



Retrieval-Augmented Generative AI for Reliable and Explainable Healthcare Decision Support: A Comprehensive Survey

Sajid Anwar¹, Dr. Pawan Meena², Yogesh Rai³

¹M.Tech Scholar Technocrats Institute of Technology Excellence

^{2,3}Asst. Professor, CSE-AIML, Technocrats Institute of Technology Excellence

Abstract - The rapid advancement of generative artificial intelligence (AI), particularly large language models (LLMs), has created significant opportunities for improving healthcare decision support. These models can interpret natural language, summarize clinical information, answer medical questions, and assist healthcare professionals in accessing complex knowledge. However, the direct application of generative models in healthcare remains challenging because their responses may contain hallucinated facts, outdated knowledge, insufficient evidence, and limited transparency. Such limitations can reduce trust and introduce substantial risks when AI-generated information is used in clinical environments. Retrieval-Augmented Generation (RAG) has emerged as a promising approach for addressing these challenges by combining the generative capabilities of LLMs with external, domain-specific knowledge retrieval. Instead of relying exclusively on parameters learned during training, a RAG system retrieves relevant evidence from trusted sources and incorporates this information into the generation process. This architecture can improve factual grounding, knowledge freshness, source traceability, and explainability.

This survey presents a comprehensive examination of RAG-enabled generative AI for reliable and explainable healthcare decision support. The paper reviews the fundamental principles of generative AI, healthcare decision support systems, and retrieval-augmented architectures. It further analyzes the major components of healthcare RAG pipelines, including medical knowledge acquisition, document preprocessing, indexing, retrieval, reranking, context integration, evidence-grounded generation, and response verification. Particular emphasis is placed on reliability issues such as hallucination mitigation, factual consistency, retrieval quality, uncertainty estimation, robustness, and knowledge updating. The survey also investigates explainability mechanisms, including source attribution, retrieval traceability, evidence-based reasoning, and clinician-oriented explanations.

In addition, the paper examines major healthcare applications of RAG, including clinical question answering, diagnostic assistance, treatment recommendation, electronic health record analysis, drug information retrieval, and personalized decision support. Existing evaluation strategies, privacy and security concerns, ethical challenges, and regulatory considerations are discussed. Finally, key research gaps are identified, and future directions are proposed for developing trustworthy, clinically useful, and human-centered RAG systems capable of supporting healthcare professionals without replacing clinical judgment.

Index Terms— Retrieval-Augmented Generation, Generative Artificial Intelligence, Large Language Models, Healthcare AI, Clinical Decision Support Systems, Explainable AI, Reliable AI, Medical Question Answering, Trustworthy AI.

I. INTRODUCTION

The healthcare sector is rapidly adopting digital technologies, resulting in the generation of large volumes of electronic health records, medical literature, clinical guidelines, diagnostic reports, and biomedical data. The increasing complexity of these heterogeneous sources creates challenges for healthcare professionals in obtaining timely and evidence-based information. Recent advances in generative artificial intelligence (AI) and large language models (LLMs) have demonstrated potential for medical question answering, clinical documentation, information synthesis, and decision support. However, conventional LLMs may generate hallucinated, outdated, or unsupported information, which can create significant risks in clinical environments.[1]

Another major limitation is the lack of transparency in generative AI systems. Healthcare professionals require not only accurate responses but also reliable evidence explaining the basis of AI-generated recommendations. [2] This requirement is particularly important for high-stakes applications involving diagnosis, treatment selection, and patient safety. Therefore, healthcare AI

systems must provide mechanisms for improving factual reliability, evidence verification, and interpretability.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address these limitations by integrating external knowledge retrieval with generative models. Instead of relying exclusively on information stored within model parameters, RAG retrieves relevant information from sources such as medical guidelines, biomedical literature, drug databases, institutional protocols, and authorized clinical records. The retrieved evidence is incorporated into the model context, enabling more knowledge-grounded and traceable responses.[3] RAG also facilitates knowledge updating without requiring complete retraining of the underlying language model.

However, RAG does not inherently guarantee reliable healthcare decision support. The quality of generated responses depends on knowledge-source credibility, retrieval accuracy, document processing, context selection, model faithfulness, and verification mechanisms. Irrelevant, outdated, incomplete, or contradictory evidence may still lead to incorrect outputs. Similarly, source citation alone does not ensure meaningful explainability; the relationship between retrieved evidence and generated recommendations must remain clear.[5]

Although research on healthcare RAG is expanding, existing studies remain fragmented across applications, architectures, datasets, evaluation methods, and reliability mechanisms. There is therefore a need for a comprehensive analysis that jointly considers reliability and explainability in healthcare-oriented RAG systems.[7]

This survey examines the architecture, applications, evaluation methods, reliability challenges, and explainability mechanisms of RAG-based healthcare decision support. Particular emphasis is placed on hallucination mitigation, factual consistency, evidence relevance, knowledge freshness, uncertainty estimation, source attribution, and evidence traceability. The study also identifies current research gaps and future directions for developing reliable, transparent, privacy-aware, and clinically useful RAG systems.

