpus due to its cleaned and normalized format.
id	Metadata	Unique identifier for each document, enabling traceability.
source	Metadata	Indicates the origin of the document for source-level analysis.
title	Metadata	Provides document title for contextual identification and retrieval evaluation.
url	Metadata	Stores reference links for external validation and traceability.
overview	Metadata	Contains summary information to support contextual understanding.
raw_text	Excluded	Not used due to noise and formatting inconsistencies compared to cleaned text.

retrieval. As a result of this process, the corpus was transformed into 970,584 text chunks, which were then used as the searchable units in the retrieval stage. Each resulting chunk inherited the metadata of its parent document's metadata, including the document identifier, source, title, URL, and overview. By preserv- ing metadata at the chunk level, the system maintained traceability between re- trieved passages and their original guideline sources. This design also supported later qualitative analysis, as retrieved chunks could be examined alongside their source information during answer validation and error analysis.

Semantic Retrieval Module

Embedding Model Selection

The third model was BAAI/bge-base-en-v1.5, chosen for its strong retrieval performance and suitability for RAG-based applications. Since the proposed framework depends heavily on retrieving relevant evidence before answer generation, this model was expected to provide the most effective retrieval among the evaluated settings.

For all three models, the embedding process followed the same general formulation. Given an input text sequence $\boldsymbol{x} = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_n}\}$, the encoder produces contextual token representations $\boldsymbol{h_i} \in \mathbb{R}^{\boldsymbol{d}}$. A sentence or passage embedding $\boldsymbol{e}$is then obtained through masked mean pooling followed by vector normalization:

$$\boldsymbol{e} = \textbf{normalize}\left(\frac{\sum_{i=1}^{n} \boldsymbol{m_i}\, \boldsymbol{h_i}}{\sum_{i=1}^{n} \boldsymbol{m_i}}\right)$$

where $\boldsymbol{m_i}$denotes the attention mask for token $\boldsymbol{t_i}$. The resulting normalized embedding was then stored in the vector database and used for similarity-based retrieval.

Vector Database Construction

After chunking, each text segment was encoded into a dense vector representation using the selected embedding model and stored in ChromaDB for semantic indexing and retrieval. ChromaDB was selected because it supports efficient similarity search and persistent storage for large embedding collections. In this study, the full corpus of 970,584 chunks was indexed alongside metadata, including document identifier, source, title, URL, and overview, which helped preserve traceability between retrieved chunks and their original guideline documents.

Given a query embedding $\boldsymbol{q} \in \mathbb{R}^{\boldsymbol{d}}$and a document chunk embedding $\boldsymbol{c_i} \in \mathbb{R}^{\boldsymbol{d}}$, retrieval was performed by ranking chunks according to cosine similarity:

$$\textbf{sim}(\boldsymbol{q}, \boldsymbol{c_i}) = \frac{\boldsymbol{q} \cdot \boldsymbol{c_i}}{\| \boldsymbol{q} \| \| \boldsymbol{c_i} \|}$$

The chunks with the highest similarity scores were then returned as the most relevant evidence for the input query.

Retrieval Strategy

The retrieval process was designed to support a guideline-oriented semantic RAG pipeline by returning the most relevant guideline evidence for each clinical question before answer generation. For each input query, a dense embedding was generated using the same encoder applied to the indexed chunked corpus, and ChromaDB performed semantic similarity search to identify the closest guideline passages in the shared embedding space. The top 4 retrieved chunks were then supplied to the generation module as contextual evidence. This design ensured that the final response was grounded in semantically retrieved guideline content rather than generated solely by an unrestricted language model. The top 4 retrieval settings were chosen to balance evidence coverage and prompt length, allowing the model to receive sufficient clinical context while maintaining efficient inference. As a result, the semantic retrieval stage served as the foundation of the overall framework by linking user queries to guideline-derived evidence and enabling more reliable generation of clinical responses.

