Review

Evaluation Methods for Inference-Time Retrieval-Augmented and Graph Retrieval-Augmented Large Language Models in Health Care: Scoping Review

Yuhan Zhao*, PhD; Yiqun Miao*, PhD; Rongrong Guo, PhD; Yuan Luo, PhD; Huiying Wang, PhD; Ying Wu, PhD

School of Nursing, Capital Medical University, Beijing, China
*these authors contributed equally

Corresponding Author:

Ying Wu, PhD
School of Nursing, Capital Medical University
No. 10 Xitoutiao, Youanmenwai, Fengtai District
Beijing 100069
China
Phone: 86 13910789837
Email: helenywu@vip.163.com

Abstract

Background: Inference-time retrieval augmentation is increasingly used to improve the traceability and verifiability of large language model (LLM) applications in health care. Evaluation practices for text-based retrieval-augmented generation (RAG) and graph-structured RAG (GraphRAG) systems remain heterogeneous, which limits comparison across studies and complicates judgments about clinical readiness.

Objective: This review mapped evaluation methods for inference-time retrieval-augmented and graph-structured retrieval-augmented LLM systems in health care and characterized how evaluation constructs are defined, operationalized, and reported across system layers and evaluation-setting categories.

Methods: We conducted a scoping review in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews), with search reporting informed by PRISMA-S (PRISMA literature search extension). Searches were conducted through May 14, 2026, in PubMed (MEDLINE), Web of Science Core Collection, IEEE Xplore, ACM Digital Library, arXiv, and medRxiv, with backward and forward citation tracking of included studies. Eligible records described health care–relevant LLM systems using inference-time RAG and reported at least 1 evaluation component. Data were charted on study characteristics, system design, retrieval-layer evaluation, evidence linkage, safety-related and GraphRAG-specific evaluation, and selected reporting and governance characteristics. We also constructed an evidence-and-gap map cross-classifying evaluation-setting categories with key evaluation domains.

Results: A total of 157 studies met the inclusion criteria. Clinical question answering was the most frequently represented application (89/157, 56.7%), followed by clinical decision support (70/157, 44.6%). Most evaluations were conducted in offline-only settings (140/157, 89.2%), whereas 17/157 (10.8%) studies reported workflow-facing, prospective, or deployment-level evaluation. Independent retrieval-layer evaluation was reported in 47/157 (29.9%) studies. Grounding and faithfulness evaluation was reported in 41/157 (26.1%) studies, and fine-grained evidence verification was reported in 22/157 (14%) studies. Human evaluation was reported in 94/157 (59.9%) studies, but interrater reliability was reported in 26/94 (27.7%) studies. LLM-as-judge evaluation was reported in 41/157 (26.1%) studies, with bias-control measures reported in 15/41 (36.6%) studies. Formal safety-related evaluation was reported in 45/157 (28.7%) studies. Among 27 (17.2%) GraphRAG studies, intermediate-artifact evaluation was reported in 11/27 (40.7%) studies, and graph construction evaluation was reported in 6/27 (22.2%) studies. The evidence-and-gap map showed limited coverage of fine-grained verification, contradiction handling, safety evaluation, LLM-as-judge safeguards, GraphRAG construction evaluation, and GraphRAG intermediate-artifact evaluation in workflow-facing, prospective, or deployment-level settings.

Conclusions: Evaluation of health care RAG and GraphRAG systems has expanded rapidly, yet reporting and operational definitions remain inconsistent across evaluation layers. Current evidence remains concentrated in offline evaluation, with limited workflow-facing, prospective, or deployment-level assessment of retrieval quality, fine-grained evidence linkage, safety, LLM-as-judge safeguards, GraphRAG construction quality, and GraphRAG intermediate artifacts. This review maps

these gaps across evaluation-setting categories and translates them into synthesis-informed evaluation considerations. These findings suggest that future evaluation may need to move beyond end-to-end benchmark performance toward more transparent, layer-specific, safety-oriented, and clinically contextualized assessment before workflow-facing implementation.

Trial Registration: OSF Registries mtf5x; https://osf.io/mtf5x/overview

J Med Internet Res 2026;28:e90046; doi: 10.2196/90046

Keywords: large language models; natural language processing; artificial intelligence; retrieval-augmented generation; GraphRAG; information storage and retrieval; scoping review; evaluation studies as topic; hallucination; clinical decision support systems

Introduction

Rationale

Large language models (LLMs) are increasingly explored in health care for tasks such as clinical decision support, patient education, documentation, administrative workflows, and broader biomedical use cases [1-6]. However, study designs, evaluation end points, and reporting practices remain heterogeneous, and deployment in high-stakes medical settings is constrained by the tendency of these models to produce factually incorrect, internally inconsistent, or insufficiently supported statements [7,8]. Such reliability limitations raise patient safety concerns and can undermine clinician trust in automated systems [9,10]. Improving factual accuracy and enabling verifiable outputs have therefore become central objectives for medical artificial intelligence (AI) research [11]. Accordingly, evaluation is not only a measure of technical performance but also a prerequisite for judging whether health care LLM systems are sufficiently transparent, safe, and interpretable before use in progressively more clinical or workflow-facing settings.