The remainder of this paper is organized as follows: Section II discusses the technical foundations of generative AI, healthcare decision support, and RAG. Section III presents the architecture of healthcare RAG systems. Sections IV and V discuss reliability and explainability, respectively. Section VI reviews healthcare applications, while Section VII discusses evaluation and benchmarking. Section VIII addresses privacy, security, ethical, and regulatory challenges. Section IX provides a comparative analysis, followed by research gaps and future directions in Section X and the conclusion in Section XI.

The major contributions of this survey are summarized as follows:

1. **Comprehensive conceptual overview:** The paper presents a unified overview of generative AI, healthcare decision support, and Retrieval-Augmented Generation, establishing the technical foundation required for understanding healthcare-specific RAG systems.

2. **Healthcare RAG architecture analysis:** A detailed examination of the end-to-end RAG pipeline is provided, including knowledge source selection, document preprocessing, indexing, retrieval, reranking, context integration, response generation, and verification.[4]
3. **Reliability-focused assessment:** The survey analyzes major factors affecting the reliability of RAG-generated healthcare information, including hallucination, retrieval failure, factual inconsistency, outdated knowledge, uncertainty, and robustness.
4. **Explainability and transparency framework:** The study investigates how evidence retrieval, source attribution, and traceable generation can contribute to more interpretable and clinically acceptable AI-assisted decision support.
5. **Application-oriented review:** Major healthcare applications of RAG are examined to demonstrate its potential role across clinical question answering, diagnosis support, treatment planning, medical information retrieval, electronic health records, and personalized decision support.
6. **Identification of research gaps and future directions:** The paper highlights unresolved challenges related to multimodal information integration, personalized retrieval, clinical validation, uncertainty-aware generation, privacy-preserving architectures, standardized evaluation, and human-AI collaboration.

The central argument of this survey is that Retrieval-Augmented Generation should not be viewed merely as a technique for improving the information available to a language model. In healthcare, RAG has the potential to serve as a foundational architecture for developing evidence-aware and transparent AI systems. Its effectiveness, however, depends on the quality of the retrieved knowledge, the ability of the generative model to remain faithful to that knowledge, and the capacity of the overall system to communicate evidence, uncertainty, and limitations to human users. Consequently, future healthcare RAG systems must move beyond optimizing language fluency and retrieval accuracy alone and instead focus on clinically meaningful measures of trustworthiness, safety, explainability, and human-centered usability.[6]

The remainder of this paper is organized as follows. Section II presents the background and technical foundations of generative AI, healthcare decision support systems, and Retrieval-Augmented Generation. Section III examines the architecture and workflow of RAG-based healthcare decision support systems. Section IV discusses reliability challenges and methods for improving factual accuracy, robustness, and evidence grounding. Section V focuses on explainability and transparency mechanisms. Section VI reviews major healthcare applications of RAG. Section VII presents evaluation metrics and benchmarking strategies. Section VIII discusses privacy, security, ethical, and regulatory considerations. Section IX provides a comparative analysis of existing approaches. Section

X identifies key research gaps and future research directions. Finally, Section XI concludes the survey and summarizes the potential of reliable and explainable RAG systems for next-generation healthcare decision support. [8]

II. METHODOLOGY OF THE SURVEY

A systematic literature-oriented approach was adopted for this survey. Recent research was examined across major scholarly sources, including literature indexed in PubMed, Web of Science, IEEE Xplore, and related academic databases. A recent 2025 scoping review of medical and nursing RAG searched PubMed, Web of Science, IEEE Xplore, and arXiv for studies published between November 2022 and May 2025. From 917 retrieved articles, 67 met its inclusion criteria.

Study	Year	Focus	Major Finding
Amugongo et al.	2025	Healthcare RAG systematic review	Diverse RAG approaches; limited standardized evaluation
Liu et al.	2025	Biomedical RAG	RAG significantly improved performance over baseline LLMs in included studies
Miao et al.	2025	Medical/nursing RAG	67 studies met inclusion criteria from 917 retrieved records
Bedi et al.	2025	Healthcare LLM evaluation	Current evaluation approaches require broader dimensions
WHO	2024/2025	AI governance	Emphasizes safety, ethics, accountability and responsible use

III. BACKGROUND AND TECHNICAL FOUNDATIONS

A. Generative Artificial Intelligence and Large Language Models

Generative artificial intelligence refers to computational systems capable of producing new content, including text, images, audio, code, and structured information, based on patterns learned from training data. In recent years, the development of transformer-based architectures has significantly advanced the ability of AI systems to understand and generate natural language. Large language models represent one of the most influential developments within generative AI because they can process

extensive textual contexts and generate coherent responses across a wide range of tasks. [9]

An LLM estimates a probability distribution over sequences of tokens. Conceptually:

where:

x = input prompt,

y = generated response.

The model relies on knowledge learned during training. This creates three major limitations for healthcare:

1. Knowledge cutoff.
2. Potential hallucination.

Lack of direct evidence attribution.

