ORIGINAL RESEARCH **Open access**

Retrieval-Augmented Generation for Real-Time Clinical Question Answering: A Framework Integrating Electronic Health Records and Clinical Guidelines

Mohammed Al-Farsi[1*], Salim Al-Harthy[1], Nasser Al-Rawahi[2]

Abstract

Clinicians often need rapid, evidence-based answers that integrate patient-specific electronic health records (EHRs) with clinical guidelines, but existing decision support tools are limited in real-time personalization. While large language models (LLMs) offer strong medical reasoning, they are prone to hallucinations and lack direct access to local EHR data, making them unsafe for standalone clinical use; meanwhile, traditional retrieval systems cannot synthesize coherent, context-aware responses. This paper proposes a retrieval-augmented generation (RAG) framework that combines dual-source retrieval from both institutional EHRs and clinical guideline databases. The system includes an EHR indexer, a guideline repository, a semantic retriever, an LLM-based generator, and a safety filter for hallucination mitigation. By grounding outputs in retrieved patient data and evidence-based recommendations, the model improves factual reliability, explainability, and clinical trustworthiness. Overall, the framework enables safe, real-time clinical question answering by integrating LLM reasoning with verified medical sources, with future validation planned on public EHR and guideline datasets.

Keywords Clinical decision support, Large language models, Electronic health records, Retrieval-augmented generation, Medical question answering, Hallucination mitigation

*Correspondence:
Mohammed Al-Farsi
mohammed.alfarsi@gmail.com
[1] Department of Intelligent Healthcare Analytics, Sultan Qaboos University, Muscat, Oman
[2] Department of Clinical AI Engineering, German University of Technology in Oman, Muscat, Oman

Introduction

Clinicians face substantial information needs during patient care, ranging from diagnostic questions about symptom complexes to therapeutic decisions about medication selection and prognostic inquiries about expected disease trajectories. Current solutions such as UpToDate, PubMed, and specialty society guidelines provide authoritative but non-personalized information, requiring clinicians to mentally integrate generic recommendations with specific patient characteristics from EHRs [1, 2]. This manual integration process is time-consuming and error-prone, particularly in high-acuity settings where rapid decision-making is essential.

Large language models have demonstrated remarkable performance on medical question answering benchmarks such as MedQA and PubMedQA, sometimes exceeding human expert accuracy on standardized examinations [3, 4]. However, standalone LLMs exhibit critical limitations for clinical deployment: they hallucinate factual content, possess outdated knowledge due to training data cutoffs, and cannot access institutional EHR data containing patient-specific laboratory values, medications, and clinical history [5, 6]. These limitations render ungrounded LLMs unsuitable for real-time clinical decision support where patient safety depends on accurate, current, personalized information.

Retrieval-augmented generation addresses both limitations by retrieving relevant contextual information from external knowledge sources before generating an answer [7, 8]. In the clinical domain, RAG can retrieve from two complementary sources: the patient's own EHR data providing personalized context, and clinical guidelines providing authoritative evidence-based recommendations [9, 10]. This dual retrieval approach grounds LLM generation in both patient-specific and population-level evidence, substantially reducing hallucination risk while maintaining the generative flexibility of LLMs.

This paper presents a conceptual framework for real-time clinical question answering that integrates EHR retrieval, clinical guideline retrieval, and LLM generation into a unified RAG architecture. The framework is designed for emerging AI applications without experimental implementation, focusing instead on architectural principles, component specifications, and evaluation strategies.

Background

Clinical information needs

Clinicians generate numerous questions during patient care, including diagnostic inquiries (e.g., "What is the differential diagnosis for this presentation?"), therapeutic questions (e.g., "What is the first-line antibiotic for community-acquired pneumonia in this patient with penicillin allergy?"), and prognostic queries (e.g., "What is the expected recovery time following this surgical procedure given the patient's comorbidities?") [11]. Studies of clinical workflow demonstrate that unanswered questions are common, with clinicians often unable to find answers during patient encounters due to time constraints and inefficient search processes [1]. Current workflows typically require manual searching of multiple systems—EHR for patient data, UpToDate for guidelines, and PubMed for primary literature—creating cognitive burden and delaying decision-making.

