ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Retrieval-augmented generation and agentic AI in healthcare: Applications, challenges, and future directions

Bihong Zhu, Xin Kang*, Kazuyuki Matsumoto, and Minoru Yoshida

Department of Information Science and Intelligent Systems, Faculty of Science and Engineering, Tokushima University, Information Science South Building, 2-1 Minamijyousanjima-cho, Tokushima, Japan

Abstract

Large language models have shown growing potential in healthcare, but their clinical use remains constrained by hallucinations, outdated knowledge, limited traceability, and weak alignment with evidence-based practice. Retrieval-augmented generation and emerging agentic artificial intelligence (AI) frameworks offer a promising path to address these limitations by combining external medical knowledge, iterative reasoning, and tool-supported decision support. This review examines the progression from foundational medical retrieval-augmented generation (RAG) systems to more adaptive agentic frameworks, with emphasis on retrieval design, knowledge integration, reasoning control, and healthcare-oriented deployment. Key application areas including medical question answering, radiology support, rare disease reasoning, and clinical decision support are reviewed alongside the technical and practical challenges that continue to limit routine use. These challenges include retrieval noise, factual inconsistency, verification difficulties, privacy and governance concerns, and the lack of standardized evaluation. The review also highlights future directions in multimodal integration, human oversight, evidence grounding, and clinically responsible implementation. By organizing current developments around applications, challenges, and future directions, this article provides a healthcare-oriented perspective on the opportunities and limitations of RAG and agentic AI.

Keywords: Retrieval-augmented generation; Agentic AI; Healthcare; Clinical decision support; Medical question answering; Evidence grounding

***Corresponding author:**
Xin Kang
(kang-xin@is.tokushima-u.ac.jp)

Citation: Zhu B, Kang X, Mastsumoto K, Yoshida M. Retrieval-augmented generation and agentic AI in healthcare: Applications, challenges, and future directions. *Artif Intell Health.* 2026;3(4):026170040. doi: 10.36922/AIH026170040

Received: April 24, 2026

Revised: May 25 2026

Accepted: May 29, 2026

Published online: June 26, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The integration of artificial intelligence (AI) into healthcare has progressed from early computer vision applications for medical imaging, exemplified by CheXNet's radiologist-level pneumonia detection on chest X-rays[1], to sophisticated language-based systems capable of clinical reasoning and decision support.[2,3] Deep learning techniques have demonstrated remarkable capabilities across multiple medical domains, including computer vision for radiological interpretation, natural language processing for electronic health record analysis, reinforcement learning for surgical assistance, and genomic sequence analysis.[2] However, the deployment of large language models (LLMs) in clinical settings faces substantial challenges related to factual accuracy, knowledge

currency, and alignment with evidence-based medicine principles.[4–10]

Traditional LLMs encode knowledge implicitly within their parameters, creating inherent limitations including susceptibility to hallucination, inability to access updated medical literature, and lack of transparency in reasoning processes.[11,12] These limitations are particularly problematic in healthcare contexts where accuracy, accountability, and traceability are paramount.[3,13] Retrieval-augmented generation (RAG) has emerged as a promising solution, combining parametric knowledge stored in model weights with non-parametric knowledge accessed through external retrieval mechanisms.[11,14]

The evolution from basic RAG to agentic RAG represents a paradigm shift in how AI systems interact with medical knowledge.[15,16] While traditional RAG employs fixed, single-pass retrieval processes, agentic RAG introduces autonomous agents capable of planning, iterative retrieval, tool integration, and multi-step reasoning.[16,17] This transition enables systems to emulate the hypothetico-deductive reasoning characteristic of clinical experts, actively formulating queries, evaluating evidence utility, and refining search strategies to address diagnostic gaps.[17]

This review examines how retrieval-augmented generation in healthcare has evolved from foundational medical RAG systems to more adaptive agentic frameworks, and what that transition means for practical use in clinical settings.

1.1. Research gap and contribution of this review

Despite the rapid growth of research on healthcare AI, medical large language models, and retrieval-augmented generation (RAG), the existing review literature remains fragmented. Prior reviews often focus on broad healthcare AI adoption, single clinical specialties, or general LLM applications, while reviews specifically discussing RAG are frequently oriented toward general-domain architectures rather than healthcare-specific deployment requirements. As a result, important issues such as evidence grounding, traceability, verification, governance, and clinically responsible deployment are often discussed only partially or in isolation.

Recent healthcare-focused RAG reviews have provided valuable summaries of emerging medical retrieval systems and biomedical language model integration strategies.[14,15] For example, Abo El-Enen *et al.*[14] primarily emphasized medical RAG applications, implementation considerations, and healthcare adoption challenges, whereas Li *et al.*[15] focused more heavily on retrieval strategies, biomedical large language models, and technical optimization approaches for medical RAG systems. However, the existing review literature still provides comparatively limited discussion of how healthcare RAG is evolving toward more adaptive agentic reasoning architectures capable of iterative retrieval, retrieval verification, tool use, graph-enhanced reasoning, and dynamic evidence refinement within clinically grounded workflows.

This review addresses that gap by synthesizing foundational medical RAG systems, agentic extensions, technical modules, application settings, and deployment-oriented challenges within a unified healthcare-oriented framework. In particular, the review emphasizes the architectural transition from conventional retrieval-augmented generation toward increasingly adaptive reasoning environments that combine retrieval grounding with iterative reasoning and evidence-aware orchestration.

Specifically, this review makes three contributions. First, it traces the progression from medical LLMs and conventional retrieval-augmented generation to more adaptive agentic frameworks that support iterative retrieval, tool use, verification, and multi-step reasoning. Second, it compares current developments across key application domains, giving primary emphasis to clinical decision support and rare disease reasoning, while treating medical question answering mainly as benchmark background and radiology as a multimodal extension. Third, it examines the practical constraints that continue to limit clinical deployment, including retrieval quality, factual consistency, verification, reproducibility, privacy, governance, and evaluation.

1.2. Organization of the review

The remainder of this review is organized as follows. Section 2 describes the review methodology, including the search sources, eligibility criteria, and scope of the literature synthesis. Section 3 introduces medical LLMs as the immediate background for retrieval-augmented systems in healthcare. Section 4 reviews the basic principles and medical adaptations of RAG, and section 5 examines the emergence of agentic RAG. Section 6 discusses advanced retrieval paradigms. Section 7 analyzes representative healthcare applications, section 8 discusses major challenges to trustworthy and clinically responsible deployment, section 9 outlines future directions, and section 10 concludes the review.

2. Methodology

2.1. Literature search strategy

A structured literature review methodology was adopted to identify studies related to retrieval-augmented generation (RAG), agentic AI systems, and healthcare-oriented

large language model applications. Literature retrieval was conducted using PubMed, Scopus, Web of Science, IEEE Xplore, and Google Scholar. The search covered publications between January 2016 and February 2026 in order to capture both the emergence of medical LLMs and the more recent development of retrieval-grounded and agentic reasoning architectures.

Search queries combined keywords related to healthcare AI, retrieval augmentation, agentic reasoning, multimodal medical systems, and clinical decision support. Representative search terms included "medical RAG," "retrieval-augmented generation in healthcare," "agentic AI," "clinical large language models," "multimodal medical foundation models," and "medical reasoning systems." Reference lists of relevant studies and recent review articles were additionally screened to identify potentially eligible publications.

2.2. Eligibility criteria and study classification

Studies were considered eligible for the primary synthesis if they satisfied at least one of the following criteria: (1) implementation of a healthcare-oriented RAG system, (2) development of an agentic retrieval or reasoning framework for medical tasks, (3) evaluation of retrieval-grounded clinical decision support systems, or (4) empirical assessment of multimodal medical reasoning architectures.

To improve methodological consistency, included literature was further divided into two categories: (i) primary empirical studies; and (ii) background and contextual literature.

Primary empirical studies were required to include implemented system architectures together with quantitative evaluation, benchmark comparison, clinical validation, physician assessment, or deployment-oriented experimentation. By contrast, survey articles, conceptual overviews, taxonomy papers, benchmark descriptions, and methodological perspective articles were not treated as primary evidence sources. Instead, these studies were retained selectively to support conceptual framing, technical background discussion, and methodological interpretation.

Studies were excluded if they: (i) lacked healthcare relevance, (ii) did not involve retrieval-grounded or reasoning-oriented architectures, (iii) contained only speculative discussion without system-level implementation, or (iv) lacked sufficient methodological detail for interpretation.

This classification strategy was introduced specifically to address the distinction between conceptual discussion and empirical evidence synthesis.[14,15]

2.3. PRISMA-oriented study selection process

Although the present review is qualitative rather than meta-analytic in nature, the literature screening process followed a PRISMA-oriented workflow to improve transparency and reproducibility. Records identified from multiple databases underwent sequential stages including duplicate removal, title and abstract screening, full-text eligibility assessment, and final qualitative inclusion (Figure 1).The initial search identified 1,452 records. After removal of 324 duplicate entries, 1,128 studies underwent title and abstract screening. A total of 972 records were excluded during preliminary screening because they were unrelated to healthcare applications, lacked retrieval-grounded architectures, or focused on general-purpose language modeling without medical relevance.