In healthcare, LLMs have demonstrated potential in several activities, including medical question answering, clinical documentation, report summarization, patient communication, biomedical literature analysis, and clinical decision support. Their ability to process unstructured language is particularly valuable because a substantial portion of healthcare information exists in textual form. Clinical notes, discharge summaries, radiology reports, medical publications, treatment guidelines, and patient communications contain important information that may be difficult to process using conventional structured-data approaches. [10]

However, language generation should not be directly interpreted as evidence of factual understanding or clinical reasoning. An LLM may generate a highly fluent and convincing response even when the underlying information is incomplete or incorrect. This limitation is particularly important in healthcare, where linguistic plausibility and clinical correctness are fundamentally different properties. A response may appear medically reasonable while containing inaccurate statements, unsupported conclusions, or inappropriate recommendations.

B. Healthcare Decision Support Systems

A healthcare decision support system is designed to assist healthcare professionals, patients, or administrators in making informed decisions by analyzing relevant information and presenting useful recommendations, alerts, predictions, or explanations. These systems do not necessarily replace human decision-makers; instead, their primary objective is to provide computational assistance that improves the availability, organization, and interpretation of relevant knowledge. [11]

Traditional clinical decision support systems often rely on predefined rules, knowledge bases, statistical models, or machine learning algorithms. Rule-based systems can provide transparent recommendations but may require substantial manual effort to create and maintain. Machine learning systems can identify complex patterns within healthcare data, but their predictions may be difficult to interpret and may depend heavily on the characteristics of the training dataset. [12]

The emergence of generative AI introduces new capabilities for healthcare decision support. Instead of producing only a numerical prediction or predefined alert, a generative system can communicate information using natural language. For example, such a system may summarize relevant clinical evidence, answer questions about treatment alternatives, explain possible

interactions among medications, or synthesize information from multiple medical documents.

RAG can contribute to these requirements by connecting generative models with curated knowledge sources. Instead of relying exclusively on internal model knowledge, the system can retrieve information relevant to a particular question or clinical context. This enables the decision support process to become more evidence-aware and potentially more adaptable to changes in healthcare knowledge.

C. Fundamentals of Retrieval-Augmented Generation

Retrieval-Augmented Generation is an AI architecture that integrates information retrieval with generative modeling. The central principle is to retrieve relevant external knowledge before or during the generation of a response. The retrieved information is then incorporated into the context provided to the language model. [14]

A general RAG system can be represented conceptually as:

User Query → Query Processing → Knowledge Retrieval → Context Selection → Generative Model → Evidence-Grounded Response

Let a user query be represented as (q), and let a knowledge repository contain a collection of documents:

$$D = \{d_1, d_2, d_3, \dots, d_n\}$$

The retrieval component identifies a subset of relevant documents:

$$R(q) = \{d_i \in D \mid S(q, d_i) > \tau\}$$

Where ($R(q)$) represents the retrieved document set and (τ) represents a relevance threshold.

The selected documents are combined with the original query to construct an augmented context:

$$C = f(q, R(q))$$

Where ($f(.)$) represents the context construction process.

The generative model then produces an output:

$$y = G(q, f(q, R(q)))$$

Where (G) represents the generative language model and (y) represents the generated response.

The purpose of this process is to provide the model with external evidence that can support its response. In healthcare, the knowledge repository may include clinical practice guidelines, medical research articles, institutional protocols, drug information, biomedical databases, and other validated sources. [13]

The quality of the final output depends on both retrieval and generation. A highly capable language model cannot reliably compensate for completely irrelevant evidence, while an effective retrieval system may still fail if the generative model misinterprets the retrieved context. Therefore, RAG should be understood as an integrated pipeline rather than a simple combination of two independent components.

D. Knowledge Retrieval and Representation

The retrieval component is responsible for identifying information relevant to a user query or clinical problem. Traditional information retrieval techniques often use lexical matching, where documents are ranked according to the occurrence and importance of query terms. Such approaches can be effective when the terminology in the query closely matches the terminology used in the knowledge source. [15]

However, healthcare language frequently contains abbreviations, synonyms, specialized terminology, and variations in clinical expression. A clinician may use different terms to describe the same condition, while a patient may express symptoms using nontechnical language. Semantic retrieval addresses this challenge by representing queries and documents as numerical vectors, commonly referred to as embeddings. Let the embedding function be represented as ($E(\cdot)$).

Documents with higher similarity scores are considered more relevant to the query. These representations can be stored within vector databases to support efficient similarity-based retrieval.

For healthcare applications, purely semantic similarity may not always be sufficient. Clinical relevance can depend on factors such as patient characteristics, disease severity, treatment history, publication quality, guideline authority, and evidence recency. Consequently, advanced healthcare RAG systems may combine lexical retrieval, semantic retrieval, metadata filtering, knowledge graphs, and reranking models. [16]

A hybrid retrieval approach can be represented as:

$$S_{\text{hybrid}}(d, q) = \alpha S_{\text{lexical}}(d, q) + (1 - \alpha) S_{\text{semantic}}(d, q)$$

Where:

- $S_{\text{hybrid}}(d, q)$ = Combined relevance score of document d for query q
- $S_{\text{lexical}}(d, q)$ = Lexical similarity score
- $S_{\text{semantic}}(d, q)$ = Semantic similarity score
- α = Weight assigned to lexical retrieval, $0 \leq \alpha \leq 1$
- $1 - \alpha$ = Weight assigned to semantic retrieval

This approach can improve retrieval robustness by combining exact terminology matching with conceptual similarity.