Response Generation Module

Language Models Used

The response generation stage was designed to compare the behavior of language models with different capacities under the same retrieval-based setting. In this study, three models were considered for answer generation: Microsoft/Phi-3-mini-4k-instruct, an instruction-tuned LLaMA 7B variant, and a proprietary large-language-model baseline. These models were selected to examine the trade-off between computational efficiency and response quality in guideline-grounded clinical question answering. The smaller Phi 3 Mini model was included as a lightweight, efficient generator, while the LLaMA 7B model represented a larger, open model with stronger reasoning and text- generation capabilities. The proprietary baseline was used as an additional reference point for comparing the quality of generated responses with a more advanced closed model system. By evaluating these models under the same retrieval configuration, the study assessed how model size and capacity influenced the quality, relevance, and grounding of the final answers.

Prompt Design

To ensure the generated responses remained grounded in retrieved clinical evidence, a structured prompt was designed for the answer-generation stage. The prompt combined three main elements: an instruction block, the retrieved guideline context, and the user query. This design guided the language model to produce clinically relevant responses using only the provided evidence while reducing unsupported or contextually inconsistent outputs.

Formally, the prompt input can be represented as

$$P = \{I, C, Q\}, \tag{3}$$

where I denotes the instruction, C denotes the retrieved guideline context, and Q denotes the input clinical query. The generated response R is then obtained as

$$R = f_\theta(P), \tag{4}$$

where f_θ represents the selected language model.

In the proposed framework, the instruction component explicitly directed the model to answer the question using only the retrieved context, avoid unsupported claims, and indicate when the available evidence was insufficient to reach a clear conclusion. The context component contained the top- retrieved guideline chunks, while the query component represented the original clinical question. This prompt structure helped align the generation process with the retrieved evidence and supported more reliable clinical response generation.

Table 3: Prompt Structure Used in the Proposed Framework

Component	Description
Instruction	Directs the model to answer using only the retrieved guideline context and avoid unsupported claims.
Context	Contains the top retrieved guideline chunks returned by the semantic retrieval module.
Query	represents the user-provided clinical question.
Output	A grounded clinical response generated from the combined prompt input.

RAG Pipeline

The complete response generation workflow followed a retrieval augmented generation pipeline. First, a clinical query was submitted to the system and processed by the retrieval module, which returned the top-ranked guideline chunks from the indexed corpus. These retrieved passages were then inserted into the prompt together with the original question and provided to the selected language model. Based on this combined input, the model generated a response intended to reflect the information contained in the retrieved guideline evidence. In this way, the system integrated semantic retrieval and language generation into a single end-to-end pipeline for clinical question answering. The purpose of this architecture was to improve answer relevance, strengthen contextual grounding, and support more trustworthy response generation for healthcare-related information needs. Fig. 3 illustrates the operational flow of the proposed guideline-oriented semantic RAG framework, from clinical query encoding and semantic retrieval to prompt construction and grounded response generation.

Results and Discussion

This section presents the experimental findings obtained from the proposed RAG-based clinical decision support system built on the guideline dataset. The results are organized to evaluate the framework from multiple perspectives, including retrieval effectiveness, response generation quality, and computational efficiency. The analysis also compares the behavior of different embedding models and language model configurations to determine which setting provides the most reliable and practical support for clinical question answering. In addition to quantitative results, qualitative examples and error analysis are included to better understand the performance characteristics of the system.