The remaining 156 articles underwent full-text assessment. Eighty-six studies were subsequently excluded because they lacked empirical evaluation, provided insufficient implementation detail, or focused exclusively on conceptual discussion without implemented medical systems. Ultimately, 70 studies were included in the final qualitative synthesis.

To further improve transparency, representative studies included in the synthesis were categorized according to architectural focus, retrieval strategy, evaluation setting, and healthcare application domain.

2.4. Methodological quality assessment

Because the included literature spans heterogeneous study designs ranging from benchmark-oriented experiments to deployment-focused healthcare systems, formal quantitative risk-of-bias scoring was not feasible. Nevertheless, methodological quality was evaluated qualitatively using principles adapted from PROBAST, TRIPOD-AI, and recent AI/ML reporting recommendations.

Particular attention was given to:

(i) external validation;
(ii) transparency of retrieval pipelines;
(iii) reproducibility of embedding and reranking configurations;
(iv) evidence traceability;
(v) robustness across retrieval settings;
(vi) clinical realism of evaluation scenarios;
(vii) physician involvement and deployment relevance.

Studies relying exclusively on closed benchmark datasets without external validation were interpreted more cautiously than studies incorporating physician

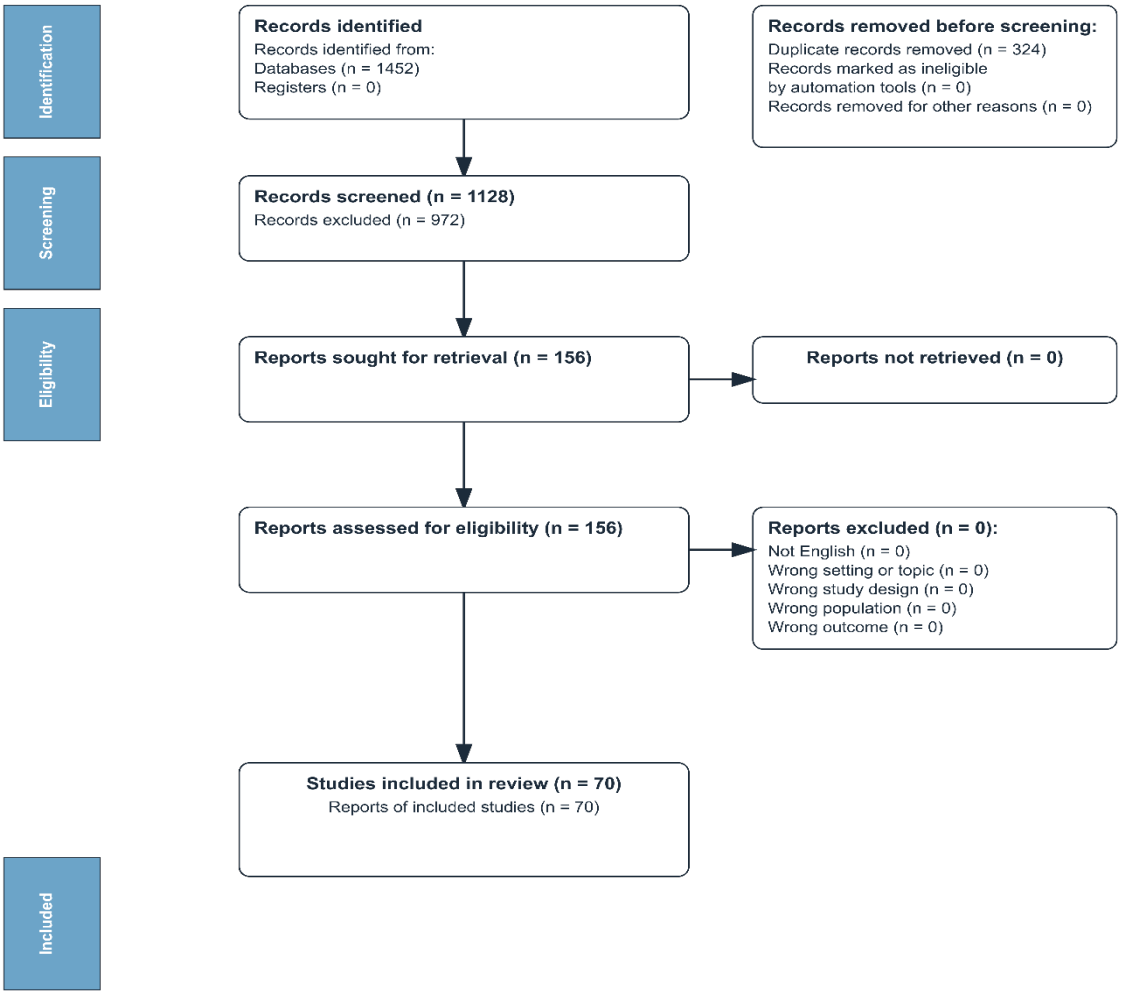

Figure 1. The overall study selection process

review, real-world workflows, or deployment-oriented evaluation.[17–22]

This quality-oriented interpretation strategy was incorporated throughout the review to avoid treating all benchmark improvements as equally clinically meaningful.

2.5. Evidence synthesis strategy

Given the substantial heterogeneity across datasets, retrieval architectures, evaluation metrics, and clinical application domains, quantitative meta-analysis was not considered methodologically appropriate. Instead, this review adopted a structured qualitative synthesis strategy.

The final synthesis organized included studies into several major thematic categories:

(i) medical large language models and retrieval foundations;
(ii) conventional medical RAG systems;
(iii) agentic and iterative retrieval frameworks;
(iv) graph-enhanced and multimodal architectures;
(v) clinical deployment and evaluation studies.

This thematic organization was designed to support comparison across retrieval paradigms while simultaneously highlighting the transition from static retrieval pipelines toward adaptive, reasoning-oriented healthcare AI systems.

Representative studies included in the final synthesis are summarized in Table 1.

3. Medical LLMs

3.1. Foundation models for healthcare

The development of medical LLMs marks an important transition in healthcare AI, because general-purpose models often lack the domain knowledge, terminology control, and safety orientation required for clinical use.[4,5,37–39] Medical LLMs attempt to close this gap through specialized pre-training on biomedical literature, clinical corpora, and healthcare-specific instruction data.[37–39]

Med-PaLM and Med-PaLM 2 represent major milestones in medical LLM development.[23] Med-PaLM improved performance on medical question answering benchmarks, while Med-PaLM 2 further increased MedQA performance to 86.5% through stronger base models, medical-domain tuning, and improved reasoning and grounding strategies.[23] Human evaluation also suggested that clinicians preferred many Med-PaLM 2 responses to comparator answers across multiple clinical quality dimensions.[23]

Smaller domain-specific models offer an alternative path that may better support privacy-conscious and resource-constrained settings.[37,38] BioMedLM, for example, showed that a 2.7B-parameter model trained on PubMed abstracts and full articles can remain competitive on biomedical tasks, including MedMCQA and medical genetics benchmarks, while offering a more transparent and efficient deployment profile.[38,41]

Table 1. Representative studies included in the qualitative synthesis

Review theme	Representative studies	System characteristics	Evaluation setting	Key contribution
Medical LLM foundations	4,23	Medical instruction-tuned LLMs	MedQA, physician evaluation	Established medical LLM baselines
Conventional medical RAG	19,24,25	Dense retrieval + grounded generation	MedQA, PubMedQA	Reduced hallucination and improved evidence grounding
Agentic and iterative RAG	17,26,27	Iterative retrieval and reasoning loops	Clinical QA benchmarks	Improved multi-step diagnostic reasoning
Graph-based and structured retrieval	28–30	Knowledge graph-guided retrieval	Medical reasoning benchmarks	Enhanced structured evidence integration
Multimodal and radiology systems	31–33	Vision-language retrieval and reasoning	Radiology and multimodal tasks	Improved multimodal evidence alignment
Verification and trustworthy AI	20,34	Verification-oriented reasoning	Medical QA and safety evaluation	Improved reasoning faithfulness and reproducibility
Clinical deployment and governance	10,18,35,36	Workflow-oriented healthcare AI	Clinical simulation and workflow analysis	Focused on traceability, usability, and governance

Abbreviations: AI: Artificial intelligence; LLM: Large language model; RAG: Retrieval-augmented generation.

BioGPT introduced a biomedical generative Transformer trained on large-scale literature[39], addressing limitations in earlier biomedical language models that were stronger in representation than in generation, such as BioBERT[44] and PubMedBERT.[45] It achieved 78.2% accuracy on PubMedQA and showed strong performance on biomedical text generation and relation extraction tasks.[39,42,44,45]

PMC-LLaMA represents an open-source route for adapting general-purpose foundation models to medicine through targeted data construction and domain-oriented tuning.[37] Despite its relatively modest size, it performed competitively across medical benchmarks and illustrated the practical value of transparent, openly modifiable models in healthcare research settings[37] (Table 2).