To mitigate factual unreliability, researchers have increasingly adopted inference-time retrieval-augmented generation (RAG) [12]. This paradigm retrieves evidence from external knowledge sources such as clinical guidelines, biomedical literature, institutional protocols, and electronic health records (EHRs) during the generation process. By conditioning generation on retrieved context, these systems aim to improve accuracy and support traceability of outputs to source evidence [13]. The approach includes text-based RAG and emerging graph-structured RAG (GraphRAG) [14]. Many retrieval-augmented systems retrieve text using dense vector similarity or hybrid sparse-dense retrieval, whereas graph-based approaches use graph structure to condition inference-time evidence retrieval or organization for generation [15,16]. Graph-based methods can also introduce unique intermediate artifacts such as retrieved paths, retrieved subgraphs, and graph community summaries, which create additional evaluation targets beyond end-to-end task performance. However, retrieval augmentation does not by itself ensure that retrieved evidence is relevant, that generated claims are faithfully supported by that evidence, that citation and source attributions are correct, or that outputs are clinically safe. RAG and GraphRAG systems therefore require evaluation methods that distinguish retrieval quality, evidence linkage, end-to-end output quality, safety-related behavior, and readiness for more clinically realistic settings.

Despite the rapid expansion of health care retrieval-augmented architectures, particularly since 2024, evaluation methodologies remain fragmented. While systematic and scoping reviews have mapped general LLM applications in clinical medicine, patient education, and broader health care settings [2-6], other reviews and guidance papers have focused more directly on testing, evaluation, and reporting of health care LLM applications [1,6,11,17]. General RAG and GraphRAG reviews have also summarized retrieval-augmented architectures, retriever-generator integration, robustness issues, graph-based indexing, graph-guided retrieval, graph-enhanced generation, and emerging evaluation frameworks [12,15,16,18-20]. These syntheses are important, but they have not primarily been designed to provide a layer-specific synthesis of health care evaluation methods for inference-time RAG and GraphRAG systems, including verification granularity, independent retrieval-layer evaluation, formal safety-related evaluation, and evaluation-setting context. Many studies emphasize end-to-end performance metrics, which can obscure the distinct contributions and failure modes of retrieval and generation components [18,19].

RAG evaluation literature has further highlighted that retrieval-augmented systems pose distinctive evaluation challenges because their behavior depends on both retrieval and generation components as well as on dynamic external knowledge sources [18,19]. A related review of medical RAG literature likewise suggests that research in this area has concentrated on technical implementations and clinical applications, whereas evaluation commonly relies on automated metrics or broad human judgments, with less explicit attention to bias and safety [21]. In addition, it is often unclear to what extent evaluation protocols incorporate clinical validity, safety assessment, and testing in settings that approximate real clinical use [17,22,23]. Evaluation methods for inference-time RAG and GraphRAG systems therefore warrant dedicated synthesis, particularly to clarify how layer-specific constructs are operationalized, how evaluation coverage varies across evaluation-setting categories and key evaluation domains, and where methodological gaps remain within health care domains. To our knowledge, no previous review has specifically mapped how evaluation methods for health care inference-time RAG and GraphRAG systems are operationalized across retrieval-layer evaluation, grounding and faithfulness evaluation, citation and source correctness evaluation, formal safety-related evaluation, human evaluation, automated metrics, LLM-as-judge evaluation, Graph-

RAG construction and intermediate-artifact evaluation, and evaluation-setting categories.

Objectives

Given the heterogeneity in task definitions, data sources, and evaluation end points, we conducted a scoping review to map the evidence and characterize evaluation practices rather than to estimate pooled effects. The primary objective of this review was to systematically map evaluation methods used for inference-time RAG and GraphRAG systems in health care. Specifically, this review aimed to characterize evaluation constructs, measurement approaches, and study designs across system layers, including retrieval-layer evaluation, grounding and faithfulness evaluation, citation and source correctness evaluation, verification granularity, end-to-end task outcomes, human evaluation, automated metrics, LLM-as-judge evaluation and related safeguards, formal safety-related evaluation, GraphRAG construction evaluation, and GraphRAG-specific intermediate-artifact evaluation. By synthesizing these practices into a layer-specific taxonomy and mapping evaluation coverage across evaluation-setting categories and key evaluation domains, this review aimed to identify methodological gaps, differentiate the evaluation needs of text-based RAG and GraphRAG systems, and inform more transparent, setting-appropriate, and safety-oriented evaluation and reporting in future research.

Methods

Protocol and Registration

This scoping review was designed to systematically map and characterize evaluation methods used in health care LLM systems using inference-time RAG, with specific attention to both text-based RAG and GraphRAG approaches. The review was conducted in accordance with methodological guidance from the Joanna Briggs Institute for scoping reviews and reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist (Checklist 1) [24,25]. A protocol specifying the research questions, eligibility criteria, information sources, search strategy structure, screening procedures, data charting items, and coding framework was developed a priori and registered on the Open Science Framework (OSF) through OSF Registries before study selection and data charting were initiated (registration ID: MTF5X).

The objective was to identify and synthesize how evaluation constructs are defined and operationalized across system layers, to summarize study designs and evaluation modalities, to map evaluation coverage across descriptive evaluation-setting categories, and to characterize reporting practices relevant to the interpretation of evaluation methods, including governance indicators related to real patient data, deidentification, and ethics approval, exemption, or waiver. Specifically, this review sought to map how evaluation was performed at the retrieval-layer, grounding and faithfulness, evidence-verification, safety-related, GraphRAG-specific, and end-to-end system levels, how evaluation coverage varied across evaluation-setting categories, and how these evaluations were reported across heterogeneous health care settings. Given the expected heterogeneity in evaluation targets, end points, and study settings, a scoping review approach was selected to support comprehensive mapping and structured evidence synthesis rather than quantitative effect estimation.

This scoping review addressed 5 research questions aligned with the structure of the Results section.

First, what health care tasks and evaluation-setting categories are represented in studies evaluating inference-time RAG and GraphRAG LLM systems?

Second, what system design characteristics relevant to evaluation are reported, including knowledge source types, retrieval approaches, reranking stages, GraphRAG-related system features, and generator or evaluator model choices?