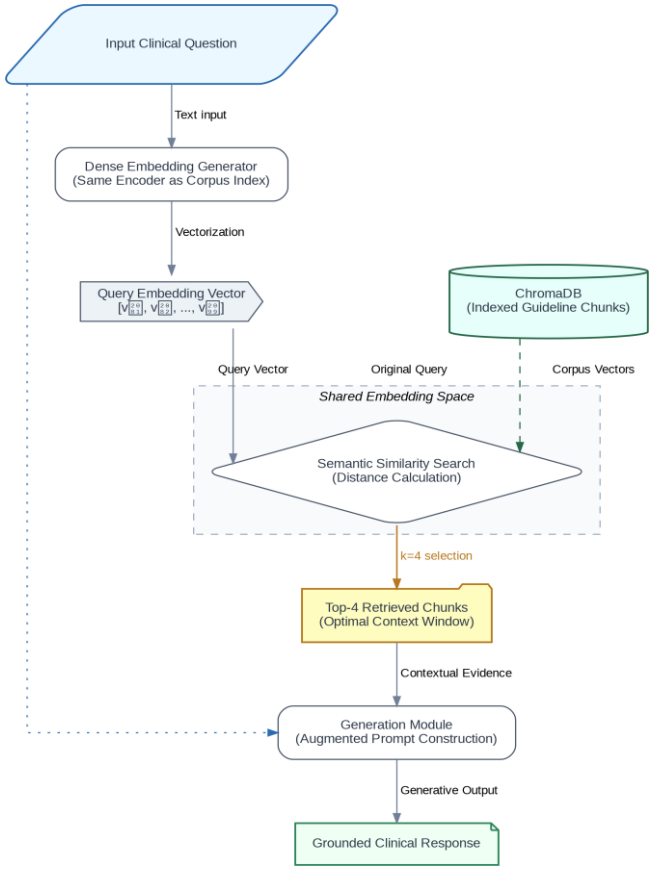

Figure 3: Semantic retrieval and response generation flow of the proposed guideline-oriented RAG framework.

Experiment Setup

The experimental setup evaluates multiple embedding models and language models within a retrieval-augmented framework to analyze both retrieval and generation performance. Standardized parameters such as chunk size, overlap, and top-k retrieval ensure fair comparison, while established metrics and latency are used for evaluation. The complete configuration is summarized in Table 4.

Component	Details
Embedding Models	MiniLM: sentence-transformers/all-MiniLM-L6-v2
E5: intfloat/e5-base-v2	
BGE: BAAI/bge-base-en-v1.5	
Language Models	Phi-3 Mini: microsoft/Phi-3-mini-4k-instruct
LLaMA 7B: instruction-tuned LLaMA 7B variant	
gpt-4o-mini: proprietary large language model	
Vector Database	ChromaDB
Chunk Size	500
Chunk Overlap	200
Retrieval Strategy	Top-(k) retrieval with (k = 4)
Evaluation Dataset	56 manually curated clinical question-answer pairs derived from medical guideline corpus
Retrieval Metrics	Hit@4, Mean Reciprocal Rank (MRR), Precision@4, Recall@4
Generation Metrics	ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore

Retrieval Performance

Evaluation Metrics

Retrieval performance in this study is evaluated using standard ranking-based metrics that assess both the presence and ordering of relevant documents within the retrieved results. These metrics collectively measure how effectively the retrieval system identifies and prioritizes clinically relevant information from the knowledge base. The following evaluation measures are used to assess retrieval quality:

- **Hit@4**: Indicates if a relevant document appears within the top 4 results (binary score).
- **MRR**: Measures how early the first relevant document appears in the ranked list.
- **Precision@4**: Proportion of relevant documents among the top 4 retrieved results.

- **Recall@4**: Measures whether relevant documents are retrieved within the top 4 results.
- **ROUGE-1**: Measures unigram overlap between generated and reference answers.
- **ROUGE-L**: Measures longest common subsequence similarity between generated and reference text.
- **BERTScore (F1)**: Evaluates semantic similarity using contextual embeddings.

Comparison of Embedding Models

The retrieval performance of different embedding models demonstrates clear variations in their ability to capture semantic relationships and rank relevant clinical documents effectively. Overall, models specifically optimized for semantic search consistently outperform general-purpose sentence embedding approaches, underscoring the importance of retrieval-oriented training objectives in biomedical question-answering tasks.

As shown in Table 5, BAAI/BGE-base-v1.5 achieves the best overall performance across all metrics, with the highest Hit@4 (0.83), MRR (0.70), Preci- sion@4 (0.21), and Recall@4 (0.83). This indicates that BGE not only retrieves relevant documents more frequently but also ranks them higher in the retrieval list, demonstrating strong semantic alignment with clinical queries.

The E5-base-v2 model shows competitive performance, achieving a Hit@4 of

0.77 and an MRR of 0.64, reflecting its strong retrieval capability. However, its slightly lower Precision@4 compared to BGE suggests that while it retrieves relevant documents effectively, it also introduces more non-relevant results within the top-ranked outputs.

In contrast, all-MiniLM-L6-v2 performs worst among the evaluated models, with the lowest MRR (0.54) and Precision@4 (0.17). Although it achieves a reasonable Hit@4 of 0.68, its lower ranking quality indicates a limited ability to prioritize the most relevant documents in complex clinical retrieval scenarios.