Large language models in medicine

LLMs have shown impressive capabilities on medical question answering benchmarks, with models such as GPT-4 achieving passing scores on the United States Medical Licensing Examination and demonstrating proficiency on biomedical research questions [3, 4, 12]. These models can explain medical concepts, generate differential diagnoses, and summarize clinical information with fluency that mimics human expert reasoning [13, 14]. However, critical limitations persist: LLMs hallucinate plausible but incorrect information, their knowledge is frozen at training time and becomes outdated, and they have no inherent access to local EHR data containing the specific patient context necessary for personalized medicine [5, 6]. These limitations are not merely technical nuisances but fundamental safety hazards for clinical deployment [15, 16].

Retrieval-augmented generation

RAG addresses the grounding problem by retrieving relevant documents from an external knowledge corpus before generating an answer, then conditioning the LLM on both the retrieved context and the user's question [7, 8]. The retrieval component typically uses dense passage retrieval, where a bi-encoder model maps queries and documents into a shared embedding space, enabling efficient approximate nearest neighbor search over millions of documents [17]. Unlike standalone LLMs that rely solely on parametric knowledge, RAG systems can access up-to-date information from external sources and provide citations traceable to retrieved documents, substantially reducing hallucination rates in conversational and structured output tasks [18, 19].

Clinical guidelines as knowledge base

Clinical practice guidelines represent authoritative, evidence-based recommendations for patient care, developed by specialty societies such as the American College of Cardiology, American Society of Clinical Oncology, and Infectious Diseases Society of America [20]. These guidelines are available in both structured formats (e.g., decision trees, algorithmic pathways) and unstructured formats (e.g., narrative text, PDF documents) [21, 22]. Commercial products such as UpToDate and DynaMed provide curated, updated guideline summaries, while primary guidelines are published in medical journals and society websites [23]. For RAG systems, guidelines offer an ideal knowledge source because they are authoritative, regularly updated, and explicitly designed to answer clinical questions with graded levels of evidence and recommendation strength.

Framework Overview

High-level architecture

The proposed framework processes a clinician's natural language question through a five-stage pipeline: first, the question is encoded for retrieval; second, the EHR retriever fetches relevant patient data from structured and unstructured records; third, the guideline retriever fetches relevant evidence-based recommendations; fourth, the LLM generator produces an answer conditioned on both retrieved contexts; and fifth, a safety filter verifies grounding and filters harmful content before presenting the answer to the clinician [7, 8, 24]. This architecture ensures that every generated statement can be traced to either the patient's EHR or an authoritative guideline, enabling transparency and clinical verification.

Figure 1 illustrates the full dual-source RAG architecture through which clinician questions are transformed into traceable answers by sequential query interpretation, parallel EHR and guideline retrieval, grounded generation, and mandatory safety verification.

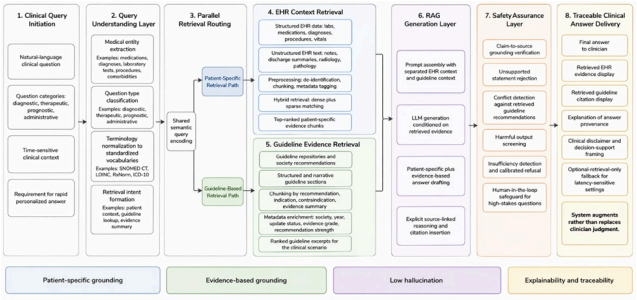

Figure 1. Conceptual Architecture of Dual-Source Retrieval-Augmented Generation for Real-Time Clinical Question Answering

Core assumptions

The framework assumes that healthcare institutions have access to digitized EHR data in structured formats (medications, laboratory results, diagnoses, procedures, vital signs) and unstructured formats (clinical notes, discharge summaries, radiology reports) that can be indexed for retrieval [25]. It further assumes that clinical guidelines are available in machine-readable formats or can be converted from PDFs through optical character recognition and structuring pipelines. Real-time processing assumes acceptable latency thresholds of less than five seconds for non-urgent questions, recognizing that some complex retrievals may require additional time without compromising clinical utility.

Design principles

Four design principles guide the framework: patient-specific grounding ensuring that answers incorporate individual patient data; evidence-based grounding ensuring that recommendations align with authoritative guidelines; low hallucination achieved through retrieval verification and citation requirements; and explainability enabling clinicians to inspect retrieved sources that support each answer component [8]. These principles prioritize safety and transparency over maximal generative flexibility, recognizing that clinical applications demand verifiable outputs rather than creative or novel responses [15, 16].