Third, how is retrieval-layer evaluation operationalized, including whether independent component-level evaluation is performed, which retrieval metrics are reported, how relevance labels or reference evidence are constructed, and how retrieval unit and granularity are specified?

Fourth, how are grounding and faithfulness outcomes evaluated, including how these constructs are operationalized, how outputs are linked to retrieved evidence, what units of analysis are used for verification, and whether citation and source correctness evaluation or conflict and contradiction handling is assessed?

Finally, what broader evaluation modalities and reporting practices are used, including human evaluation procedures, automated end point selection, LLM-as-judge evaluation procedures, formal safety-related evaluation, GraphRAG construction evaluation, GraphRAG intermediate-artifact evaluation, and selected governance-related elements?

Ethical Considerations

Ethical approval was not required for this scoping review because it synthesized data from publicly available studies and did not involve human participants.

Eligibility Criteria

Eligibility criteria were defined a priori using the Population, Concept, and Context (PCC) framework recommended for scoping reviews [25]. In this review, the population corresponded to implemented health care–relevant LLM systems, the concept to inference-time RAG and its evaluation, and the context to health care application settings. Records were eligible if they described a health care–relevant system in which an LLM produced generative outputs and implemented RAG by retrieving external evidence during inference and incorporating retrieved material into generation, including text-based retrieval as well as GraphRAG approaches in which graph structure was used for inference-time evidence retrieval, organization, or assembly. Records were required to report at least 1 evaluation component (eg, retrieval performance, grounding and faithfulness, citation and source correctness, end-to-end task performance, human

evaluation, formal safety-related evaluation, or GraphRAG-specific evaluation).

We excluded studies describing training-only knowledge injection without inference-time retrieval conditioning, retrieval systems without generative outputs, studies without empirical evaluation of an implemented system, narrative reviews, conference abstracts without accessible full text, and records for which full text could not be obtained. When multiple records described the same underlying study, a single record was retained for synthesis to avoid double counting. We preferentially retained the peer-reviewed version when it adequately described the evaluation methods; otherwise, the record providing the most complete description of the study design, system implementation, and evaluation procedures was retained as the primary synthesis record, with companion reports consulted as needed for clarification only.

Information Sources

We searched PubMed (MEDLINE), Web of Science Core Collection, IEEE Xplore, and the ACM Digital Library. Searches were conducted on May 14, 2026, and were limited to records published or posted from January 1, 2024 to May 14, 2026. Searches were limited to English-language records. This period was selected to focus on contemporary evaluation practices from 2024 onward, alongside the broader uptake of frontier LLMs and specialized health care RAG frameworks. This window was intended to reflect recent evaluation paradigms for inference-time retrieval architectures in clinical domains rather than foundational natural language processing (NLP) tasks that predated their widespread health care use. To capture emerging work disseminated ahead of journal publication, we additionally searched arXiv and medRxiv using equivalent concept blocks adapted to platform-specific syntax and the same search cutoff. We also performed backward and forward citation tracking for all included studies using the same eligibility criteria to identify additional eligible records [26].

Search Strategy

The search strategy was developed iteratively by the review team and combined controlled vocabulary and free-text terms for (1) LLMs; (2) inference-time RAG, including GraphRAG approaches; (3) health care context; and (4) evaluation-related concepts. Controlled vocabulary was used where supported by the database, and syntax, fields, and limits were adapted to each platform. To align the search strategy with the review objective of synthesizing evaluation practices, the database queries included evaluation-related terms, including evaluation, benchmarking, metrics, grounding and faithfulness, hallucination, safety-related evaluation, and citation and source correctness, to improve identification of studies that explicitly reported evaluation methods, benchmarking designs, or evidence-verification procedures. Because the review focused specifically on evaluation practices, these terms were included to improve precision for studies reporting explicit evaluation methods; backward and forward citation tracking was used to mitigate the risk of missing eligible studies whose evaluation components were not captured in searchable titles, abstracts, or keywords. Literature searching and search reporting were conducted and documented in accordance with the PRISMA-S (PRISMA literature search extension) [26]; an item-by-item PRISMA-S checklist, full database-specific search strategies, and supplementary search procedures are provided in Checklist 2 and Multimedia Appendix 1. No study registries, structured website handsearching, contact-based supplementary identification, or formal search peer review was undertaken; these PRISMA-S items are reported explicitly in Checklist 2.

Selection of Sources of Evidence

Retrieved records were deduplicated in EndNote (Clarivate) and screened in a dedicated platform. Moreover, 2 reviewers (YZ and YM) independently screened titles and abstracts and then full texts of records classified as potentially eligible or uncertain. Reviewers completed a calibration exercise before full screening. Discrepancies were resolved through a consensus-seeking discussion between the 2 reviewers (YZ and YM); if consensus could not be reached, a third senior reviewer (YW) adjudicated the final inclusion decision. Reasons for exclusion at the full-text stage were recorded and are reported in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Records identified through all information sources and citation tracking were screened using the same eligibility criteria and selection procedures.

Data Charting Process

A standardized data-charting form was developed a priori to operationalize the review questions and support consistent charting of system characteristics, evaluation constructs, operational definitions, descriptive evaluation-setting categories, evaluation domains, and reporting elements. The form was refined iteratively through team discussion and was piloted on a prespecified sample of included studies to calibrate interpretation of fields and coding rules. Following calibration, one reviewer (YZ) charted data from all included studies and a second reviewer (YM) independently verified all entries against the full texts and available supplementary materials. Discrepancies were resolved through a consensus-seeking discussion. If consensus was not reached, a third senior reviewer (YW) adjudicated the final coding to maintain consistency across the review.