3.2. Clinical knowledge encoding

Clinical knowledge in medical LLMs must be judged along several dimensions, including factual accuracy, reasoning quality, comprehension, and safety.[4,23] Human evaluation protocols have therefore become important complements to benchmark scores, because they reveal where clinically fluent responses still fall short of reliable expert-level use.[23]

Parameter-efficient instruction tuning has become a practical strategy for aligning medical LLMs to new tasks with limited exemplars or restricted computing budgets.[37,38]

Table 2: Comparison of medical large language models

Model	Parameters	Training data	Key features	Performance
Med-PaLM[4]	540B	MultiMedQA datasets	Instruction prompt tuning	67.6% on MedQA[40]
Med-PaLM 2[23]	340B	Medical domain fine-tuning	Chain of retrieval	86.5% MedQA
BioMedLM[38]	2.7B	PubMed abstracts & articles	Small, privacy-preserving	57.3% on MedMCQA[38,41]
BioGPT[39]	355M/ 1.5B	Biomedical literature	Generative pre-training	78.2% on PubMedQA[39,42]
PMC-LLaMA[37]	13B	4.8M papers + 30K textbooks	Open-source, lightweight	Competitive with ChatGPT on some tasks[37,43]

Even so, safer deployment still depends on robust evaluation and supervision, because prompt alignment alone does not resolve persistent problems of factual drift, incomplete reasoning, or hidden bias in clinical settings.[4,23]

Progress toward expert-level medical question answering also depends on stronger long-form reasoning, better workflow fit, and more realistic evaluation conditions.[23] Med-PaLM 2 showed clear gains across several datasets and adversarial settings, but these results also reinforced that benchmark improvement does not automatically translate into routine clinical usefulness.[23]

4. RAG in medicine

4.1. Fundamental RAG architecture

Retrieval-augmented generation combines parametric model knowledge with non-parametric external evidence, allowing responses to be conditioned on retrievable sources rather than only on what is stored in model weights.[11,12,18] In healthcare, this design is attractive because it can support more current, inspectable, and context-sensitive responses than parameter-only systems[18] (Figure 2).

The retriever component usually relies on dense passage retrieval or related embedding-based architectures that map questions and passages into a shared search space.[46] Compared with sparse methods such as BM25, dense retrievers have shown clear gains on open-domain retrieval benchmarks, while the generator conditions its output on the returned passages to produce more specific and better-supported responses.[46]

REALM was an early and influential example of retrieval-augmented language model pre-training.[12] By coupling a latent retriever to language model training and inference, it showed how external document access could make knowledge use more modular and less dependent on purely parametric storage.[12]

4.2. Medical RAG benchmarks

Evaluating medical RAG requires benchmarks that capture both clinical content and retrieval behaviour.[19,40,41] MIRAGE, for example, was designed specifically for retrieval-augmented medical question answering and showed that RAG can improve accuracy by up to 18% over strong prompting baselines across multiple model and corpus combinations.[19]

MedQA remains a widely used benchmark because it provides board-style medical examination questions in multiple languages[40], while MedMCQA offers a much larger multi-subject benchmark covering thousands of medical topics.[41] Together, these resources help compare retrieval design and reasoning control under standardized conditions, even though they remain simpler than real clinical use[40,41] (Table 3).

4.3. Medical-specific adaptations

Medical RAG requires domain-specific adaptations because healthcare retrieval must handle specialized terminology, heterogeneous evidence, and the expectation of evidence-based answers.[49–51] Hybrid retrieval strategies that combine semantic and lexical signals have proved useful in biomedical settings, where neither surface matching nor embeddings alone are always sufficient.[49,50]

Clinical entity augmented retrieval (CLEAR) is one example of this adaptation.[52] By retrieving over structured clinical entities rather than only over dense embeddings, it improved extraction accuracy while sharply reducing token usage and inference time in large-scale clinical note processing.[52]

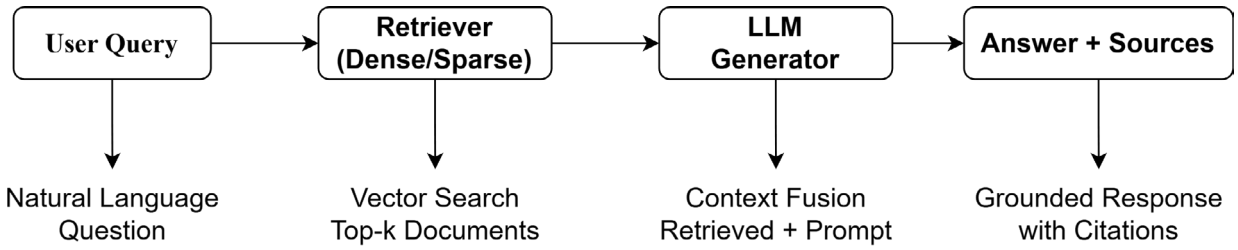

Figure 2. Basic RAG pipeline

Table 3. Medical RAG application scenarios

Application domain	Key requirements	RAG benefits	Example systems
Medical QA benchmarks	Factual accuracy, evidence use	Benchmarking retrieval gains	MIRAGE[19]
Clinical decision support	Guideline adherence, traceability	Evidence-linked recommendations	MedRAG[24], hepatology agentic Graph RAG[47]
Rare disease reasoning	Multi-step search, phenotype synthesis	Traceable diagnostic chains	Deep-DxSearch[17], DeepRare[48]
Radiology and multimodal support	Image-text alignment, cross-modal context	Bridging multimodal evidence	RAGSemRad[31]

Rationale-guided retrieval augmented generation (RAG[2]) addresses three recurring problems in medical RAG: vulnerability to irrelevant context, poorly targeted queries, and retriever bias toward particular source corpora.[51] The framework combines a lightweight filtering model trained on perplexity-based rationale labels, rationale-guided query generation, and balanced retrieval across multiple biomedical corpora. Across three medical question-answering benchmarks, it improved strong baseline models by up to 6.1%.

5. Agentic RAG

5.1. Agentic retrieval and adaptive clinical reasoning

Traditional RAG systems generally follow a fixed retrieve-then-generate pipeline in which retrieval occurs only once before response generation.[17] Although this architecture improves access to external knowledge, it remains limited when medical reasoning requires sequential evidence exploration, uncertainty resolution, or dynamic hypothesis refinement during complex clinical tasks.[17,53]

Agentic RAG extends conventional pipelines by introducing autonomous or semi-autonomous reasoning behaviors capable of dynamically coordinating retrieval, planning, tool use, and evidence evaluation throughout the reasoning process.[54] Rather than treating retrieval as a static preprocessing step, agentic architectures allow the system to iteratively refine search trajectories according to intermediate reasoning states and evolving evidence requirements.[54]

This capability is particularly important in healthcare because clinical reasoning frequently depends on progressive hypothesis testing rather than direct answer retrieval. Diagnostic interpretation often requires integrating heterogeneous evidence sources, resolving conflicting findings, and adapting subsequent information-seeking strategies according to partially observed clinical patterns. Consequently, the value of agentic retrieval may lie less in autonomy itself than in the ability to support clinically grounded reasoning trajectories through iterative evidence interaction.[17,35,36,55–57]

Several recent systems illustrate this transition toward adaptive medical reasoning architectures. Agentic retrieval frameworks have been explored for clinical decision support, multi-step medical question answering, and evidence-grounded diagnostic reasoning.[35,36,55–57] These systems commonly incorporate iterative retrieval, reasoning decomposition, retrieval verification, or dynamic

planning mechanisms that allow evidence gathering to evolve alongside the reasoning process itself.[35,36,55–57]

Compared with conventional single-pass medical RAG pipelines, agentic architectures may improve evidence completeness, reasoning flexibility, and handling of diagnostically ambiguous scenarios. However, they also introduce additional challenges related to orchestration complexity, computational overhead, reasoning controllability, and verification of intermediate reasoning trajectories.[54,57]

Taken together, these developments suggest that healthcare RAG systems are increasingly evolving from static retrieval pipelines toward adaptive evidence-grounded reasoning environments capable of supporting more complex forms of clinical inference (Figure 3).

5.2. Multi-agent collaboration

Multi-agent systems introduce collaborative reasoning patterns that resemble expert consultation in clinical settings.[35,36,56] By distributing subtasks such as retrieval, synthesis, and adjudication across different agents, these systems can make complex evidence use more structured and more inspectable.[35,36,56]

Experimental results on four benchmark medical QA datasets show that Discuss-RAG consistently outperforms MedRAG, improving answer accuracy by up to 16.67% on BioASQ and 12.20% on PubMedQA. This collaborative approach addresses the lack of modeling for humanlike reasoning behaviors during information retrieval and reduces reliance on suboptimal medical corpora.[36]

AIPatient illustrates a different application of agentic design.[35] It uses multiple task-specific agents to simulate patient interaction and complex reasoning, achieving high question-answering accuracy while also showing educational value in studies with medical learners.[35]

MedRAG-Agent represents another multi-agent design, combining query decomposition, knowledge-graph navigation, document retrieval, and verification within one healthcare-oriented pipeline.[57] Reported gains in MEDQA accuracy and hallucination reduction suggest that multi-agent structure can help when retrieval and reasoning must be coordinated rather than treated as separate stages[57] (Figure 4).