E. Document Processing and Knowledge Indexing

Before information can be retrieved effectively, healthcare documents must undergo preprocessing and indexing. Medical knowledge sources may vary substantially in structure and format. They may include scientific articles, clinical guidelines, electronic health records, laboratory reports, institutional documents, or structured biomedical databases. A common preprocessing workflow includes document collection, cleaning,

segmentation, metadata extraction, embedding generation, and indexing. Large documents are typically divided into smaller segments or chunks because language models and retrieval systems have limitations regarding the amount of information that can be efficiently processed at one time.

The selection of chunk size represents an important design consideration. If segments are too small, clinically meaningful context may be lost. Conversely, excessively large segments may contain irrelevant information and reduce retrieval precision. Therefore, healthcare RAG systems may use context-aware chunking strategies that preserve logical sections, medical concepts, or relationships among clinical findings. Metadata can further improve retrieval quality. Relevant metadata may include publication date, document type, medical specialty, evidence level, source organization, patient population, and clinical category. Such information enables filtering and prioritization before generation. [16]

F. Retrieval-Augmented Generation in Healthcare

The application of RAG in healthcare differs from general-purpose information retrieval because the quality and authority of information sources are critically important. An incorrect restaurant recommendation or product description may have limited consequences, whereas unsupported medical information may influence healthcare decisions. Therefore, healthcare RAG systems should emphasize knowledge governance and source validation.

A healthcare-oriented RAG architecture can be conceptualized as consisting of five major layers:

1. **Knowledge Layer:** Contains validated medical knowledge, clinical guidelines, biomedical literature, drug information, institutional protocols, and authorized patient information.
2. **Retrieval Layer:** Identifies and ranks information relevant to the clinical query using lexical, semantic, hybrid, or graph-based retrieval techniques.
3. **Reasoning and Generation Layer:** Uses an LLM or other generative model to synthesize retrieved information into a coherent response.
4. **Verification Layer:** Evaluates whether the generated response is consistent with the retrieved evidence and identifies unsupported or uncertain statements.
5. **Interaction and Explanation Layer:** Presents the response together with evidence sources, confidence information, limitations, and explanations appropriate for the intended user.

This layered architecture highlights an important principle: trustworthy healthcare RAG cannot depend solely on the language model. Reliability must be considered throughout the entire information lifecycle, from knowledge acquisition to retrieval, generation, verification, and human interaction.

G. Relationship Between RAG, Reliability, and Explainability

The relevance of RAG to healthcare can be understood through its potential contributions to three interconnected objectives: knowledge grounding, reliability, and explainability.

Knowledge grounding refers to connecting generated information with external evidence. Rather than producing an answer entirely from internal model representations, the system uses retrieved documents as contextual support.

Reliability refers to the degree to which a system consistently provides accurate, relevant, safe, and evidence-supported information. In a RAG architecture, reliability is influenced by the quality of the knowledge base, retrieval accuracy, evidence coverage, model faithfulness, and verification mechanisms.

Explainability refers to the ability of users to understand the basis of a system's output. RAG can improve transparency by identifying the documents or evidence used during response generation. However, meaningful explainability requires more than displaying references. The system should ideally communicate the relationship between the evidence and the generated recommendation. [17]

H. Summary of Technical Foundations

Generative AI and LLMs provide powerful capabilities for processing and synthesizing healthcare information, but their direct use in clinical environments is constrained by hallucination, outdated knowledge, limited transparency, and uncertainty. Healthcare decision support systems require stronger guarantees of factual relevance and evidence traceability than many general-purpose AI applications. Retrieval-Augmented Generation provides a technical foundation for addressing these limitations by integrating external knowledge retrieval with generative models. A typical RAG system involves query processing, document retrieval, context construction, language generation, and potentially output verification. In healthcare, additional requirements arise regarding knowledge authority, evidence quality, privacy, clinical relevance, and explainability.

The technical foundations discussed in this section establish the basis for examining how RAG can be systematically designed for healthcare decision support. The next section presents the architecture and operational workflow of healthcare-specific RAG systems, with particular emphasis on knowledge acquisition, preprocessing, retrieval, reranking, context integration, evidence-grounded generation, and verification.

IV. RAG ARCHITECTURE FOR HEALTHCARE DECISION SUPPORT

A Retrieval-Augmented Generation (RAG) system for healthcare integrates external medical knowledge with a generative language model to produce evidence-grounded decision support. Unlike conventional LLM-based systems, the RAG architecture retrieves relevant information dynamically before generating a response. A typical healthcare RAG framework consists of five major stages: knowledge acquisition, document processing, information retrieval, response generation, and output verification.

A. Knowledge Acquisition and Preparation: The knowledge base may include clinical guidelines, peer-reviewed medical literature, drug databases, hospital protocols, and electronic health records. These sources are collected, cleaned, segmented, and transformed into machine-readable representations. Metadata such as publication date, medical specialty, and evidence level can also be stored to improve retrieval quality.