Charting was conducted at the study level and allowed multilabel coding when studies reported multiple tasks, knowledge sources, evaluation-setting categories, or evaluation modalities. Coding was based on full-text review and prespecified category definitions rather than title, abstract, or terminology alone. When information relevant to a field was explicitly reported, fields were coded according to the relevant category definitions. When information was not reported, fields were coded as “not reported.” Fields were coded as “not applicable” when the item was structurally irrelevant to the study design or evaluation approach. Ambiguous cases were resolved through full-text review and reviewer consensus rather than retained as a separate coding category. Coding decisions and adjudication notes were documented to preserve an audit trail.

Data Items

Data were charted across five domains: (1) study characteristics and application settings, including application task categories and descriptive evaluation-setting categories; (2) system design characteristics relevant to evaluation, including knowledge source type, retrieval granularity, retrieval approach, reranking, generation model choice, evaluator model choice, and GraphRAG-related system features; (3) retrieval-layer evaluation, including retrieval metrics, relevance labeling procedures, reference evidence construction, and reporting of document-level or passage-level granularity; (4) grounding and faithfulness evaluation, including terminology used, operationalization approach, unit of analysis, fine-grained evidence verification, citation and source correctness, and handling of conflicting or contradictory evidence; and (5) evaluation modalities and reporting-related study characteristics, including human evaluation design and rater characteristics, interrater reliability (IRR) reporting, automated metrics, LLM-as-judge procedures and bias-control measures, formal safety-related evaluation, GraphRAG construction and intermediate-artifact evaluation, and governance indicators when explicitly described.

Operational Definitions and Coding Framework

To support consistent and reproducible synthesis, we applied an explicit coding framework with operational definitions for evaluation layers, constructs, and study attributes. The framework was developed before full data charting, refined during calibration, and then applied uniformly across all included studies. Key operational definitions and primary coding units used in this review are summarized in Table 1.

Table 1. Operational definitions and coding units used in this review.

Construct	Working definition in this review	Primary unit of assessment	Included under this construct	Not counted as this construct
Retrieval-layer evaluation	Explicit assessment of the quality or performance of retrieved evidence independent of final generated outputs.	Retrieved item, ranked set, document, passage, node, path, or subgraph	Retrieval metrics, relevance judgments, structured assessment of retrieval quality, comparison of retrieved evidence sets	End-to-end task performance reported without explicit retrieval-layer assessment
Grounding and faithfulness evaluation	Assessment of whether generated answers, claims, statements, citations, or sources were supported by retrieved evidence or retrieved context, and whether generated outputs remained constrained by that evidence.	Response, answer, claim, statement, citation, source, or sentence	Evidence-support assessment, answer faithfulness assessment, source-supported output checking, claim verification against retrieved evidence, and assessment of unsupported additions relative to retrieved context	Factual correctness assessed only against an external reference standard without explicit linkage to retrieved evidence
Citation and source correctness evaluation	Assessment of whether cited or displayed sources existed, were accurate, and supported the corresponding answer content.	Citation, source, source excerpt, or linked answer segment	Source attribution checks, citation support checks, source existence checks, evidence-to-answer linkage assessment	Citation presence alone without verification that the cited or displayed source supports the answer content
Formal safety-related evaluation	Assessment in which safety, harm, unsafe recommendations, hallucination-related risk, undertriage, suicide risk, harmful content, or comparable safety outcomes were included as formal evaluation dimensions or end points.	Response, recommendation, decision, triage output, risk classification, or system behavior	Safety scores, harmfulness ratings, unsafe recommendation assessment, hallucination risk assessment, undertriage or overtriage harm evaluation, harmful content blocking	General discussion of safety risks without formal evaluation; accuracy or guideline concordance reported without a safety-related end point
Evaluation-setting category	Descriptive classification of evaluation conditions by their relationship to real-world health care use, rather than an ordinal maturity level.	Study or evaluation setting	Offline-only evaluation, simulated vignette or case evaluation, workflow pilot or user study, prospective clinical study, real-world deployment or postdeployment monitoring	General study setting description without enough information to classify the evaluation setting
GraphRAG[a] minimum criterion	Explicit use of graph structure at inference time to retrieve, organize, or assemble evidence that conditions generation.	System-level design feature	Retrieval of nodes, paths, subgraphs, graph-structured evidence assembly, graph-derived summaries, or graph-derived community summaries used at inference time	Knowledge graph use limited to background knowledge representation, training-time enrichment, or architecture description without graph-structured inference-time retrieval or evidence assembly
Graph construction evaluation	Explicit evaluation of graph construction quality or graph content quality.	Graph, node, edge, relation, triple, or graph-derived schema	Assessment of node correctness, edge correctness, relation quality, triple extraction quality, graph completeness, or graph construction accuracy	Reporting graph size, graph architecture, or graph construction workflow without evaluating graph quality

Construct	Working definition in this review	Primary unit of assessment	Included under this construct	Not counted as this construct
GraphRAG intermediate-artifact evaluation	Explicit evaluation of intermediate graph-related artifacts produced or used during graph-structured retrieval-augmented generation.	Node, edge, path, subgraph, graph-derived summary, retrieved graph context, or provenance-linked graph artifact	Assessment of retrieved nodes, retrieved paths, retrieved subgraphs, graph-derived summaries, graph-based context, or provenance-linked graph artifacts	End-to-end output evaluation of a GraphRAG system without assessment of graph-related intermediate artifacts

[a]GraphRAG: graph-structured retrieval-augmented generation.

Evaluation targets were conceptualized as nonmutually exclusive layers reflecting the multicomponent nature of inference-time RAG systems, including retrieval-layer evaluation, grounding and faithfulness evaluation, and end-to-end task evaluation. In addition, studies were coded for cross-cutting evaluation modalities and reporting dimensions, including human evaluation, formal safety-related evaluation, efficiency and implementation-readiness indicators, graph construction evaluation, GraphRAG intermediate-artifact evaluation, and reporting and governance dimensions.