5.3. End-to-end agentic training

End-to-end training of agentic RAG systems through reinforcement learning makes it possible to optimize both retrieval behavior and diagnostic reasoning together. Deep-DxSearch, for example, trains the model as an active investigator that formulates queries, evaluates intermediate

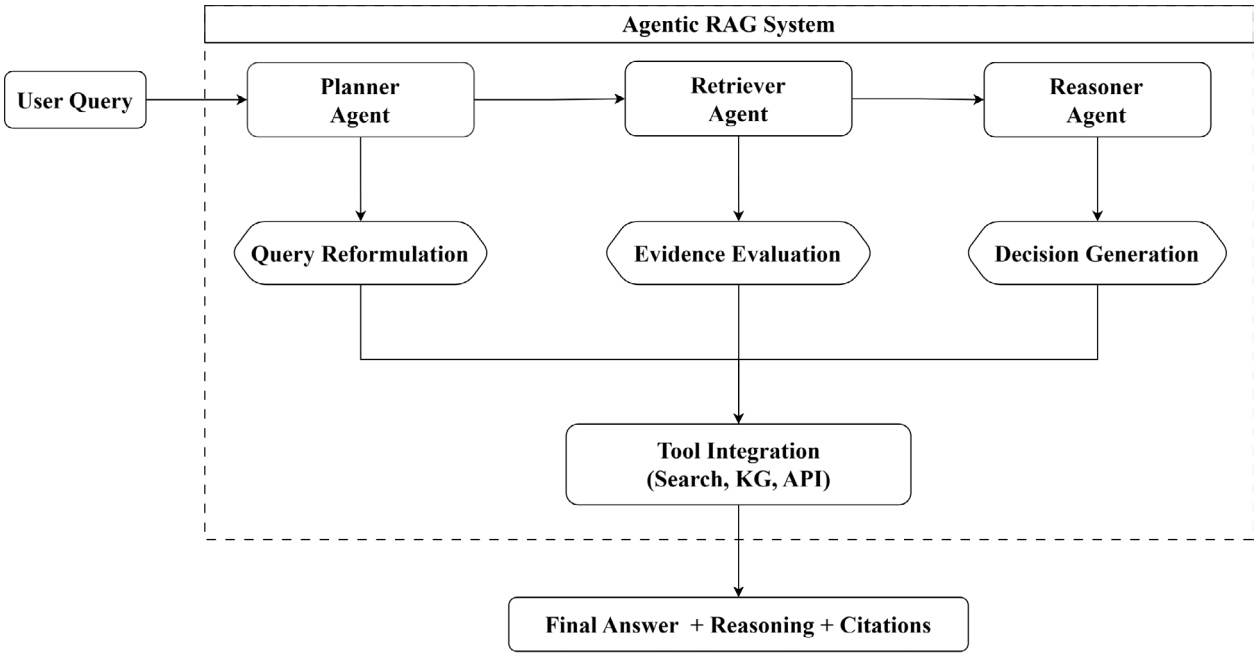

Figure 3. Agentic RAG architecture

evidence, and revises its search path while pursuing a diagnosis.[17]

Reported results suggest that this end-to-end agentic training can outperform prompt-engineering and training-free RAG baselines on both in-distribution and out-of-distribution diagnostic benchmarks.[17] In addition, physician-assisted evaluation on real-world cases indicates that the framework may improve practical diagnostic performance, not only benchmark scores.[17]

DeepRare extends this logic to rare disease diagnosis by integrating many specialized tools, phenotype resources, and current knowledge sources within one evidence-linked reasoning pipeline.[48] Its use of heterogeneous inputs and ranked, traceable hypotheses reflects the type of diagnostic assistance that is particularly relevant to sparse and difficult clinical cases.[48]

6. Advanced retrieval and reasoning architectures

6.1. Graph-enhanced retrieval architectures

Graph-enhanced retrieval architectures extend conventional retrieval-augmented generation by introducing structured relationships among biomedical entities, clinical concepts, and retrieved evidence.[58] Rather than retrieving isolated passages independently, these systems organize evidence through graph-aware structures that support semantic linkage, multi-hop reasoning, and relational knowledge exploration.[58] This design is particularly relevant in healthcare because diagnostic and therapeutic reasoning often depends on explicit relationships among symptoms, diseases, medications, laboratory findings, and treatment pathways.

Medical Graph RAG (MedGraphRAG) applies this idea to healthcare by building hierarchical graph structures from extracted entities and foundational medical knowledge, then using graph-aware retrieval to balance global awareness with efficient indexing.[28] MedSumGraph[59] further extends graph-based retrieval by adding structured medical summaries and optimized prompting. This design helps the model interpret domain-specific knowledge without additional training and improves multiple-

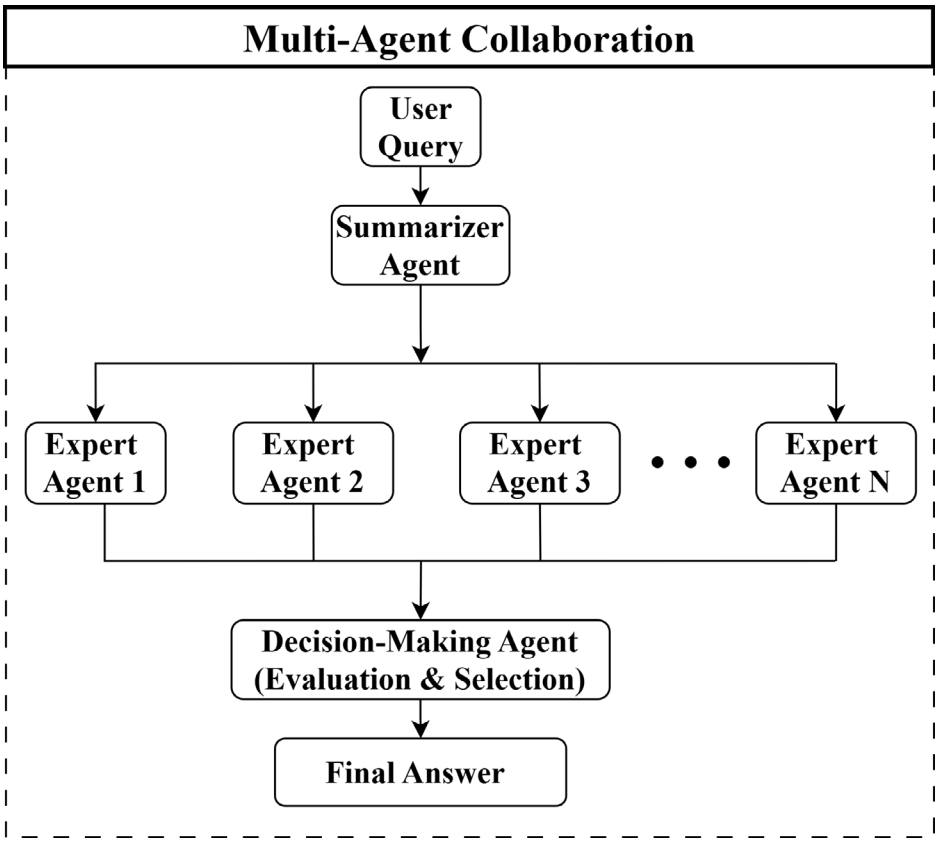

Figure 4. Multi-agent collaboration flow

choice medical QA performance relative to several strong baselines.[59]

LLM-KGMQA proposes a multi-hop medical question-answering framework that uses a knowledge graph to support entity linking and stepwise reasoning.[30] Its reported gains across one-hop to three-hop reasoning tasks illustrate the value of structured retrieval when questions require explicit relation chaining.[30] Agentic Medical Graph-RAG (AMG-RAG) combines graph construction, graph updating, reasoning, and external evidence retrieval within one agentic framework.[27] On MEDQA and MEDMCQA, it reported competitive performance while also addressing the practical need to keep medical knowledge structures current over time.[27]

KGAREVION adds a verification layer to graph-based reasoning by generating candidate triplets from latent model knowledge and then checking them against a grounded knowledge graph.[60] This design improved accuracy across a broad set of biomedical question-answering models while reducing unsupported graph content.[60]

Taken together, graph-enhanced systems illustrate a broader transition from passage-level retrieval toward structured evidence reasoning. Compared with conventional dense retrieval pipelines, graph-based architectures may improve interpretability and multi-hop clinical reasoning, although they also introduce additional complexity related to graph construction, ontology maintenance, and computational scalability.

6.2. Iterative and recursive retrieval frameworks

Conventional medical RAG systems generally rely on single-pass retrieval in which evidence is collected only once before generation. However, many clinical reasoning tasks require progressive evidence refinement, follow-up search, and hypothesis revision. Iterative and recursive retrieval frameworks address this limitation by enabling multi-round evidence gathering and adaptive query reformulation during reasoning.[17,26]

Iterative retrieval methods enable multi-round evidence gathering when a single query is unlikely to surface enough relevant support. i-MedRAG exemplifies this design by allowing the model to generate follow-up queries based on what earlier retrieval rounds returned.[26] Across clinical-vignette-style benchmarks, i-MedRAG improved performance over conventional one-pass RAG and showed how follow-up search can help build more informative reasoning chains for difficult medical questions.[26]

BioRAG extends iterative retrieval into life science reasoning by decomposing questions and combining search with a domain-specific knowledge hierarchy, thereby supporting more structured step-by-step evidence gathering.[25] Recurrence generation-augmented retrieval (RGAR)[29] retrieves both factual and conceptual knowledge from dual sources and lets these sources refine one another through iteration. Its reported results on factual-aware medical QA benchmarks suggest that iterative interaction between knowledge types can outperform traditional one-pass medical RAG pipelines.[29]

These architectures suggest that retrieval in healthcare AI is increasingly evolving from a static preprocessing stage into an adaptive reasoning process. Iterative retrieval may improve evidence completeness and diagnostic exploration, although it also increases orchestration complexity, latency, and computational cost.