Table 1 clarifies the distinct functional roles, safety contributions, and failure consequences of each architectural component in the proposed dual-source RAG framework.

Table 1. Functional Differentiation of Core Components in Dual-Source RAG for Clinical Question Answering

Component	Primary Function	Principal Input	Principa Output
Query understanding layer	Interprets clinician intent and extracts medically relevant entities	Natural-language clinical question	Structure retrieva query wit normalize concept and questio type
EHR retrieval module	Identifies patient-specific structured and unstructured evidence	Current patient record, historical EHR documents, normalized query	Ranked patient-specific evidenc chunks
Guideline retrieval module	Identifies authoritative recommendations relevant to the case	Guideline corpus, metadata, normalized query	Ranked guidelin excerpts with provenan and

			evidence grade
Prompt construction layer	Separates and organizes retrieved evidence for generation	EHR evidence plus guideline evidence plus clinician question	Structure prompt w source delineatio
LLM generation module	Synthesizes a fluent answer using retrieved context	Structured grounded prompt	Draft clinical answer with citations and reasonin trace
Grounding verification layer	Tests whether answer claims are supported by retrieved sources	Draft answer plus retrieved evidence	Verified rejected answer segment
Safety filter layer	Screens for harmful or contradictory outputs	Verified answer plus clinical risk signals	Safe answer, warning, refusal
Clinician-facing interface	Presents answer and supporting evidence transparently	Safe verified answer plus citations	Traceabl decision support output

EHR Indexing and Retrieval

Unstructured EHR data

Unstructured clinical text represents a rich source of patient information, including history of present illness, review of systems, physical examination findings, assessment and plan from progress notes, as well as radiology and pathology reports [25]. For retrieval, these documents require preprocessing including de-identification to remove protected health information, sentence or paragraph chunking for granular retrieval, and metadata tagging with document type, date, and author role. The dense retriever encodes these chunks into embedding vectors that capture semantic meaning, enabling retrieval of clinically relevant passages even when exact keyword matches are absent [17].

Structured EHR data

Structured EHR data—medications with dosages and administration routes, laboratory results with reference ranges and flags for abnormal values, diagnoses with standardized codes, procedures with dates, and vital signs with trends—provides quantifiable patient information essential for clinical reasoning [25]. For retrieval, structured data can be linearized into natural language templates (e.g., "Patient's most recent hemoglobin A1c was 7.2% on 2024-01-15") and indexed alongside unstructured text, or encoded separately using specialized retrievers designed for tabular data. Laboratory trends over time, medication changes, and diagnostic trajectories are particularly valuable for answering questions about disease progression and treatment response.

Hybrid retrieval

Medical text presents unique retrieval challenges because clinical concepts may be expressed using different terminology (e.g., "myocardial infarction" versus "heart attack") while keyword matching remains important for specific terms such as medication names or laboratory codes [17]. Hybrid retrieval combining dense semantic retrieval (capturing conceptual similarity) with sparse keyword retrieval (BM25 capturing exact term matches) typically outperforms either approach alone in biomedical domains [7]. A reranking stage using a cross-encoder model can further improve relevance by directly computing query-document relevance scores for top candidates, albeit at higher computational cost suitable for re-ranking rather than initial retrieval [17].

Clinical Guidelines Integration

Guideline chunking and indexing

Clinical guidelines vary substantially in length from brief recommendations to comprehensive documents exceeding 100 pages, requiring thoughtful chunking strategies that preserve semantic coherence while enabling granular retrieval [22, 23]. Sentence-level chunks provide maximum

granularity but may lose surrounding context, while section-level chunks preserve the logical structure of recommendations, including indications, contraindications, evidence summaries, and strength of recommendations. Metadata enrichment is essential, tagging each chunk with source guideline title, issuing society, publication date, update status, evidence grade (e.g., Level A, Class I), and recommendation strength to enable filtering and prioritization during retrieval [21, 22].

Retrieval from guidelines

Given a clinician's question about a specific patient, the guideline retriever identifies the most relevant recommendation excerpts by encoding the question and guideline chunks into a shared embedding space [17]. Retrieved excerpts can be prioritized by recency (favoring updated guidelines over older versions) and evidence grade (favoring high-quality evidence such as randomized controlled trials over expert opinion) [21, 22]. When multiple guidelines address the same clinical scenario—for example, conflicting recommendations from different specialty societies—the framework can retrieve and present all relevant excerpts, enabling clinicians to adjudicate discrepancies rather than forcing a single recommendation [23].