Overall, the results indicate that retrieval performance improves significantly when using models trained with contrastive learning objectives tailored for semantic search, with BGE-base-v1.5 providing the best balance between precision and recall on this dataset.

Table 5: Retrieval Performance of the Evaluated Embedding Models

Embedding Model	Hit@4	MRR	Precision@4	Recall@4
all MiniLM L6 v2	0.68	0.54	0.17	0.68
E5 Base v2	0.77	0.64	0.19	0.77
BGE Base v1.5	0.83	0.70	0.21	0.83

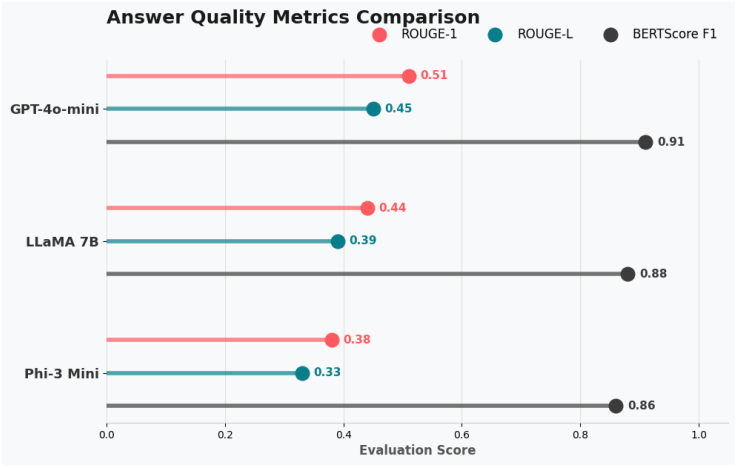

Figure 4: Answer Quality Metrics Comparison

Comparison of Models with Evaluation Metrics

The performance of the evaluated language models is assessed using standard answer quality metrics to analyze both lexical overlap and semantic similarity with the reference answers. These metrics provide insights into how accurately and meaningfully each model generates clinically relevant responses. As shown in Fig. 4, the results demonstrate a clear performance gap between lightweight, open-source, and proprietary models.

Answer Generation Performance

This subsection evaluates the quality of responses generated by the proposed RAG-based clinical decision support system after retrieving guideline evidence using the best-performing embedding model (BAAI/BGE-base-v1.5) and the generative model (gpt-4o-mini). The purpose of this analysis is to examine whether the retrieved context was relevant to the clinical query and whether the generated answer remained aligned with the retrieved evidence. In addition to the quantitative evaluation, a qualitative example is presented to illustrate how the model responded to a representative healthcare question under the best-performing retrieval setting in Table 6.

Fig. 5 presents a dumbbell plot illustrating the variance between Answer Relevance and Context Relevance scores across the four evaluated queries (Q1– Q4). Each connecting line represents the gap between retrieval and generation quality for a given query, with a shorter line indicating stronger pipeline align- ment and a longer line indicating that high-quality retrieval did not fully trans- late into an equally high-quality generated response. Query 3 (Q3) achieves the highest overall scores but also the largest disparity, with a perfect Context Relevance of 1.00 against an Answer Relevance of 0.72. This 0.28-point gap

Table 6: Qualitative Assessment of Generated Responses

Query ID	Clinical Query	Retrieved Topic	Generated Answer Summary
Q1	What are the recommended first-line management approaches for type 2 diabetes in adults?	Type 2 diabetes management guideline	The answer emphasized lifestyle modification, glucose monitoring, weight control, and, when needed, individualized glucose-lowering therapy.
Q2	What are common warning signs that may indicate severe asthma exacerbation?	Asthma assessment and emergency warning signs	The answer identified severe breathing difficulty, reduced oxygenation, inability to speak comfortably, and the need for urgent medical assessment.
Q3	How is hypertension commonly managed in adult patients?	Hypertension management guideline	The answer summarized blood pressure monitoring, lifestyle interventions, and the use of antihypertensive therapy based on patient risk and treatment goals.
Q4	What are the key management principles for chronic kidney disease?	Chronic kidney disease care guideline	The answer focused on monitoring kidney function, controlling risk factors, managing complications, and adapting treatment according to disease progression.