Retrieval-layer evaluation was coded as present only when a study reported an explicit assessment of retrieved evidence quality independent of final generated outputs. This included quantitative retrieval metrics, relevance judgments of retrieved items, or structured evaluation of retrieval performance. Studies that reported only end-to-end task performance or qualitative examples without explicit assessment of retrieved evidence were not coded as reporting retrieval-layer evaluation.

For the purposes of this review, grounding and faithfulness evaluation was defined as assessment of whether generated answers, claims, statements, citations, or sources were supported by retrieved evidence or retrieved context, and whether generated outputs remained constrained by that evidence. Because terminology varied across studies, grounding and faithfulness evaluations were identified based on described verification procedures rather than author-reported labels. Citation and source correctness, claim verification against retrieved evidence, and assessment of unsupported additions relative to retrieved context were included when they explicitly evaluated alignment between generated outputs and retrieved sources. Hallucination-related assessments were included under grounding and faithfulness only when unsupported content was evaluated with reference to retrieved evidence or retrieved context. This operationalization was informed by previous RAG evaluation literature that assesses answer faithfulness and related evidence-linked dimensions [18,20].

For each distinct grounding or evidence-linkage evaluation component, we recorded the most granular verification unit explicitly described, including response- or answer-level, claim- or statement-level, citation- or source-level, and sentence-level verification. Because individual studies could report multiple evaluation components at different verification units, the summary granularity categories were nonmutually exclusive. Fine-grained evidence verification was coded when studies reported claim-, statement-, citation-, source-, or sentence-level verification. When a relevant verification unit was not explicitly described, it was coded as "not reported." Fields were coded as "not applicable" when the item was structurally irrelevant to the study design or evaluation approach. Ambiguous cases were resolved through full-text review and reviewer consensus rather than retained as a separate coding category.

Evaluation-setting categories were coded to describe the relationship between evaluation conditions and real-world health care use [22]. These descriptive, nonmutually exclusive categories included offline-only evaluation, simulated vignette or case evaluation, workflow pilot or user study, prospective clinical study, and real-world deployment or postdeployment monitoring. These categories were not treated as ordinal maturity levels. When a study reported multiple evaluation-setting categories, all applicable categories were recorded.

GraphRAG systems were identified using a minimum criterion requiring explicit use of graph structure at inference time to retrieve, organize, or assemble evidence that conditioned generation [27]. This included retrieval of nodes, paths, subgraphs, graph-derived summaries, or graph-derived community summaries. Studies that referenced the use of a knowledge graph without inference-time graph-structured retrieval or evidence assembly were not classified as GraphRAG. Classification was based on the reported inference-time role of graph structure in evidence retrieval or assembly rather than on the mere presence of a knowledge graph within the broader system architecture. Graph construction evaluation was coded separately when studies evaluated graph nodes, edges, relations, triples, or knowledge-graph construction quality. GraphRAG intermediate-artifact evaluation was coded when studies evaluated retrieved nodes, retrieved paths, retrieved subgraphs, graph-derived summaries, graph-based context, or provenance-linked graph artifacts.

When a study reported multiple tasks, multiple knowledge sources, or multiple evaluation methods, all applicable categories were coded. Information relevant to a field but not explicitly reported was coded as "not reported." Items that were structurally irrelevant to a study design or evaluation approach were coded as "not applicable." Coding decisions and adjudication notes were documented to preserve traceability [11].

Critical Appraisal of Individual Sources of Evidence

Consistent with scoping review methodology, we did not conduct formal methodological quality appraisal or risk-of-bias assessment [24,25]. This was because the aim of the

review was to map the range and characteristics of evaluation practices rather than to estimate intervention effects or exclude studies on the basis of methodological quality.

Reporting-Related Assessment

We conducted a structured assessment of evaluation reporting completeness to characterize transparency of evaluation practices and to support interpretation of methodological gaps across the evidence base. This assessment was descriptive rather than evaluative and was intended to characterize reporting transparency, not to rate methodological quality. Reporting assessment focused on whether key information necessary to interpret and compare evaluation results was explicitly described [11].

For each included study, we recorded reporting of key system and evaluation elements, including (1) retrieval and knowledge source description, including corpus or source, retrieval unit and granularity, retriever, and reranking components; (2) grounding and faithfulness evaluation, including definition, operationalization, and unit of analysis; (3) human evaluation procedures, including rater background and IRR; (4) automated evaluation procedures, including automated metrics, LLM-as-judge evaluation, and bias-control measures specifically applied when an LLM was used as an evaluator; (5) formal safety-related evaluation; and (6) GraphRAG construction evaluation and GraphRAG intermediate-artifact evaluation when applicable.

For studies involving real patient data, we additionally recorded reporting of ethical governance elements, including deidentification procedures and institutional review board approval, exemption, or waiver [11,22]. No studies were excluded on the basis of reporting completeness or ethical reporting. Findings were synthesized descriptively to identify common reporting gaps and areas where standardized reporting guidance could improve transparency and comparability in future studies.

Synthesis of Results

Given the heterogeneity in health care tasks, system architectures, evaluation targets, and outcome measures, quantitative meta-analysis was not appropriate. We therefore synthesized findings using descriptive statistics and narrative synthesis to map evaluation practices across included studies [24]. For the structured charting domains, we summarized counts and proportions of studies reporting the corresponding evaluation construct or design element.