6.3. Retriever–generator coordination and fusion

As healthcare RAG systems become increasingly complex, recent research has shifted from optimizing retrievers and generators independently toward improving coordination between retrieval and reasoning components.[46,61] In clinical settings, retrieval quality directly affects downstream reasoning fidelity, while generator behavior influences how retrieved evidence is interpreted, prioritized, and integrated into final responses.

The retriever component determines which external materials are even available to the generator, so errors at this stage propagate throughout the entire pipeline.[46] Medical-specific retrievers must therefore account for biomedical terminology, corpus heterogeneity, and the practical need to rank both highly similar and clinically complementary evidence.[49,50] MedBioRAG, for instance, combines semantic and lexical retrieval to improve biomedical document ranking and downstream question answering.[50]

Other systems focus on retrieval efficiency and evidence organization. CLEAR retrieves over clinically meaningful entities rather than over full-note embeddings.[52] In large-scale note processing, this reduced computational cost while maintaining or improving extraction quality.[52]

On the generation side, fusion mechanisms increasingly attempt to integrate retrieved evidence with parametric model knowledge in a clinically interpretable manner.[11,61] Sequence-to-sequence architectures remain useful because they can aggregate information across multiple passages while maintaining flexible conditional generation.[61] MedRGAG proposes a unified retrieval-generation framework that combines retrieved documents with generated background knowledge[62], while MedGENIE shows that retrieval-guided medical generation can benefit from explicitly balancing external evidence and parametric completion.[63]

These coordination challenges become especially important in multimodal settings. Semantic feedback-based RAG for radiology report generation (RAGSemRad)[31] combines fine-grained semantic retrieval with semantic assessment to strengthen alignment between imaging input and generated reports. This makes it a useful example of how retrieval–generation coordination becomes increasingly critical once systems must reason across heterogeneous medical modalities.[31]

Taken together, these studies suggest that future healthcare RAG systems may increasingly function as tightly coordinated retrieval–reasoning environments rather than loosely connected retriever and generator modules.

6.4. Verification, filtering, and trust calibration

Because healthcare applications require high reliability and evidence traceability, recent medical RAG systems increasingly incorporate verification, filtering, and trust-calibration mechanisms designed to reduce unsupported reasoning and improve grounded clinical inference.[20–22,34,51,63]

Process reward modeling addresses a central weakness of medical reasoning systems: the difficulty of localizing errors to specific reasoning steps. Med-PRM tackles this by verifying intermediate reasoning against retrieved medical guidance and scoring the process rather than only the final answer.[20] Across several medical QA benchmarks and open-ended diagnostic tasks, Med-PRM improved base-model performance and showed that stepwise reward design can strengthen smaller models as well as stronger policy backbones.[20]

ALIGNRAG reframes the problem as one of reasoning alignment, namely whether an LLM's internal reasoning stays constrained by retrieved evidence.[64] By contrast, Self-Refine offers a more general iterative refinement baseline.[65] The comparison highlights why retrieval-sensitive critique is especially important in evidence-bound settings such as healthcare.[64,65]

Filtering and quality-control modules are increasingly important because retrieval can harm performance when weakly matched or distracting passages enter the context window.[21,51,63] In healthcare settings, poorly filtered context can make an answer appear evidence-based while actually pushing the model toward unsupported or misleading conclusions.

Recent work addresses this problem through rationale-guided filtering, retrieval reordering, and selective context construction. RAG^2 uses rationale-based filtering and balanced retrieval across biomedical corpora[51], while long-context studies show that reordering and controlled context assembly can reduce the impact of hard negatives and overlong evidence lists.[21,63]

Related trustworthiness frameworks such as TRUST-SCORE and TRUST-ALIGN[22,34] shift attention from surface fluency to grounded evidence use, calibrated uncertainty, refusal behavior, and attribution quality. For healthcare deployment, filtering, trust calibration, and abstention should therefore be treated as core architectural components rather than optional refinements.

Taken together, these developments suggest that trustworthy healthcare RAG requires not only accurate retrieval, but also explicit mechanisms for evidence validation, uncertainty management, and clinically interpretable reasoning control.

6.5. Architectural trends and comparative analysis

The architectural evolution of healthcare RAG systems reveals several converging trends: graph-enhanced retrieval, iterative evidence refinement, retrieval–generator coordination, and verification-oriented reasoning. Although these approaches differ substantially in implementation strategy, they collectively reflect a transition from static document retrieval toward adaptive evidence-grounded reasoning systems.

Importantly, no single architecture appears universally optimal across all healthcare tasks. Graph-based systems may improve structured reasoning but increase maintenance complexity. Iterative retrieval frameworks improve evidence exploration while introducing additional latency and orchestration cost. Verification-oriented architectures strengthen trustworthiness but require more sophisticated evaluation infrastructure.

These trade-offs suggest that future clinically deployable healthcare RAG systems will likely integrate multiple complementary mechanisms depending on workflow requirements, modality complexity, and safety constraints.

7. Applications

7.1. Medical question answering as benchmark background

Medical question answering remains one of the most widely used benchmark settings for healthcare RAG because it enables controlled evaluation of factual grounding, retrieval quality, and hallucination reduction.[19,40,41] Benchmarks such as MIRAGE[19] and MedQA are therefore useful primarily as comparative evaluation frameworks rather than as direct indicators of clinical readiness.

Recent systems incorporating iterative retrieval,

evidence grounding, and agentic reasoning have further improved benchmark-oriented medical QA performance.[36,57,60] However, examination-style datasets often simplify uncertainty, longitudinal context, and real-world clinical complexity.[40,41] Consequently, the primary value of medical QA lies in evaluating retrieval and reasoning strategies that may later transfer into more clinically consequential applications such as decision support and rare disease diagnosis.

7.2. Clinical decision support

Clinical decision support is the strongest near-term application for RAG in healthcare because it requires more than fluent answer generation.[24,47] Useful systems must connect patient-specific information with guidelines, retrieved evidence, and explicit reasoning steps that clinicians can inspect. This makes evidence grounding, traceability, and controlled updating central design requirements rather than optional improvements.

MedRAG illustrates this direction by combining retrieval with knowledge-graph-elicited reasoning for healthcare copilot use.[24] Instead of relying only on parametric recall, it retrieves diagnosis and treatment recommendations from a structured diagnostic hierarchy, which better supports differential reasoning and guideline-aware output. A self-correcting agentic graph RAG in hepatology extends this logic through a retrieve-evaluate-refine loop, reporting strong faithfulness, context recall, and answer relevancy in a focused decision-support setting.[47]

Related multi-agent decision-support systems suggest that the most promising designs are those that adapt retrieval strategy to query type, knowledge source, and required level of justification.[47,57] Compared with medical question answering benchmarks, clinical decision support better reflects the practical demand for systems that not only answer correctly, but also show where their recommendations come from and how they should be verified before use.

7.3. Rare disease reasoning

Rare disease reasoning is a particularly appropriate setting for agentic RAG because the diagnostic process is usually sparse, multi-step, and evidence-hungry.[17,48] Cases often require synthesis across phenotype descriptions, laboratory findings, genetic results, and scattered literature, so single-pass retrieval is frequently insufficient. This makes iterative search, query reformulation, and explicit hypothesis refinement more valuable than in routine benchmark-style tasks.

Deep-DxSearch demonstrates this pattern clearly.[17] By training the model as an active investigator that formulates queries, evaluates evidence utility, and refines its search trajectory, the system improves common and rare disease diagnosis beyond prompt-engineering and training-free RAG baselines.[17] Its reported gains on real-world cases are important not only because of accuracy, but because they show how retrieval and reasoning can be co-optimized for diagnostic work.

DeepRare extends this line of work by integrating a wide set of specialized tools, phenotype resources, and up-to-date knowledge sources for rare disease diagnosis.[48] Its use of heterogeneous clinical inputs and ranked, evidence-linked hypotheses reflects the type of traceable diagnostic assistance that healthcare settings are more likely to value.[48] Among current applications, rare disease reasoning is therefore one of the clearest demonstrations of why agentic retrieval can matter beyond benchmark performance.

7.4. Radiology and multimodal extension

Radiology and multimodal interpretation have rapidly emerged as important frontiers in healthcare RAG research because many clinical workflows require integrating imaging findings with textual reports, laboratory results, prior examinations, and broader patient context.[28,31] Compared with text-only retrieval systems, multimodal healthcare RAG architectures must additionally address cross-modal alignment, semantic consistency, visual grounding, and heterogeneous evidence integration.