B. Retrieval and Reranking: When a clinical query is received, the system converts the query into an appropriate representation and retrieves the most relevant documents from the knowledge repository. Dense retrieval, keyword-based retrieval, or hybrid approaches may be employed. A reranking mechanism can further prioritize documents according to semantic relevance, source credibility, and clinical applicability. [19]

The retrieved evidence can be expressed as:

$$R(q) = \text{TopK}_{d_i \in D} (S(q, d_i))$$

Where:

- $R(q)$ = Set of top-K retrieved documents for query q
- q = Input clinical query
- $D = \{d_1, d_2, \dots, d_n\}$ = Medical knowledge repository
- d_i = i -th document in the repository
- $S(q, d_i)$ = Similarity score between query q and document d_i
- K = Number of documents retrieved

C. Context-Augmented Generation: The selected documents are combined with the user query and provided as contextual input to the generative model. The model then synthesizes the available evidence into a coherent response. In healthcare applications, generation should be constrained to the retrieved information whenever possible to reduce unsupported claims and hallucinated content.

D. Evidence Verification: An additional verification layer can assess whether the generated response is supported by the retrieved documents. This stage may identify factual inconsistencies, unsupported statements, conflicting evidence, or insufficient information. If reliable evidence is unavailable, the system should communicate uncertainty rather than generate a confident but potentially incorrect response. [18]

E. Explainable Response Presentation: The final response should provide appropriate source attribution and evidence traceability. Instead of presenting only a generated recommendation, the system can indicate the supporting medical sources and relevant evidence. Such transparency enables clinicians to review the basis of the generated output and supports informed human decision-making.

V. RELIABILITY IN HEALTHCARE RAG SYSTEMS

Reliability is a fundamental requirement for deploying Retrieval-Augmented Generation (RAG) systems in healthcare. Since AI-generated information may influence clinical decisions, the system must provide accurate, relevant, evidence-supported, and consistent responses. RAG can improve reliability by grounding

generated responses in external medical knowledge; however, retrieval and generation errors can still affect the final output.

A. Hallucination Mitigation: Hallucination occurs when a generative model produces information that is unsupported or factually incorrect. RAG can reduce this problem by providing relevant evidence within the generation context. However, retrieved evidence does not guarantee factual accuracy. Therefore, healthcare RAG systems should incorporate response validation and evidence-consistency mechanisms.

B. Retrieval Accuracy: The quality of a generated response depends strongly on the relevance of retrieved information. Irrelevant or incomplete documents can result in incorrect conclusions. Retrieval performance can be evaluated using measures such as Precision@K, Recall@K, and Mean Reciprocal Rank (MRR). Hybrid retrieval and reranking techniques can improve the selection of clinically relevant evidence.

C. Factual Consistency: A reliable system should ensure that generated statements remain consistent with the retrieved evidence. Let (G) represent the generated response and (E) represent the retrieved evidence. A factual consistency function can be expressed conceptually as:

$$C(G, E) = \frac{N_{\text{supported}}}{N_{\text{total}}}$$

Where:

- $C(G, E)$ = **Factual Consistency Score**
- G = Generated response
- E = Retrieved evidence
- $N_{\text{supported}}$ = Number of claims in G supported by E
- N_{total} = Total number of factual claims in G

A higher value indicates that a greater proportion of generated claims are supported by available evidence.

D. Knowledge Freshness: Medical knowledge continuously changes due to new research, clinical trials, treatment guidelines, and regulatory updates. RAG enables knowledge repositories to be updated independently of the underlying language model. Consequently, systems can incorporate newer evidence without completely retraining the generative model.

E. Uncertainty and Human Oversight: Healthcare RAG systems should recognize situations in which sufficient evidence is unavailable or contradictory. Instead of producing highly confident responses under uncertainty, the system should communicate limitations and recommend appropriate human verification. Final clinical decisions should remain under qualified healthcare professionals, with AI serving as a decision-support mechanism rather than an autonomous replacement.

VI. PROPOSED RELIABLE HEALTHCARE RAG FRAMEWORK

The proposed Clinical RAG (Retrieval-Augmented Generation) architecture is designed to provide reliable and evidence-based medical responses. First, the user or clinician submits a medical query. The system performs query processing and rewriting to

clarify and improve the query. It then retrieves relevant information through two approaches: keyword retrieval, which finds exact terms, and vector retrieval, which identifies semantically similar information.

The retrieved documents are passed to a reranker, which prioritizes the most relevant sources. An evidence filter removes unreliable or irrelevant information and ensures that trusted medical knowledge and clinical guidelines are considered. The context builder combines the selected evidence into a structured context for the LLM/Generative AI.