RAG Architecture

Retriever component

The dense retriever employs a bi-encoder architecture that independently encodes queries and documents into fixed-dimensional embedding vectors, enabling efficient approximate nearest neighbor search over pre-computed document embeddings [7, 17]. For medical domain adaptation, the retriever can be fine-tuned on clinical question-answer pairs or medical information retrieval datasets such as emrQA or PubMedQA to improve semantic matching of medical terminology [2, 3]. An embedding dimension of 768 to 1024 balances retrieval quality with storage efficiency, supporting indexing of millions of EHR chunks and guideline excerpts while maintaining sub-second retrieval latency [17].

Generator component

The LLM generator conditions on a prompt that concatenates retrieved EHR data, retrieved guideline excerpts, and the clinician's original question, producing a fluent answer that synthesizes patient-specific and evidence-based information [7, 8]. Model sizes between 7 billion and 70 billion parameters offer a practical trade-off between reasoning capability and inference latency, with smaller models suitable for local deployment behind institutional firewalls and larger models potentially requiring cloud infrastructure [9, 10]. Instruction tuning on medical question answering datasets teaches the model to follow clinical prompt formats, cite retrieved sources, and appropriately refuse to answer when retrieved context is insufficient.

Prompt design

The prompt structure explicitly separates system instructions, retrieved context, and user question, with the system instruction defining the model's role as a clinical decision support tool that must ground all answers in provided sources [8]. Retrieved EHR data appears first, formatted with clear section headers and date stamps, followed by retrieved guideline excerpts with source citations including society name, publication year, and evidence grade [22, 23]. The user question then appears, and the model is instructed to generate an answer that explicitly references specific retrieved passages, cites sources using numeric references, and states when retrieved information is insufficient to answer confidently [18, 19].

Real-Time Query Processing

Query understanding

Before retrieval, the system analyzes the clinician's natural language question to extract medical entities (medications, diagnoses, laboratory tests, procedures) and classify question type into diagnostic, therapeutic, prognostic, or administrative categories [2]. Entity extraction identifies patient-specific references such as "this patient's creatinine" that require EHR retrieval focused on the current encounter, while question type classification determines which guideline sections are most relevant—therapeutic questions prioritize treatment recommendation sections whereas diagnostic questions prioritize differential diagnosis algorithms [11]. This preprocessing step also normalizes medical terminology to standardized vocabularies (RxNorm for medications, LOINC for laboratory tests, SNOMED CT for diagnoses) to improve retrieval matching.

Latency optimization

Real-time clinical use requires end-to-end latency under five seconds for non-urgent questions, achieved through a combination of pre-computation and inference optimization [7, 17]. EHR and guideline embeddings are pre-computed offline and stored in vector indexes, reducing retrieval to a fast nearest neighbor search rather than expensive on-the-fly encoding. LLM inference benefits from quantization (reducing model weights from 16-bit to 8-bit or 4-bit precision) and speculative decoding (using a smaller draft model to generate candidate tokens that the larger model verifies in parallel) [9, 10]. For latency-sensitive settings such as emergency departments or operating rooms, the framework can optionally return retrieval results alone without generation, providing clinicians with relevant source excerpts within one second.

Safety and Hallucination Mitigation

Grounding verification

After generation, the framework verifies that each factual claim in the answer can be traced to at least one retrieved source, rejecting answers that contain unsupported statements [18, 19, 24]. This verification can be implemented as a second LLM call that checks answer-sentence against retrieved passages, or as a rule-based system requiring explicit citation markers in the generated output. When retrieved context is insufficient to answer a question completely, the model is instructed to state its limitations explicitly (e.g., "The retrieved EHR does not contain recent laboratory values for this patient") rather than hallucinating missing information [18, 24]. Questions that fall entirely outside the scope of retrieved guidelines or EHR data trigger a refusal response directing clinicians to alternative resources such as specialist consultation or primary literature.

Safety filtering

Post-generation guardrails apply content safety filters to detect and block harmful outputs including treatment recommendations that contradict retrieved guidelines, dangerous medication combinations, or inappropriate diagnostic assertions [15, 16]. Medical disclaimers are appended automatically to all answers, stating that the system provides decision support rather than autonomous recommendations and that final clinical decisions rest with the treating physician [13, 14]. For high-stakes questions identified by question type classification (e.g., "Should I administer thrombolytics?" in suspected stroke), the framework requires explicit clinician approval before displaying the generated answer, creating a human-in-the-loop safeguard that preserves clinical accountability [11].