In radiology settings, retrieval quality becomes inseparable from multimodal consistency because generated outputs must remain aligned not only with retrieved textual evidence, but also with imaging findings themselves. Systems such as RAGSemRad[31] demonstrate that fine-grained semantic retrieval can improve radiology report generation by filtering irrelevant context and strengthening alignment between images and narrative output. Related multimodal medical QA and graph-based systems further suggest that multimodal retrieval may improve the interpretability and reliability of rationale-supported clinical responses.[28,31]

Recent medical vision-language models have substantially accelerated progress in multimodal medical reasoning. LLaVA-Med adapts conversational multimodal learning to biomedical imaging tasks through medical visual instruction tuning, enabling image-grounded dialogue and clinically oriented multimodal interaction.[32] Similarly, CheXAgent introduces a clinically specialized vision-language assistant capable of radiographic interpretation, visual dialogue, and multimodal reasoning across chest imaging tasks.[33] These systems illustrate how multimodal medical foundation models may increasingly

combine image understanding, conversational reasoning, and retrieval-grounded evidence integration within unified clinical AI architectures.

Another emerging direction involves retrieval-enhanced multimodal reasoning in which imaging evidence is dynamically integrated with external biomedical knowledge during inference. Such architectures may improve diagnostic contextualization because radiographic interpretation often depends on combining image findings with disease knowledge, patient history, and evolving diagnostic hypotheses. This trend also aligns closely with recent agentic reasoning frameworks capable of adaptively coordinating multimodal retrieval according to intermediate reasoning states.

Beyond radiology, multimodal healthcare RAG may eventually support integrated reasoning across pathology slides, genomics, wearable monitoring, surgical video analysis, and longitudinal electronic health records. However, these systems also introduce substantial challenges related to interpretability, computational cost, modality imbalance, evidence attribution, and clinically reliable cross-modal reasoning.[28,31–33]

8. Challenges

8.1. Evidence grounding and factual validity

The first deployment barrier is not whether a model can generate plausible medical language, but whether its claims remain anchored to retrievable and clinically relevant evidence.[14,18] Hallucination remains a serious problem in healthcare, and retrieval alone does not remove it. If the search process surfaces weak, outdated, or poorly matched passages, the generator may still produce confident but unsupported conclusions.

This means that factual validity depends on the whole pipeline rather than on the mere presence of a retrieval module.[14,18] Retrieval-aware self-reflective systems can help by conditioning generation on explicit retrieval decisions and internal critique, but they remain sensitive to corpus quality, passage selection, and evidence relevance.[21,22]

For healthcare use, the practical question is therefore whether the system can show that an answer is grounded in appropriate evidence, not merely whether it cites something. Deployment-facing evaluation should distinguish between fluent unsupported output, weakly supported output, and recommendations that are genuinely traceable to current medical sources.

8.2. Verification, traceability, and evaluation

Even when relevant evidence is retrieved, a second barrier remains: the reasoning process may drift away from that evidence.[20,64] Reasoning alignment problems arise when an LLM produces conclusions that sound coherent but are not faithfully constrained by the retrieved record. This is especially important in medicine, where intermediate reasoning steps may determine whether a recommendation appears safe or unsafe to a clinician.

Verification mechanisms therefore matter as much as retrieval quality. Approaches such as critique-driven alignment[64] and process reward modeling exemplified by Med-PRM[20] are important because they evaluate reasoning step by step rather than only scoring final answers. These methods move the field closer to traceable support, where a clinician can inspect not just the output, but the path by which it was reached.

Evaluation also remains underdeveloped. Reproducibility studies show that performance can shift with changes in embedding models, hardware settings, retrieval precision, or pipeline configuration.[21,22,34] For that reason, stronger evaluation in healthcare should include benchmark performance, evidence traceability, reproducibility, and expert review, rather than relying on single aggregate scores.

8.3. Workflow Fit and clinical integration

A technically strong RAG pipeline may still fail in practice if it does not fit clinical workflow.[10,18] Healthcare environments require systems that are timely, legible, easy to validate, and compatible with existing documentation practices. If retrieval takes too long, presents excessive context, or produces reasoning chains that are cumbersome to inspect, routine use becomes unlikely regardless of benchmark quality.

This issue is particularly visible in decision-support settings, where the value of a system depends on how well retrieved evidence can be incorporated into real diagnostic or treatment discussions. Local and enterprise RAG architectures may help by improving privacy control and system integration[10,18], but they do not solve the broader problem of workflow fit. Outputs still need to be reviewable, clinically interpretable, and aligned with how physicians actually use guidelines, reports, and patient records.

For these reasons, deployment studies should examine usability, latency, updateability, and human review burden alongside accuracy. A healthcare RAG system that is correct but awkward to use may be less valuable than a slightly simpler system that is easier to supervise and integrate into care processes.

8.4. Governance, accountability, and clinical oversight

Healthcare deployment of RAG systems involves

challenges that extend beyond model accuracy alone. In clinical environments, retrieval-grounded systems must operate within institutional workflows that require data security, patient privacy protection, evidence traceability, and regulatory compliance. Unlike general-purpose AI systems, healthcare RAG architectures may directly influence diagnostic interpretation, treatment planning, and clinical communication, thereby increasing the importance of governance-oriented safeguards.[10,18]

Another major concern involves accountability during retrieval-grounded clinical reasoning. Because generated responses may depend on dynamically retrieved evidence rather than only parametric model knowledge, healthcare institutions must ensure that retrieved sources remain transparent, auditable, and clinically appropriate throughout the reasoning process. Errors introduced through outdated retrieval indexes, incomplete evidence selection, or poorly governed retrieval pipelines may produce clinically misleading outputs despite apparently coherent reasoning.[18,21]

Healthcare deployment therefore requires explicit governance mechanisms covering auditability, knowledge-source version control, access management, and clinician accountability. Recent healthcare-focused analyses emphasize that retrieval-grounded AI systems should support transparent evidence logging, institutional oversight, and controlled updating of medical knowledge sources in order to remain clinically trustworthy.[10,18] Governance concerns additionally extend beyond privacy protection to include bias monitoring, escalation pathways, and human review responsibility in decision-support settings.[10,18]

Clinical oversight also remains essential because healthcare RAG systems may exhibit variable reasoning behavior across retrieval settings, prompting strategies, and institutional data environments. Consequently, deployment-oriented evaluation should include not only benchmark performance, but also reproducibility assessment, workflow integration analysis, and monitoring of clinically unsafe reasoning patterns.[34] Human-in-the-loop supervision and physician verification therefore remain important safeguards, particularly for high-risk diagnostic and therapeutic applications.

9. Future directions

9.1. Multimodal and cross-modal integration

Future healthcare RAG systems are increasingly moving beyond text-only retrieval pipelines toward integrated multimodal reasoning architectures capable of jointly processing imaging, pathology, clinical narratives, laboratory findings, and longitudinal patient records. Such systems aim to support evidence-grounded reasoning across heterogeneous clinical modalities rather than relying exclusively on textual retrieval.[31]

Recent multimodal medical foundation models such as LLaVA-Med[32] and CheXAgent[33] illustrate how conversational reasoning can be combined with clinically grounded visual interpretation. These systems extend retrieval-grounded reasoning into radiology and multimodal diagnostic settings by integrating visual evidence, semantic retrieval, and interactive language generation within unified reasoning pipelines.[32,33]

Semantic feedback-based RAG for radiology report generation (RAGSemRad) further demonstrates that multimodal retrieval requires not only access to heterogeneous evidence sources, but also mechanisms capable of aligning generated outputs with clinically meaningful image representations.[31] Compared with conventional text-only medical QA systems, multimodal healthcare RAG therefore introduces additional challenges related to cross-modal consistency, visual grounding, temporal alignment, and computational scalability.

Future multimodal healthcare systems may ultimately integrate imaging, pathology, genomics, wearable monitoring, and longitudinal electronic health records into unified evidence-grounded reasoning environments. However, ensuring traceability and interpretability across heterogeneous modalities will remain a major challenge for clinically deployable multimodal AI systems.[31–33]

9.2. Human-in-the-loop systems

Human oversight will remain central even as retrieval and reasoning become more adaptive. In healthcare, human-in-the-loop design is not simply a safeguard added after model development. It is part of the operating logic of a responsible system, because clinicians still need to guide evidence use, inspect reasoning, and decide whether retrieved support is sufficient for action.[66,67]

Future systems should therefore be designed to make intervention easy rather than exceptional. Review checkpoints, editable retrieval traces, and explicit escalation paths are more valuable for deployment than the appearance of full automation, particularly in high-stakes decision-support settings.[66,67]

9.3. Standardized evaluation frameworks

Standardized evaluation remains underdeveloped. Existing benchmarks are useful, but they do not yet capture enough of the conditions that matter in clinical use, such as changing evidence bases, heterogeneous patient records,

reproducibility across pipeline configurations, and the burden placed on human reviewers.[19,68]

More mature evaluation should therefore combine benchmark accuracy with evidence traceability, expert review, reproducibility testing, and workflow-oriented assessment. Without this broader view, reported gains may say more about a particular setup than about readiness for responsible healthcare deployment.[19,68]

9.4. Trustworthy deployment and governance frameworks

Future healthcare RAG systems will likely require integrated governance frameworks capable of supporting transparent evidence attribution, longitudinal audit trails, and clinician-in-the-loop supervision throughout the reasoning pipeline.[10,18] As retrieval-grounded AI systems become increasingly embedded within clinical workflows, governance mechanisms must evolve beyond static privacy protection toward dynamic oversight structures that support real-time monitoring, escalation management, and institution-level accountability.