The LLM generates an answer, while a citation check verifies that claims are properly supported by the retrieved evidence. An uncertainty check identifies potentially unreliable or ambiguous information. Finally, clinical review provides human oversight before producing the final response. **The proposed architecture consists of eight layers.**

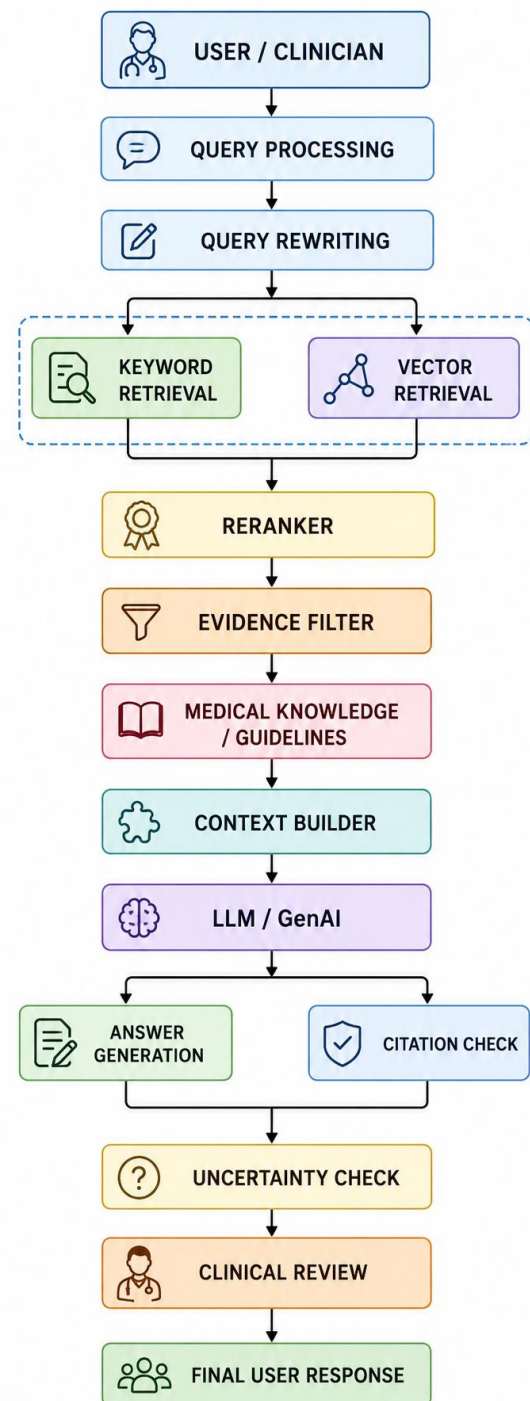


Fig. Healthcare RAG Framework

VII. APPLICATIONS OF RAG IN HEALTHCARE

Retrieval-Augmented Generation (RAG) can support a wide range of healthcare applications by combining generative AI with external and domain-specific medical knowledge. Its ability to retrieve relevant evidence before generating responses makes it suitable for information-intensive clinical tasks.

A. Clinical Question Answering: RAG-based systems can answer clinical questions by retrieving relevant information from medical guidelines, research publications, and institutional knowledge bases. Evidence-grounded responses can assist

healthcare professionals in quickly accessing relevant medical information.

B. Diagnostic Decision Support: RAG can support diagnostic processes by retrieving information related to symptoms, diseases, diagnostic criteria, and clinical findings. The retrieved evidence can help clinicians consider relevant possibilities while maintaining human control over the final diagnosis.

C. Treatment and Medication Support: Healthcare RAG systems can retrieve treatment guidelines, drug information, contraindications, and medication-related evidence. Such systems may assist clinicians in evaluating treatment options and identifying potentially relevant medication information. [20]

D. Electronic Health Record Analysis: Electronic Health Records (EHRs) contain large amounts of structured and unstructured information. RAG can retrieve relevant patient-specific records and combine them with external medical knowledge to generate concise summaries and context-aware responses.

E. Biomedical Literature Retrieval: RAG can improve access to the rapidly expanding biomedical literature. Researchers and healthcare professionals can use natural-language queries to retrieve relevant publications and generate evidence-based summaries, reducing the effort required to manually examine large collections of scientific documents.

F. Personalized Healthcare Support: By combining patient-specific information with appropriate medical knowledge, RAG systems can support personalized information delivery. However, patient data must be handled using appropriate privacy, security, and access-control mechanisms.

VIII. EVALUATION METRICS AND BENCHMARKING

Evaluation is essential for determining whether Retrieval-Augmented Generation (RAG) systems can provide reliable and clinically useful healthcare decision support. Conventional language-generation metrics alone are insufficient because healthcare applications require assessment of retrieval quality, factual correctness, evidence grounding, and clinical relevance.

A. Retrieval Evaluation: Retrieval performance determines whether appropriate evidence is supplied to the generative model. Common metrics include Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG). These measures evaluate the relevance and ranking of retrieved medical documents.

B. Generation Evaluation: Generated responses can be evaluated using metrics such as factual accuracy, relevance, completeness, and semantic similarity. Automated measures may provide initial assessments, but they should be complemented by expert evaluation because medically appropriate responses may require contextual interpretation.

C. Faithfulness and Groundedness: Faithfulness measures whether the generated response is supported by the retrieved evidence. A response containing medically plausible information but unsupported claims should not be considered fully reliable. Claim-level verification can therefore be used to determine the

proportion of statements adequately supported by retrieved sources.

D. Clinical Evaluation: Clinical evaluation should involve qualified healthcare professionals who assess the correctness, usefulness, safety, and applicability of generated responses. Expert evaluation is particularly important for diagnostic and treatment-related applications where errors may have significant consequences.