Studies were retained as the unit of analysis, and multilabel coding was permitted when a study reported multiple tasks, evaluation layers, knowledge sources, or evaluation-setting categories. Results were organized to reflect the layer-specific evaluation framework defined a priori. Synthesis addressed application settings and evaluation contexts; system design patterns relevant to evaluation; retrieval-layer evaluation; grounding and faithfulness evaluation; evaluation modalities, including human evaluation and LLM-as-judge evaluation; formal safety-related evaluation; GraphRAG construction and intermediate-artifact evaluation; descriptive evaluation-setting categories; and selected reporting and governance observations.

Where informative, cross-tabulations were used to support descriptive comparisons across task categories, evaluation-setting categories, evaluation domains, and system design features; no statistical hypothesis testing was performed. Evaluation-setting categories were used descriptively and were not analyzed as an ordinal maturity scale. Findings were summarized using tables and figures to support transparency and interpretability. Each included study contributed equally to the synthesis.

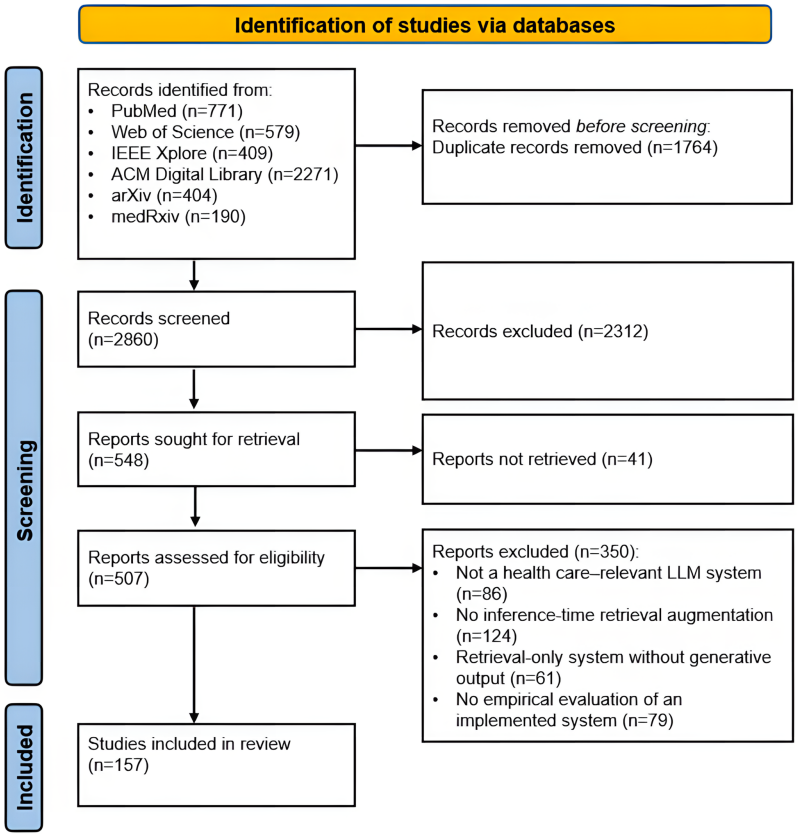
To visualize evaluation gaps, we constructed an evidence-and-gap map by cross-classifying descriptive evaluation-setting categories with key evaluation domains. Evaluation-setting categories included offline-only evaluation, simulated vignette or case evaluation, workflow pilot or user study, prospective clinical study, and real-world deployment or postdeployment monitoring. Evaluation domains included retrieval-layer evaluation, fine-grained evidence verification, citation and source correctness evaluation, conflict or contradiction handling, formal safety-related evaluation, human evaluation with IRR reporting, LLM-as-judge evaluation with bias-control measures, GraphRAG construction evaluation, and GraphRAG intermediate-artifact evaluation. Evaluation-setting categories and evaluation domains were coded as nonmutually exclusive; therefore, individual studies could contribute to more than one cell.

Results

Search Results and Study Selection

A total of 157 studies met the inclusion criteria and were included in this scoping review [28-184]. The process of study identification, screening, and inclusion is summarized in the PRISMA-ScR flow diagram (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram of the study selection process. This diagram outlines the systematic literature search and screening procedure conducted in accordance with PRISMA-ScR guidelines. It details the number of records identified from databases and preprint platforms (PubMed, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, arXiv, and medRxiv), the number of duplicates removed, and the stepwise exclusion reasons applied during screening and full-text eligibility assessment. A final total of 157 studies met the inclusion criteria for the scoping review. Searches were conducted through May 14, 2026. Backward and forward citation tracking was performed for all included studies. LLM: large language model.



Study Characteristics and Application Settings

The included studies (N=157 [28-184]) covered a broad range of health care tasks and evaluation contexts (Table 2; Table S1 in Multimedia Appendix 2). Clinical question answering was the most frequently studied application, reported in 89 (56.7%) studies. Other commonly reported applications included clinical decision support tasks (n=70, 44.6%), patient or caregiver education (n=27, 17.2%), and summary or report generation (n=24, 15.3%). Task categories were not mutually exclusive, and some systems addressed multiple downstream tasks. Overall, the evidence base remained concentrated in question answering and decision-support applications, with smaller clusters of patient education, report generation, medical visual question answering, and administrative or operational support applications.

Table 2. Characteristics of included studies and application settings (N=157). Percentages use all included studies (N=157) as the denominator. Application task categories, evaluation-setting categories, and knowledge-source categories were coded as multilabel categories; percentages therefore are not expected to sum to 100% within those blocks. Offline-only evaluation indicates the absence of workflow-facing, prospective, or deployment-level evaluation; studies could also be coded as simulated vignette or case evaluation when applicable. Evaluation-setting categories were descriptive, nonmutually exclusive categories rather than ordinal maturity levels. Full study-level coding is provided in Table S1 in Multimedia Appendix 2.