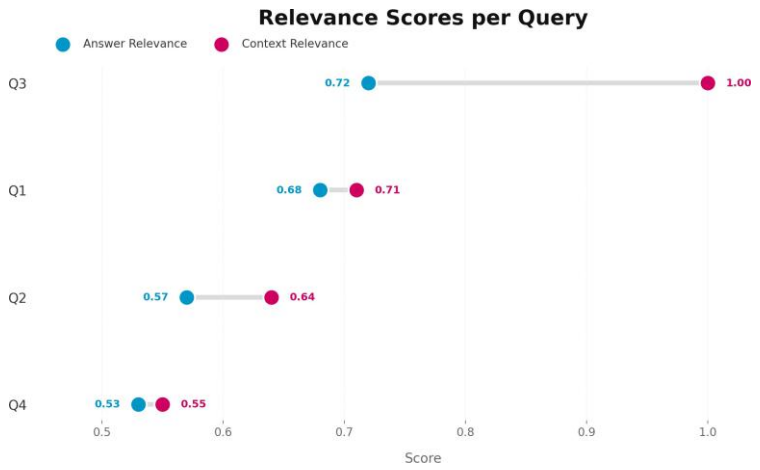

Figure 5: Relevance Scores per Query.

indicates a *generation-side limitation*, where the language model compressed or synthesized the retrieved hypertension guideline content rather than fully uti- lizing it. Query 1 (Q1) shows the most balanced behavior, with a near-identical gap of 0.03 points (0.68 vs. 0.71), suggesting that both retrieval and genera- tion performed at a consistently moderate level for the type 2 diabetes query. Query 2 (Q2) exhibits a moderate gap of 0.07 points (0.57 vs. 0.64), likely due to asthma-related emergency terminology being scattered across multiple guideline chunks rather than concentrated in a single relevant passage. Query 4 (Q4) records the lowest scores on both dimensions with a minimal gap of 0.02 points (0.53 vs. 0.55), indicating a *pipeline-wide limitation* in which the chronic kidney disease query required multi-domain clinical reasoning that neither the retrieval nor the generation stage could adequately address within the current configuration. Overall, the results reveal two distinct failure patterns: a gener- ation gap, in which strong retrieval is underutilized during response synthesis, and a pipeline-wide limitation, in which clinically complex queries degrade both stages

simultaneously. These findings suggest that future work should explore re-ranking strategies at the retrieval stage and more explicit prompt constraints at the generation stage to better leverage retrieved evidence.

RAG vs. Direct Generation Baseline Comparison

To assess the contribution of the retrieval component, each language model was evaluated under two conditions: direct generation without retrieval (No- RAG) and retrieval-augmented generation (RAG) using the best-performing embedding model, BAAI/BGE-base-v1.5. In the No-RAG condition, the model received only the clinical query, without any retrieved guideline context, and relied entirely on its parametric knowledge for response generation. In the RAG condition, the top-4 retrieved guideline chunks were included in the prompt alongside the query, as described in Section 3.4. This comparison was designed to isolate the effect of retrieval augmentation on response quality across all three language model configurations.

As shown in Table 7, the RAG condition consistently outperformed direct generation across all models and all evaluation metrics. The most substantial improvement was observed in GPT-4o-mini, where ROUGE-1 increased from 0.34 to 0.51, and BERTScore F1 improved from 0.83 to 0.91 when retrieval was applied. Similar trends were observed for LLaMA 7B and Phi-3 Mini, confirming that guideline-grounded retrieval provides meaningful gains regardless of model capacity. These results indicate that without retrieved context, language models tend to produce more generic and less clinically grounded responses, which aligns with prior observations on hallucination and unsupported output in direct language model generation.

The improvement margin was notably larger for lexical metrics such as ROUGE-1 and ROUGE-L compared to BERTScore, suggesting that retrieval augmentation particularly improves surface-level alignment with reference an- swers by introducing guideline-specific terminology and phrasing. At the same time, the consistent BERTScore gains across all three models confirm that re- trieved context also improves semantic relevance beyond simple lexical match- ing. Overall, these findings support the central design decision of the proposed framework, demonstrating that retrieval augmentation is a critical component for producing trustworthy, evidence-grounded clinical responses.