E. Explainability Evaluation: Explainability can be assessed by examining source attribution, evidence relevance, transparency, and user understanding. Clinicians should be able to identify the evidence supporting an AI-generated recommendation and determine whether the explanation is sufficient for informed decision-making.

F. Benchmarking Requirements: A comprehensive healthcare RAG benchmark should evaluate the complete pipeline rather than only the language model. Future benchmarks should include diverse medical specialties, realistic clinical scenarios, recent evidence, uncertainty cases, conflicting sources, and patient-specific contexts.

IX. PRIVACY, SECURITY, ETHICAL, AND REGULATORY CHALLENGES

The deployment of Retrieval-Augmented Generation (RAG) in healthcare introduces significant concerns related to patient privacy, data security, ethical responsibility, and regulatory compliance. Since healthcare RAG systems may process sensitive clinical information, these aspects must be addressed throughout system design and deployment.

A. Patient Privacy: Healthcare RAG systems may retrieve information from Electronic Health Records (EHRs), clinical reports, and other patient-specific sources. Unauthorized access or inappropriate retrieval may expose sensitive information. Data minimization, anonymization, access control, and secure data storage are therefore essential.

B. Data and System Security: RAG architectures introduce additional security risks because external knowledge repositories directly influence generated responses. Malicious or incorrect documents may potentially affect retrieval results. Secure indexing, source validation, access management, monitoring, and protection against prompt and data manipulation are necessary to maintain system integrity.

C. Bias and Fairness: Medical datasets may contain demographic, geographic, or socioeconomic biases. If such biases are present in the knowledge base or retrieved evidence, the generated response may reproduce or amplify them. Healthcare RAG systems should therefore be evaluated across diverse patient populations to identify and reduce unfair outcomes.

D. Accountability and Human Oversight: Determining responsibility for AI-assisted decisions is an important ethical issue. RAG systems should function as decision-support tools rather than autonomous clinical authorities. Healthcare professionals should retain responsibility for evaluating AI-generated information and making final clinical decisions.

E. Regulatory Compliance: Healthcare AI systems must comply with applicable privacy, medical-device, data-protection, and AI regulations. Regulatory requirements may differ according to country, application, and level of clinical risk. Therefore, RAG systems intended for real-world healthcare deployment require appropriate documentation, validation, monitoring, and governance.

F. Ethical Transparency: Users should be informed when healthcare information is generated or assisted by AI. The system should clearly communicate its limitations, evidence sources, and uncertainty. Such transparency can reduce inappropriate reliance on automated recommendations.

Reliable healthcare RAG requires privacy-preserving data management, robust security, fairness assessment, clear accountability, regulatory compliance, and meaningful human oversight. Addressing these challenges is essential for responsible clinical adoption.

X. COMPARATIVE ANALYSIS OF EXISTING RAG APPROACHES

Healthcare RAG systems differ in their retrieval strategies, knowledge sources, generation mechanisms, and verification capabilities. A comparative analysis is useful for identifying the strengths and limitations of different architectural approaches.

Approach	Retrieval Strategy	Major Advantage	Key Limitation
Sparse RAG	Keyword-based retrieval	Effective for exact medical terminology	Limited semantic understanding
Dense RAG	Embedding-based retrieval	Captures semantic relationships	May retrieve clinically irrelevant content
Hybrid RAG	Sparse + dense retrieval	Combines lexical and semantic relevance	Higher computational complexity
Reranked RAG	Retrieval followed by reranking	Improves evidence selection	Additional processing overhead
Knowledge-Graph RAG	Graph-based retrieval	Represents relationships among medical concepts	Knowledge graph construction is complex
Multimodal RAG	Text, image, and other modalities	Supports diverse clinical information	Requires multimodal models and datasets

Approach	Retrieval Strategy	Major Advantage	Key Limitation
Verification-Aware RAG	Retrieval + response verification	Improves factual reliability	Increased latency and system complexity

Sparse retrieval is particularly useful when medical terminology must be matched precisely, whereas dense retrieval is better suited to semantically related queries. Hybrid approaches combine these advantages and can provide more robust retrieval. Reranking further improves retrieval quality by evaluating candidate documents using additional relevance criteria.

Knowledge-graph-based RAG provides structured relationships among diseases, symptoms, medications, and treatments, making it useful for complex medical queries. Multimodal RAG extends retrieval beyond textual information by incorporating medical images, charts, and other clinical data. However, these approaches generally require more sophisticated infrastructure. Verification-aware architectures introduce an additional stage for checking whether generated responses are supported by retrieved evidence. Although this can improve reliability, it increases computational requirements and response latency.

XI. RESEARCH GAPS AND FUTURE DIRECTIONS

Although Retrieval-Augmented Generation (RAG) has significant potential for healthcare decision support, several research challenges remain unresolved. Existing systems primarily focus on improving retrieval and response quality, while clinical reliability, explainability, safety, and real-world validation require further investigation.

A. Clinically Aware Retrieval: Future RAG systems should incorporate clinical context such as patient history, symptoms, age, medications, and disease severity during retrieval. Context-aware retrieval can improve the relevance of evidence for individual clinical scenarios.

B. Multimodal Healthcare RAG: Most existing RAG systems focus primarily on textual information. Future research should integrate medical images, laboratory reports, physiological signals, and structured EHR data. Multimodal RAG could provide more comprehensive decision support across complex clinical workflows.