Evaluation Strategy

Retrieval metrics

Retrieval performance is evaluated using held-out clinical questions drawn from real clinician queries, with metrics including recall@k (proportion of questions for which at least one relevant source appears in the top-k retrieved results), mean reciprocal rank (inverse rank of the first relevant result), and normalized discounted cumulative gain (accounting for graded relevance of multiple retrieved sources) [3, 17]. For EHR retrieval, relevance is defined by whether retrieved patient data actually answers the clinical question (e.g., retrieving the most recent creatinine value for a question about kidney function), while guideline retrieval relevance requires that excerpts contain the correct recommendation for the clinical scenario [22, 23]. Target thresholds include recall@5 above 0.90 for both retrieval sources, ensuring that generation has access to necessary context.

Generation metrics

Answer quality requires evaluation along multiple dimensions: factual accuracy assessed by clinician review comparing generated answers against gold-standard responses derived from EHR data and guidelines, hallucination rate measured as the proportion of answer statements unsupported by retrieved sources, answer completeness measured as the proportion of gold-standard information present in the generated answer, and citation correctness measured as whether cited sources actually support the claims they accompany [5, 6, 8]. Clinician evaluation panels comprising attending physicians from relevant specialties rate answers on Likert scales for clinical appropriateness, safety, and actionability [13, 14]. A hallucination rate below 5% and factual accuracy above 95% represent minimum acceptable thresholds for clinical deployment.

Clinical validation

Real-world pilot studies with practicing clinicians assess the framework's practical utility using time-motion analysis (time saved per question compared to manual EHR and guideline searching), answer correctness verified against a reference standard created by expert panels, clinician trust measured through validated questionnaires, and usability assessed using system usability scale instruments [1, 11]. Pilot deployment on public EHR datasets such as MIMIC-III or MIMIC-IV and public guideline repositories would enable rigorous evaluation without institutional review board constraints during development [9, 10, 25]. Successful pilots would demonstrate statistically significant reductions in answer time, non-inferiority or superiority in answer correctness compared to current workflows, and clinician trust scores exceeding predefined thresholds.

Table 2 translates the proposed architecture into an evaluation logic by linking retrieval performance, generation quality, safety enforcement, and real-world clinical usefulness.

Table 2. Evaluation Matrix Linking Retrieval Quality, Generation Reliability, Safety Performance, and Clinical Utility in Real-Time Clinical RAG

Evaluation Domain	What Should Be Measured	Representative Metric or Assessment Logic	V
EHR retrieval performance	Ability to recover patient-specific evidence needed to answer the question	Recall@k, mean reciprocal rank, relevance of retrieved laboratory values, medications, diagnoses, notes	T s e
Guideline retrieval performance	Ability to recover authoritative recommendations aligned with the scenario	Recall@k, nDCG, recency-weighted relevance, evidence-grade-aware ranking quality	E re
Cross-source complementarity	Degree to which retrieved EHR evidence and guideline evidence jointly support answer construction	Proportion of questions for which both patient-specific and recommendation-specific evidence are retrieved	T (
Generative factual accuracy	Accuracy of synthesized answer relative to retrieved evidence and expert reference standard	Expert review, claim matching, correctness scoring	A m
Hallucination control	Frequency of unsupported claims in model output	Unsupported statement rate, claim-to-source verification failure rate	T (
Citation fidelity	Whether cited sources actually support the adjacent answer claims	Source-claim concordance review	r a
Answer completeness	Degree to which clinically necessary information is included without critical omissions	Comparison with expert-generated reference answers	P c o
Latency and workflow fit	Time required for end-to-end retrieval, generation, verification, and display	Total response time, module-specific latency, fallback performance	r tl

Safety adjudication performance	Ability to detect harmful, contradictory, or insufficiently grounded responses	Safety filter precision and recall, rate of appropriate refusals, escalation frequency for high-stakes questions	a
Human factors and clinical utility	Practical effect on clinician trust, usability, and time burden	Time-motion analysis, usability scales, trust questionnaires, clinician preference comparison	C u

Conclusion

This paper has presented a conceptual framework for retrieval-augmented generation that integrates dual-source retrieval from electronic health records and clinical guidelines to enable real-time, patient-specific clinical question answering. The architecture addresses the fundamental limitations of standalone large language models—hallucination, lack of patient-specific grounding, and inability to access local EHR data—by conditioning generation on retrieved context from both the individual patient's medical record and authoritative evidence-based guidelines.