Domain and item	Count, n (%)
Application task categories	
Clinical question answering	89 (56.7)
Clinical decision support	70 (44.6)
Patient or caregiver education	27 (17.2)
Summary or report generation	24 (15.3)
Medical visual question answering	7 (4.5)
Medical imaging	6 (3.8)

Domain and item	Count, n (%)
Evaluation settings	
Offline-only evaluation	140 (89.2)
Simulated vignette or case evaluation	37 (23.6)
Any workflow-facing, prospective, or deployment-level evaluation	17 (10.8)
Workflow pilot or user study	17 (10.8)
Prospective clinical study	2 (1.3)
Real-world deployment or postdeployment monitoring	3 (1.9)
Knowledge source categories	
Public or public-mixed knowledge sources	119 (75.8)
Private institutional knowledge sources, including mixed sources	27 (17.2)
Graph-structured knowledge sources	7 (4.5)
EHR[a]-related subset	
Retrieval corpus explicitly EHR-related	21 (13.4)

[a]EHR: electronic health record.

Across evaluation-setting categories, 140 (89.2%) studies reported offline-only evaluation and did not describe workflow pilots, prospective studies, or real-world deployment or postdeployment monitoring (Table 2; Table S1 in Multimedia Appendix 2). Simulated vignette or case evaluation was reported in 37 (23.6%) studies. Only 17 (10.8%) studies reported at least 1 workflow-facing, prospective, or deployment-level evaluation setting. This included workflow pilot or user study (n=17, 10.8%), prospective clinical study (n=2, 1.3%), and real-world deployment or postdeployment monitoring (n=3, 1.9%). These evaluation-setting categories were descriptive and not mutually exclusive. Figure 2 visualizes the relationship between application task category and evaluation-setting category.

Figure 2. Heat map of health care application tasks by evaluation-setting category. The heat map shows study counts across health care application tasks and evaluation-setting categories, with evaluation-setting categories ordered descriptively from offline-only evaluation to real-world deployment or postdeployment monitoring (offline-only evaluation, simulated vignette or case evaluation, workflow pilot or user study, prospective clinical study, and real-world deployment or postdeployment monitoring). Cell labels indicate the number of included studies in each task-setting combination. Because application task and evaluation-setting category were coded as nonmutually exclusive categories, individual studies could contribute to multiple cells; therefore, row and column totals do not sum to the 157-study corpus. Evaluation-setting categories were used descriptively and were not treated as ordinal maturity levels. Full study-level coding is provided in Table S1 in Multimedia Appendix 2.

Knowledge sources used for retrieval varied across studies (Table 2; Table S1 in Multimedia Appendix 2). Public or public-mixed knowledge sources were reported in 119 (75.8%) studies, whereas 27 (17.2%) studies used private institutional sources either exclusively or in combination with other sources. Graph-structured knowledge sources were coded as the primary knowledge-source category in 7 (4.5%) studies. Retrieval corpora explicitly described as EHR-related were reported in 21 (13.4%) studies.

RAG System Design Patterns Relevant to Evaluation

System architectures and retrieval configurations were heterogeneous. Dense or vector-based retrieval was commonly reported, often alongside hybrid sparse-dense retrieval, reranking, or domain-specific corpus construction. Systems meeting the prespecified GraphRAG minimum criterion were reported in 27 (17.2%) studies. This count reflects studies in which graph structure participated in inference-time retrieval, evidence organization, reasoning, or generation, rather than studies that merely referenced a knowledge graph as background knowledge.

Evaluation-Method Coverage Across Included Studies

Evaluation-method coverage across the included studies is summarized in Table 3. Overall, end-to-end task performance evaluation was reported more consistently than layer-specific retrieval assessment, evidence-linkage verification, citation and source correctness evaluation, conflict or contradiction handling, GraphRAG-specific artifact evaluation, and implementation-facing evaluation indicators.

Table 3 provides the corresponding counts and percentages for the principal evaluation domains, GraphRAG indicators, and governance-related reporting elements.

Table 3. Evaluation methods, graph-structured retrieval-augmented generation (GraphRAG) indicators, and governance-related reporting elements. Unless otherwise specified, percentages use all included studies (N=157) as the denominator. Interrater reliability uses the subgroup denominator of studies with human evaluation (n=94). Bias-control measures among large language model (LLM)–as-judge studies use the subgroup denominator of studies using LLM-as-judge evaluation (n=41). GraphRAG intermediate-artifact evaluation and graph construction evaluation use the subgroup denominator of studies meeting the GraphRAG minimum criterion (n=27). Deidentification and institutional review board (IRB) reporting use the subgroup denominator of studies using real patient data (n=49). Fine-grained evidence verification includes claim-level, statement-level, citation-level, source-level, or sentence-level verification. Evidence-verification granularity categories were nonmutually exclusive because individual studies could report distinct evaluation components at different verification units. Data are aligned with Tables S2 and S3 in Multimedia Appendix 2.

Domain and item	Studies, n/N (%)
Retrieval-layer evaluation	
Explicit retrieval-layer evaluation	47/157 (29.9)
Grounding and evidence linkage	
Grounding and faithfulness evaluation	41/157 (26.1)
Fine-grained evidence verification	22/157 (14)
Citation and source correctness evaluation	12/157 (7.6)
Conflict or contradiction handling evaluation	9/157 (5.7)
Evidence-verification granularity	
No evidence-support verification	116/157 (73.9)
Answer-level verification	20/157 (12.7)
Claim-level or statement-level verification	17/157 (10.8)
Citation-level or source-level verification	21/157 (13.4)
Evaluation modalities	
Human evaluation used	94/157 (59.9)
Interrater reliability reported among studies with human evaluation	26/94 (27.7)
Automated metrics reported	117/157 (74.5)
LLM-as-judge evaluation used	41/157 (26.1)
Bias-control measures reported among studies using LLM-as-judge evaluation	15/41 (36.6)
Safety-related evaluation	
Formal safety-related evaluation	45/157 (28.7)
GraphRAG-specific indicators	
Studies meeting GraphRAG minimum criterion	27/157 (17.2)
GraphRAG intermediate-artifact evaluation reported	11/27 (40.7)
Graph construction evaluation reported	6/27 (22.2)