In addition to model performance, future deployment frameworks may need to incorporate controlled retrieval-index updating, evidence version tracking, and retrieval reproducibility monitoring to ensure that clinical recommendations remain traceable and scientifically defensible over time.[34] These considerations are particularly important because healthcare RAG systems may continuously interact with evolving biomedical knowledge sources, institutional retrieval infrastructures, and heterogeneous patient populations.

Future trustworthy deployment may also require standardized reporting protocols describing retrieval configuration, indexing strategies, evidence provenance, and retrieval-update procedures. Such mechanisms could improve reproducibility, facilitate regulatory evaluation, and strengthen institutional confidence in retrieval-grounded clinical AI systems.[18,34]

Another important direction involves integrating governance-aware safeguards directly into the retrieval and reasoning pipeline itself. Rather than treating governance as an external administrative layer, future healthcare RAG systems may increasingly incorporate retrieval filtering, uncertainty calibration, abstention mechanisms, and human escalation pathways as intrinsic architectural components.[22,34,51]

9.5. Reasoning-oriented foundation models and continual knowledge updating

Recent advances in reasoning-oriented foundation models may further influence the evolution of healthcare RAG systems. Compared with earlier instruction-following LLMs, newer architectures such as GPT-o1/o3-style systems and DeepSeek-R1-like models increasingly emphasize extended chain-of-thought reasoning, adaptive problem decomposition, self-verification, and multi-step deliberative inference.[15,17,20,64,69,70] These capabilities may improve evidence-grounded clinical reasoning because medical decision-making often requires iterative hypothesis evaluation and complex evidence synthesis rather than single-pass response generation.

Although systematic healthcare evaluation remains limited, emerging studies suggest that reasoning-oriented models may improve difficult diagnostic reasoning, uncertainty handling, and long-context evidence integration within retrieval-grounded medical workflows. In particular, combining retrieval grounding with extended reasoning trajectories may strengthen clinical inference in scenarios involving heterogeneous evidence sources, incomplete patient information, or evolving diagnostic hypotheses. These developments also align closely with recent agentic RAG frameworks that dynamically coordinate retrieval, verification, and reasoning throughout the inference process.[20,27,64]

At the same time, continual knowledge updating is becoming increasingly important because healthcare knowledge evolves rapidly through updated clinical guidelines, newly published biomedical evidence, and changing institutional protocols. Unlike static benchmark settings, real-world medical deployment requires retrieval systems capable of maintaining temporally current and institutionally relevant knowledge bases over time.[18,27]

Future healthcare RAG architectures may therefore require mechanisms supporting continual retrieval-index updating, adaptive knowledge integration, temporal version control, and longitudinal evidence maintenance without destabilizing previously learned reasoning behaviors.[27] These requirements are especially important in medicine because outdated retrieval sources or stale biomedical knowledge may directly affect clinical reliability and evidence-grounded decision support.

However, reasoning-oriented healthcare RAG systems also introduce important challenges related to computational cost, reasoning transparency, controllability of extended inference processes, and reproducibility across retrieval configurations. Multi-agent and graph-enhanced reasoning systems may also introduce substantial computational overhead and latency constraints that could limit real-time clinical deployment. As reasoning depth and retrieval complexity continue to

increase, future healthcare deployment frameworks may require stronger verification, monitoring, and governance mechanisms capable of supporting clinically interpretable and accountable reasoning processes.

10. Conclusion

This review argues that the main value of retrieval-augmented and agentic AI in healthcare does not lie in making language models sound more informed. Its value lies in whether external evidence can be retrieved, inspected, and used to constrain reasoning in ways that are clinically meaningful. From that perspective, the strongest current case is not broad medical question answering, but applications such as clinical decision support and rare disease reasoning, where evidence grounding, multi-step search, and traceable justification clearly matter.

The progression from conventional RAG to agentic frameworks is therefore important because it expands what healthcare systems can do when a single retrieval pass is not enough. Iterative retrieval, tool use, and reasoning supervision offer real advantages for complex diagnostic and support tasks. However, routine deployment still depends less on headline benchmark scores than on verification, reproducibility, workflow fit, governance, and clinician oversight. Future evaluation frameworks for healthcare RAG systems should therefore incorporate external validation, longitudinal clinical testing, and workflow-level usability assessment in addition to benchmark accuracy. The central challenge for the next stage of the field is not simply building more capable medical agents, but building systems whose evidence use is transparent enough, controlled enough, and useful enough to earn a place in clinical practice.

Acknowledgments

None.

Funding

This work was supported by JSPS KAKENHI (Grant Number JP26K14963).

Conflict of interest

The authors declare no competing interests.

Author contributions

Conceptualization: Bihong Zhu, Xin Kang
Supervision: Xin Kang, Kazuyuki Matsumoto, Minoru Yoshida
Writing–original draft: Bihong Zhu, Xin Kang
Writing–review & editing: Xin Kang, Kazuyuki Matsumoto, Minoru Yoshida

All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Rajpurkar P, Irvin J, Zhu K, *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv.* Preprint posted online 2017.

   doi: 10.48550/arXiv.1711.05225

2. Esteva A, Robicquet A, Ramsundar B, *et al.* A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29.

   doi: 10.1038/s41591-018-0316-z

3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

   doi: 10.1038/s41591-018-0300-7

4. Singhal K, Azizi S, Tu T, *et al.* Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.

   doi: 10.1038/s41586-023-06291-2

5. Zhou H, Liu F, Gu B, *et al.* A survey of large language models in medicine: Progress, application, and challenge. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2311.05112

6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940.

   doi: 10.1038/s41591-023-02448-8

7. Ondov B, Attal K, Demner-Fushman D. A survey of automated methods for biomedical text simplification. *J Am Med Inform Assoc.* 2022;29(11):1976-1988.

   doi: 10.1093/jamia/ocac149

8. Clusmann J, Kolbinger FR, Muti HS, *et al.* The future landscape of large language models in medicine. *Commun Med.* 2023;3(1):141.

   doi: 10.1038/s43856-023-00370-1

9. Yang X, Chen A, PourNejatian N, *et al.* A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.

   doi: 10.1038/s41746-022-00742-2

10. Abd-alrazaq A, AlSaad R, Alhuwail D, *et al.* Large Language

Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ*. 2023;9:e48291.

doi: 10.2196/48291

11. Lewis P, Perez E, Piktus A, *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv.* Preprint posted online 2020.

doi: 10.48550/arXiv.2005.11401

12. Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: Retrieval-augmented language model pre-training. *arXiv.* Preprint posted online 2020.

doi: 10.48550/arXiv.2002.08909

13. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719-731.

doi: 10.1038/s41551-018-0305-z

14. Abo El-Enen M, Saad S, Nazmy T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Comput Appl*. 2025;37(33):28191-28267.

doi: 10.1007/s00521-025-11666-9

15. Li Y, Zhang W, Yang Y, *et al.* Towards agentic RAG with deep reasoning: A survey of RAG-reasoning systems in LLMs. *arXiv.* Preprint posted online 2025.

doi: 10.48550/arXiv.2507.09477

16. Neha F, Bhati D. Traditional RAG vs. Agentic RAG: A comparative study of retrieval-augmented systems. *TechRxiv.* Preprint posted online August 26, 2025.

doi: 10.36227/techrxiv.175624551.12254549/v1

17. Zheng Q, Sun Y, Wu C, *et al.* End-to-end agentic RAG system training for traceable diagnostic reasoning. *arXiv.* Preprint posted online 2025.

doi: 10.48550/arXiv.2508.15746

18. Yang R, Ning Y, Keppo E, *et al.* Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* 2025;2(1):2.

doi: 10.1038/s44401-024-00004-1

19. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking Retrieval-Augmented Generation for Medicine. In: *Findings of the Association for Computational Linguistics ACL 2024*. Kerrville, USA: Association for Computational Linguistics; 2024:6233-6251.

doi: 10.18653/v1/2024.findings-acl.372

20. Yun J, Sohn J, Park J, *et al.* Med-PRM: Medical Reasoning Models with Stepwise, Guideline-verified Process Rewards. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Kerrville, USA: Association for Computational Linguistics; 2025:16565-16582.

doi: 10.18653/v1/2025.emnlp-main.837

21. Jin B, Yoon J, Han J, Arik SO. Long-context LLMs Meet RAG: Overcoming challenges for long inputs in RAG. *arXiv.* Preprint posted online 2024.

doi: 10.48550/arXiv.2410.05983

22. Song M, Sim SH, Bhardwaj R, Chieu HL, Majumder N, Poria S. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. *arXiv.* Preprint posted online 2024.

doi: 10.48550/arXiv.2409.11242

23. Singhal K, Tu T, Gottweis J, *et al.* Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950.

doi: 10.1038/s41591-024-03423-7

24. Zhao X, Liu S, Yang SY, Miao C. MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *arXiv.* Preprint posted online 2025.

doi: 10.48550/arXiv.2502.04413

25. Wang C, Long Q, Xiao M, *et al.* BioRAG: A RAG-LLM framework for biological question reasoning. *arXiv.* Preprint posted online 2024.

doi: 10.48550/arXiv.2408.01107

26. Xiong G, Jin Q, Wang X, Zhang M, Lu Z, Zhang A. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *arXiv.* Preprint posted online 2024.