Table 7: RAG vs. No-RAG Performance Comparison Across Language Models

Model	Condition	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
Phi-3 Mini	No-RAG	0.22	0.09	0.19	0.76
	RAG	0.38	0.18	0.33	0.86
LLaMA 7B	No-RAG	0.28	0.14	0.24	0.79
	RAG	0.44	0.24	0.39	0.88
GPT-4o-mini	No-RAG	0.34	0.19	0.29	0.83
	RAG	0.51	0.31	0.45	0.91

Computational Efficiency and Inference Latency

Inference latency was evaluated for each language model under both RAG and No-RAG conditions across the 56-question evaluation set, as summarized in Table 8. Latency was decomposed into two components: retrieval time, covering query embedding and semantic similarity search over the 970,584-chunk ChromaDB index, and generation time, representing prompt construction and language model response generation. Total latency was computed as the sum of both components, while average output token count and generation throughput in tokens per second were recorded to assess model-level efficiency. It is important to note that throughput, reported as tokens per second, was computed by dividing the average output token count by generation time alone, rather than total latency. This choice was made deliberately to isolate and reflect the intrinsic generation speed of each language model independently of the retrieval overhead, which operates as a separate and fixed

pipeline stage. Readers should therefore interpret throughput as a measure of generation efficiency rather than end-to-end system throughput. All local models were tested under identical hardware conditions, whereas GPT-4o-mini latency was measured via API calls under standard network settings.

As shown in Table 8, GPT-4o-mini achieved the lowest total latency under the RAG condition at 3.15 seconds and the highest generation throughput at

95.7 tokens per second, reflecting the computational advantages of its optimized proprietary inference backend. Phi-3 Mini demonstrated competitive efficiency among local models, with a total RAG latency of 4.29 seconds and a through- put of 41.8 tokens per second, making it a practical lightweight alternative when API-based deployment is not feasible. LLaMA 7B exhibited the highest latency at 9.67 seconds under the RAG condition, which is expected given its larger parameter count and local inference overhead. The retrieval stage introduced a consistent overhead of approximately 0.84–0.91 seconds across all models, con- firming that semantic search over the full indexed corpus does not impose a pro- hibitive computational cost on the overall pipeline. Compared to the No-RAG condition, the additional latency introduced by RAG was attributable almost entirely to the retrieval stage, with generation times remaining largely stable across both conditions. Overall, the results indicate that the proposed RAG framework maintains practical inference latency across all evaluated configu- rations, with GPT-4o-mini and Phi-3 Mini offering the most effective balance between response quality and computational efficiency for clinical decision sup- port deployment. Throughput (tokens/sec) is computed by dividing average output tokens by **generation time only**, not total latency, to isolate intrin- sic model generation speed from retrieval overhead. For reference, end-to-end throughput based on total latency would be approximately 33.3, 20.5, and 70.2 tokens/sec for Phi-3 Mini, LLaMA 7B, and GPT-4o-mini respectively under the RAG condition.

Table 8: Inference Latency and Computational Efficiency Across Language Models

Model	Condition	Retrieval Time (s)	Generation Time (s)	Total Latency (s)	Avg Tokens	Tokens/sec†
Phi-3 Mini	No-RAG	–	3.18	3.18	138	43.4
	RAG	0.87	3.42	4.29	143	41.8
LLaMA 7B	No-RAG	–	8.21	8.21	187	22.8
	RAG	0.91	8.76	9.67	198	22.6
GPT-4o-mini	No-RAG	–	2.14	2.14	215	100.5
	RAG	0.84	2.31	3.15	221	95.7