C. Uncertainty-Aware Generation: Healthcare systems should explicitly identify uncertainty, incomplete evidence, and conflicting medical information. Future models should develop reliable confidence estimation and abstention mechanisms rather than generating unsupported conclusions.

D. Evidence Verification: More robust claim-level verification methods are required to determine whether generated statements are adequately supported by retrieved evidence. Automated verification combined with expert review can improve the reliability of high-risk clinical applications.

E. Privacy-Preserving RAG: Future architectures should investigate federated learning, secure retrieval, anonymization, encryption, and privacy-preserving computation to enable RAG systems to utilize sensitive healthcare information without unnecessary exposure of patient data.

F. Standardized Clinical Evaluation: There is a need for standardized benchmarks that evaluate retrieval quality, factuality, clinical usefulness, explainability, fairness, safety, and robustness simultaneously. Evaluation should involve healthcare professionals and realistic clinical scenarios rather than relying exclusively on automated language metrics.

G. Human-AI Collaboration: Future healthcare RAG systems should be designed around human-centered interaction. Instead of replacing clinicians, these systems should provide evidence, alternatives, uncertainty information, and explanations that support professional judgment.

The future of healthcare RAG depends on moving from general-purpose information retrieval toward clinically aware, evidence-grounded, privacy-preserving, multimodal, and uncertainty-aware architectures. Progress in these areas can contribute to safer and more trustworthy AI-assisted healthcare decision support.

XII. CONCLUSION

Retrieval-Augmented Generation (RAG) provides a promising framework for developing reliable and explainable generative AI systems for healthcare decision support. By connecting large language models with external medical knowledge sources, RAG can reduce dependence on static model knowledge and improve access to current, domain-specific evidence. This capability is particularly valuable in healthcare, where accuracy, transparency, and evidence-based decision-making are essential.

This survey examined the fundamental architecture of healthcare-oriented RAG systems, including knowledge acquisition, preprocessing, retrieval, reranking, context integration, generation, and verification. The analysis demonstrated that system reliability depends not only on the underlying language model but also on the quality and relevance of retrieved evidence. Hallucination mitigation, factual consistency, knowledge freshness, uncertainty communication, and human oversight are therefore critical requirements.

The survey also highlighted the importance of explainability. Source attribution, evidence traceability, and transparent communication of uncertainty can enable healthcare professionals to validate AI-generated information and maintain appropriate control over clinical decisions. Furthermore, RAG can support diverse applications, including clinical question answering, diagnostic assistance, treatment support, EHR analysis, biomedical literature retrieval, and personalized healthcare.

Despite these advantages, challenges remain in privacy protection, security, bias mitigation, multimodal integration, clinical validation, standardized evaluation, and regulatory compliance. Future research should therefore focus on clinically aware retrieval, multimodal RAG, uncertainty-aware generation,

automated evidence verification, privacy-preserving architectures, and human-centered AI collaboration. RAG should be viewed not merely as an enhancement to generative language models but as an architectural foundation for evidence-aware healthcare AI. With rigorous validation, responsible governance, and continuous human oversight, RAG-based systems can contribute to safer, more transparent, and clinically useful healthcare decision support while preserving the central role of healthcare professionals in final decision-making.

XIII. REFERENCES

- [1] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, no. 6, Art. no. e0000877, Jun. 2025, doi: 10.1371/journal.pdig.0000877.
- [2] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, and N. Liu, "Retrieval-augmented generation for generative artificial intelligence in health care," *npj Health Systems*, vol. 2, Art. no. 2, Jan. 2025, doi: 10.1038/s44401-024-00004-1.
- [3] S. Liu, A. B. McCoy, and A. Wright, "Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines," *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605–615, 2025, doi: 10.1093/jamia/ocaf008.
- [4] Y. Miao et al., "Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review," *Journal of Medical Internet Research*, 2025.
- [5] "Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations," *Medical Artificial Intelligence*, 2026.
- [6] "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *DIGITAL HEALTH*, 2025, doi: 10.1177/20552076251337177.
- [7] "Integrating fine-tuning and retrieval-augmented generation for healthcare AI systems: A scoping review," 2026.
- [8] L. Garza, A. Kotal, M. A. Grasso, and E. Umucu, "Retrieval-Augmented Framework for LLM-Based Clinical Decision Support," *arXiv preprint arXiv:2510.01363*, 2025.
- [9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [10] J. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2023.
- [11] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.

- [12] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in *Proc. 37th Int. Conf. Machine Learning*, 2020.
- [13] K. Singhal et al., "Towards expert-level medical question answering with large language models," *Nature Medicine*, vol. 29, pp. 1870–1880, 2023.
- [14] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.
- [15] M. Moor et al., "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, pp. 259–265, 2023.
- [16] S. Thirunavukarasu et al., "Large language models in medicine," *Nature Medicine*, vol. 29, pp. 1930–1940, 2023.
- [17] A. Holzinger, K. Keiblinger, K. Holub, M. Malle, A. M. C. Bloice, and M. S. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, Art. no. e1312, 2019.
- [18] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [20] T. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.