doi: 10.48550/arXiv.2408.00727

27. Rezaei MR, Fard RS, Parker JL, Krishnan RG, Lankarany M. Agentic Medical Knowledge Graphs Enhance Medical Question Answering: Bridging the Gap Between LLMs and Evolving Medical Knowledge. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Kerrville, USA: Association for Computational Linguistics; 2025:12682-12701.

doi: 10.18653/v1/2025.findings-emnlp.679

28. Wu J, Zhu J, Qi Y, *et al.* Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol 1. Kerrville, USA: Association for Computational Linguistics; 2025:28443-28467.

doi: 10.18653/v1/2025.acl-long.1381

29. Liang S, Zhang L, Zhu H, Wang W, He Y, Zhou D. RGAR: Recurrence Generation-augmented Retrieval for Factual-aware Medical Question Answering. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Kerrville, USA: Association for Computational Linguistics; 2025:4006-4033.

doi: 10.18653/v1/2025.findings-emnlp.214

30. Wang F, Shi D, Aguilar J, *et al.* LLM-KGMQA: Large

language model-augmented multi-hop question-answering system based on knowledge graph in medical field. *Res Sq.* Preprint posted online August 5, 2024.

doi: 10.21203/rs.3.rs-4721418/v1

31. Jia X, Xiong Y, Pei S, Zhang Y, Yan C, Fang Z. Semantic Feedback-Based RAG for Radiology Report Generation. *Big Data Min Anal.* 2026;9(2):393-406.

doi: 10.26599/BDMA.2025.9020037

32. Li C, Wong C, Zhang S, *et al.* LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv.* Preprint posted online 2023.

doi: 10.48550/arXiv.2306.00890

33. Chen X, Reibman A, Arora S. Sequential Recommendation Model for Next Purchase Prediction. In: *Machine Learning & Applications*. Academy & Industry Research Collaboration; 2023:141-158.

doi: 10.5121/csit.2023.131013

34. Zheng D, Liu D, Lapata M, Pan JZ. TrustScore: Reference-free evaluation of LLM response trustworthiness. *arXiv.* Preprint posted online 2024.

doi: 10.48550/arXiv.2402.12545

35. Yu H, Zhou J, Li L, *et al.* Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Commun Med.* 2025;6(1):27.

doi: 10.1038/s43856-025-01283-x

36. Dong X, Zhu W, Wang H, *et al.* Talk before you retrieve: agent-led discussions for better rag in medical QA. *arXiv*. Preprint posted online April 30, 2025.

doi: 10.48550/arXiv.2504.21252

37. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *arXiv*. Preprint posted online August 25, 2023.

doi: 10.48550/arXiv.2304.14454

38. Bolton E, Venigalla A, Yasunaga M, *et al.* BioMedLM: A 2.7B Parameter Language Model Trained on Biomedical Text. *arXiv*. Preprint posted online March 27, 2024.

doi: 10.48550/arXiv.2403.18421

39. Luo R, Sun L, Xia Y, *et al.* BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform.* 2022;23(6):bbac409.

doi: 10.1093/bib/bbac409

40. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Appl Sci.* 2021;11(14):6421.

doi: 10.3390/app11146421

41. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *arXiv*. Preprint posted online March 27, 2022.

doi: 10.48550/arXiv.2203.14371

42. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv*. Preprint posted online September 13, 2019:arXiv:1909.06146.

doi: 10.48550/arXiv.1909.06146

43. Welsby P, Cheung BMY. ChatGPT. *Postgrad Med J*. 2023;99(1176):1047-1048.

doi: 10.1093/postmj/qgad056

44. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, ed. *Bioinformatics*. 2020;36(4):1234-1240.

doi: 10.1093/bioinformatics/btz682

45. Gu Y, Tinn R, Cheng H, *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthc*. 2022;3(1):1-23.

doi: 10.1145/3458754

46. Karpukhin V, Oguz B, Min S, *et al.* Dense Passage Retrieval for Open-Domain Question Answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Kerrville, USA: Association for Computational Linguistics; 2020:6769-6781.

doi: 10.18653/v1/2020.emnlp-main.550

47. Hu Y, Xuan W, Zhou Q, *et al.* A self-correcting Agentic Graph RAG for clinical decision support in hepatology. *Front Med*. 2025;12.

doi: 10.3389/fmed.2025.1716327

48. Zhao W, Wu C, Fan Y, *et al.* An agentic system for rare disease diagnosis with traceable reasoning. *Nature*. 2026;651(8106):775-784.

doi: 10.1038/s41586-025-10097-9

49. Kumari M, Chauhan R, Jain R, Garg P. A novel context-aware retrieval framework for biomedical knowledge integration with large language models. *Inf Fusion*. 2026;127:103902.

doi: 10.1016/j.inffus.2025.103902

50. Kim S. MedBioRAG: Semantic Search and Retrieval-Augmented Generation with Large Language Models for Medical and Biological QA. *arXiv*. Preprint posted online December 10, 2025.

doi: 10.48550/arXiv.2512.10996

51. Sohn J, Park Y, Yoon C, *et al.* Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. In: *Proceedings of the 2025 Conference of the Nations of*

the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Vol 1. Kerrville, USA: Association for Computational Linguistics; 2025:12739-12753.

doi: 10.18653/v1/2025.naacl-long.635

52. Lopez I, Swaminathan A, Vedula K, *et al.* Clinical entity augmented retrieval for clinical information extraction. *NPJ Digit Med*. 2025;8(1):45.

doi: 10.1038/s41746-024-01377-1

53. Singh A. Agentic RAG Systems for Improving Adaptability and Performance in AI-Driven Information Retrieval. *SSRN*. Preprint posted online 2025.

doi: 10.2139/ssrn.5188363

54. Tian F, Fang J, Ganguly D, Meng Z, Macdonald C. Am I on the Right Track? What Can Predicted Query Performance Tell Us about the Search Behaviour of Agentic RAG. *arXiv*. Preprint posted online July 14, 2025.

doi: 10.48550/arXiv.2507.10411

55. Yao S, Zhao J, Yu D, *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv*. Preprint posted online March 10, 2023.

doi: 10.48550/arXiv.2210.03629

56. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative Agents: Interactive Simulacra of Human Behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. New York USA: ACM; 2023:1-22.

doi: 10.1145/3586183.3606763

57. Yadav V, Gaurav, Rana A, Sharma S. MedRAG-Agent: Medical Query Resolution By Employing A Multi-Agent, Knowledge Graph-Enhanced RAG-Based AI Framework. In: *2025 IEEE 6th Global Conference for Advancement in Technology (GCAT)*. New York USA: IEEE; 2025:1-4.

doi: 10.1109/GCAT66372.2025.11368379

58. Han H, Wang Y, Shomer H, *et al.* Retrieval-Augmented Generation with Graphs (GraphRAG). *arXiv*. Preprint posted online January 8, 2025.

doi: 10.48550/arXiv.2501.00309

59. Kim D, Yoo S, Jeong O. MedSumGraph: enhancing GraphRAG for medical QA with summarization and optimized prompts. *Artif Intell Med*. 2026;172:103311.

doi: 10.1016/j.artmed.2025.103311

60. Su X, Wang Y, Gao S, *et al.* KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA. *arXiv*. Preprint posted online March 3, 2025.

doi: 10.48550/arXiv.2410.04660

61. Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics; 2021:874-880.

doi: 10.18653/v1/2021.eacl-main.74

62. Li L, Zhou X, Zhang Y, Wu X. From Retrieval to Generation: Unifying External and Parametric Knowledge for Medical Question Answering. *arXiv*. Preprint posted online October 21, 2025.

doi: 10.48550/arXiv.2510.18297

63. Frisoni G, Cocchieri A, Presepi A, Moro G, Meng Z. To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol 1. Kerrville, USA: Association for Computational Linguistics; 2024:9878-9919.

doi: 10.18653/v1/2024.acl-long.533

64. Wei J, Zhou H, Zhang X, *et al.* Retrieval is Not Enough: Enhancing RAG Reasoning through Test-Time Critique and Optimization. *arXiv*. Preprint posted online October 11, 2025.

doi: 10.48550/arXiv.2504.14858

65. Madaan A, Tandon N, Gupta P, *et al.* Self-Refine: Iterative Refinement with Self-Feedback. *arXiv*. Preprint posted online May 25, 2023.

doi: 10.48550/arXiv.2303.17651

66. Jia S, Bit S, Jasodanand VH, Liu Y, Kolachalama VB. Agentic memory-augmented retrieval and evidence grounding for medical question-answering tasks. *Int J Med Inform*. 2026;212:106339.

doi: 10.1016/j.ijmedinf.2026.106339

67. Shi Y, Xu S, Yang T, *et al.* MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering. *AMIA Annu Symp Proc*. 2024;2024:1011-1020.

68. Prayitno LOMY, Nurfadilah A, Saudi SB, Tsunami WD, Sajiah AM. Conversational Agent for Medical Question-Answering Using RAG and LLM. *J Artif Intell Eng Appl*. 2025;4(3):1894-1899.

doi: 10.59934/jaiea.v4i3.1077

69. Zeng Z, Cheng Q, Yin Z, *et al.* Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective. *arXiv*. Preprint posted online December 18, 2024.

doi: 10.48550/arXiv.2412.14135

70. Guo D, Yang D, Zhang H, *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;645(8081):633-638.

doi: 10.1038/s41586-025-09422-z