The framework's key advantages include patient-specific grounding that ensures answers reflect the individual's actual laboratory values, medications, and clinical history; evidence-based grounding that aligns recommendations with current guidelines from specialty societies; low hallucination achieved through retrieval verification and citation requirements; and explainability enabled by traceable source citations. By retrieving rather than memorizing clinical knowledge, the system maintains up-to-date information without retraining and provides transparency that is essential for clinician trust and patient safety.

Several limitations require acknowledgment. The framework's performance depends critically on retrieval quality—if the retriever fails to fetch relevant EHR data or guideline excerpts, generation quality degrades regardless of LLM capability. LLM inference latency, even with optimization, may exceed acceptable thresholds for some high-acuity settings, though retrieval-only fallback modes can mitigate this concern. Most importantly, the framework requires rigorous clinical validation before deployment, including prospective studies measuring impact on clinician workflow, decision quality, and patient outcomes.

We call for implementation of this framework on public EHR datasets such as MIMIC-IV and public clinical guideline repositories, enabling the research community to evaluate retrieval and generation performance without institutional barriers. Subsequent validation in real-world clinical pilots with attending physicians would assess practical utility and safety. If successful, this RAG-based approach could transform clinical decision support from generic information retrieval to personalized, real-time, evidence-based reasoning that augments rather than replaces clinician judgment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 May 2024 Revised: 17 Jun 2024 Accepted: 17 Jul 2024
Published online: 20 January 2025

References

Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. In: Proc Conf Empir Methods Nat Lang Process Int Joint Conf Nat Lang Process (EMNLP-IJCNLP). 2019;2019:2567-77.

Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel). 2021;11(14):6421. https://doi.org/10.3390/app11146421.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194. https://doi.org/10.1038/s41746-022-00742-2.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. https://doi.org/10.1038/s41586-023-06291-2.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-74.

Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proc Conf Empir Methods Nat Lang Process (EMNLP). 2020;2020:6769-81.

Khattab O, Zaharia M. ColBERT: efficient and effective passage search via contextualized late interaction over BERT. In: Proc Int ACM SIGIR Conf Res Dev Inf Retr. 2020;2020:39-48.

Pampari A, Raghavan P, Liang J, Peng J. emrQA: a large corpus for question answering on electronic medical records. In: Proc Conf Empir Methods Nat Lang Process. 2018;2018:2357-68.

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. https://doi.org/10.2196/45312.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8.

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1.

Luo MJ, Pang J, Bi S, Lai Y, Zhao J, Shang Y, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. 2024;142(9):798-805. https://doi.org/10.1001/jamaophthalmol.2024.2303.

Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15.

Zhan Z, Zhou S, Li M, Zhang R. RAMIE: retrieval-augmented multi-task information extraction with large language models on dietary supplements. J Am Med Inform Assoc. 2025;32(3):545-54.

Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc. 2025;32(6):1032-9.

Alkhalaf M, Yu P, Yin M, Deng C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. J Biomed

Inform. 2024;156:104662. https://doi.org/10.1016/j.jbi.2024.104662.

Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res. 2024;26:e58041. https://doi.org/10.2196/58041.

Jeong M, Sohn J, Sung M, Kang J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics. 2024;40(Suppl 1):i119-29.

Ong CS, Obey NT, Zheng Y, Cohan A, Schneider EB. SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. NPJ Digit Med. 2024;7(1):364. https://doi.org/10.1038/s41746-024-01366-z.

Lopez I, Swaminathan A, Vedula K, Narayanan S, Nateghi Haredasht F, Ma SP, et al. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. 2025;8(1):45. https://doi.org/10.1038/s41746-025-01459-2.

Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia AT, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. https://doi.org/10.1038/s41746-025-01600-1.

Wada A, Tanaka Y, Nishizawa M, Yamamoto A, Akashi T, Hagiwara A, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Digit Med. 2025;8(1):395. https://doi.org/10.1038/s41746-025-01782-8.

Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. In: Findings Assoc Comput Linguist EMNLP 2021. 2021;2021:3784-803.

Ayala O, Bechard P. Reducing hallucination in structured outputs via retrieval-augmented generation. In: Proc Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Ind Track. 2024;2024:228-38.