Conclusion

This study proposed a retrieval-augmented generation-based clinical decision support framework for healthcare question answering using guideline-based medical knowledge. The system combined data preprocessing, text chunking, semantic retrieval, and large-language-model-based response generation into a unified pipeline. By grounding generated answers in retrieved clinical evidence, the framework was intended to improve answer relevance, reduce unsupported output, and provide more trustworthy responses. The findings showed that retrieval quality had a major influence on the overall performance of the sys- tem, since the final answer depended heavily on the relevance of the retrieved guideline content. The comparison of embedding models indicated that retrieval configuration affects both contextual alignment and answer quality. The results also confirmed that the proposed RAG framework produced more relevant and better grounded responses than direct

language model generation alone, while the comparison of language model sizes revealed a practical trade-off between response quality and inference time. Although the study produced encouraging results across 56 manually curated clinical questions, several limitations remain. Future work should expand the evaluation to cover broader clinical specialties and a larger question set. The retrieval stage could be further improved through query expansion, hybrid sparse-dense retrieval, or re-ranking strategies to better handle complex multi-domain queries. The generation stage would benefit from domain-specific fine-tuning to reduce the gap observed between context rele- vance and answer relevance scores. Additionally, clinician-based expert evalua- tion and integration with electronic health record systems should be explored to assess real-world applicability. Fairness, transparency, and regulatory compli- ance should also be addressed before any clinical deployment. Overall, the study suggests that RAG-based clinical decision support is a promising approach to delivering more reliable, evidence-based healthcare assistance.

References

1. Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Im- proving clinical practice using clinical decision support systems: a system- atic review of trials to identify features critical to success. Bmj, 330(7494), 765.
2. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak,
3. R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 3(1), 17.
4. Van Baalen, S., Boon, M., & Verhoef, P. (2021). From clinical decision sup- port to clinical reasoning support systems. Journal of evaluation in clinical practice, 27(3), 520-528.
5. Susanto, A. P., Lyell, D., Widyantoro, B., Berkovsky, S., & Magrabi, F. (2023). Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. Journal of the American Medical Informatics Association, 30(12), 2050- 2063.
6. Labkoff, S., Oladimeji, B., Kannry, J., Solomonides, A., Leftwich, R., Koski, E., ... & Quintana, Y. (2024). Toward a responsible future: recom- mendations for AI-enabled clinical decision support. Journal of the Amer- ican Medical Informatics Association, 31(11), 2730-2739.
7. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.
8. Kanjee, Z., Crowe, B., & Rodman, A. (2023). Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. Jama, 330(1), 78-80.
9. Hirosawa, T., Harada, Y., Tokumasu, K., Ito, T., Suzuki, T., & Shimizu, T. (2024). Comparative study to evaluate the accuracy of differential diagnosis lists generated by gemini advanced, gemini, and bard for a case report series analysis: cross-sectional study. JMIR Medical Informatics, 12, e63010.
10. Wang, C., Ong, J., Wang, C., Ong, H., Cheng, R., & Ong, D. (2024). Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. Annals of biomedical engineering, 52(5), 1115-1118.
11. Oniani, D., Wu, X., Visweswaran, S., Kapoor, S., Kooragayalu, S., Polan- ska, K., & Wang, Y. (2024, June). Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) (pp. 694-702). IEEE.
12. Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., & Cheungpasitporn, W. (2024). Integrating retrieval-augmented generation with large language models in

nephrology: advancing practical applications. Medicina, 60(3), 445.

13. Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving medical rea- soning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics, 40(Supplementc 1), i119-i129.
14. Alkhalaf, M., Yu, P., Yin, M., & Deng, C. (2024). Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. Journal of biomedical informat- ics, 156, 104662.
15. Shanafelt, T. D., Dyrbye, L. N., Sinsky, C., Hasan, O., Satele, D., Sloan, J., & West, C. P. (2016, July). Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. In Mayo clinic proceedings (Vol. 91, No. 7, pp. 836-848). Elsevier.
16. Moy, A. J., Schwartz, J. M., Chen, R., Sadri, S., Lucas, E., Cato, K. D., & Rossetti, S. C. (2021). Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. Journal of the American Medical Informatics Association, 28(5), 998-1008.
17. Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: applications, advances, and challenges. Artificial intelli- gence review, 57(11), 299.
18. Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., ... & Ting, D. S. W. (2024). Development and testing of retrieval augmented generation in large language models–a case study report. arXiv preprint arXiv:2402.01733.
19. Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diag- nostic medicine with a focus on digital pathology–a recent scoping review. Diagnostic pathology, 19(1), 43.
20. Lu, Y., Zhao, X., & Wang, J. (2024, August). ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In Pro- ceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024) (pp. 64-68).
21. Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.