Domain and item	Studies, n/N (%)
Governance among studies using real patient data	
Studies using real patient data	49/157 (31.2)
Deidentification reported among studies using real patient data	33/49 (67.3)
IRB approval, exemption, or waiver reported among studies using real patient data	23/49 (46.9)

Retrieval-Layer Evaluation

Independent retrieval-layer evaluation was reported in 47 (29.9%) studies (Table 3; Table S2 in Multimedia Appendix 2). In the remaining 110 (70.1%) studies, retrieval-layer evaluation was not reported, and evaluation was conducted at the level of final generated outputs or through qualitative examples. Thus, although all included systems used inference-time RAG, only a subset reported retrieval as a separable evaluation layer.

Among the 47 studies reporting retrieval-layer evaluation, commonly reported metric families included precision-based metrics (n=23, 48.9%), recall metrics (n=32, 68.1%), mean reciprocal rank (n=8, 17%), and normalized discounted cumulative gain (n=6, 12.8%; Table S2 in Multimedia Appendix 2). Reporting of retrieval-layer evaluation varied substantially across studies. Some studies reported standard information retrieval metrics, whereas others used context precision, context recall, retrieval accuracy, relevance judgments, or structured assessment of retrieved evidence. Studies that only compared final answer accuracy, final task performance, or qualitative examples without explicit assessment of retrieved evidence were not counted as reporting retrieval-layer evaluation.

Evaluation of Grounding and Faithfulness

Grounding and faithfulness evaluation was reported in 41 (26.1%) studies (Table 3; Table S2 in Multimedia Appendix 2). Operationalizations varied across studies. Citation and source correctness evaluation was reported in 12 (7.6%) studies, and fine-grained evidence verification was reported in 22 (14%) studies. Conflict or contradiction handling evaluation was reported in 9 (5.7%) studies. These categories were not mutually exclusive.

Studies also varied in the granularity at which evidence linkage was assessed (Table 3; Table S2 in Multimedia Appendix 2). Answer-level verification was reported in 20 (12.7%) studies. Claim-level or statement-level verification was reported in 17 (10.8%) studies. Citation-level or source-level verification was reported in 21 (13.4%) studies. Overall, 116 of 157 studies (73.9%) reported no evidence-support verification. Verification units were therefore often broad, incompletely specified, or absent.

Evaluation Modalities: Human and Automated End Points

Human evaluation was reported in 94 (59.9%) studies (Table 3; Table S3 in Multimedia Appendix 2). IRR was reported in 26 (27.7%) of the human-evaluated studies.

Automated metrics were reported in 117 (74.5%) studies (Table 3; Table S3 in Multimedia Appendix 2). These included task-performance metrics, retrieval metrics, lexical-overlap or semantic-similarity metrics, and automated evaluation frameworks, depending on the study design and evaluation target.

LLM-as-judge evaluation was reported in 41 (26.1%) studies. Among the 41 (26.1%) studies using LLM-as-judge evaluation, 15 (36.6%) studies reported at least 1 bias-control measure (Table 3; Table S3 in Multimedia Appendix 2).

Formal safety-related evaluation was reported in 45 (28.7%) studies. Safety was operationalized heterogeneously across studies, including explicit safety ratings, clinical risk or harm rubrics, medication and contraindication safety end points, harmful-content blocking, suicide risk stratification, triage harm indices, regulatory harm assessment, and hallucination-related risk evaluation. Table 4 summarizes the primary safety-related operationalization family assigned to each study with formal safety-related evaluation. Methodological safeguards, such as IRR reporting and bias-control measures for LLM-as-judge evaluation, were less frequently reported than the use of evaluation end points themselves.

Table 4. Primary operationalization families for formal safety-related evaluation among studies with formal safety-related assessment (n=45). Categories reflect the primary safety-related operationalization family assigned to each study with formal safety-related evaluation (n=45). Categories are mutually exclusive for this main-text summary, although individual studies could include secondary safety-related elements. Study-level coding is provided in Table S3 in Multimedia Appendix 2. Percentages use studies with formal safety-related evaluation (n=45) as the denominator.

Primary operationalization family	Studies, n (%)	Operationalization and evaluation modality
Explicit safety, harm, or clinical-risk scoring within expert or LLM[a] evaluation rubrics	22 (48.9)	Broad safety scoring, harmfulness ratings, clinical-risk ratings, or safety dimensions embedded in expert, clinician, user, or LLM-as-judge evaluation rubrics.
Clinical management safety end points and harm consequences	10 (22.2)	Medication safety, drug contraindication screening, antibiotic or opioid safety, refusal or escalation behavior, undertriage or overtriage harm, medical or regulatory harm, or other task-specific clinical safety end points.
Hallucination, factuality, or evidence misalignment framed as safety	6 (13.3)	Hallucination, factuality, source suitability, unsupported clinical content, or evidence misalignment explicitly treated as a safety-relevant failure mode.

Primary operationalization family	Studies, n (%)	Operationalization and evaluation modality
Harmful-content, crisis, adversarial, or misuse-safety safeguards	5 (11.1)	Harmful-content blocking, crisis safeguards, adversarial safety prompting, misuse-oriented testing, or safety monitoring for mental health or patient-facing systems.
Public-health misinformation and fact-checking safety evaluation	2 (4.4)	Fact-checking or misinformation evaluation framed as public-health risk mitigation.

[a]LLM: large language model.

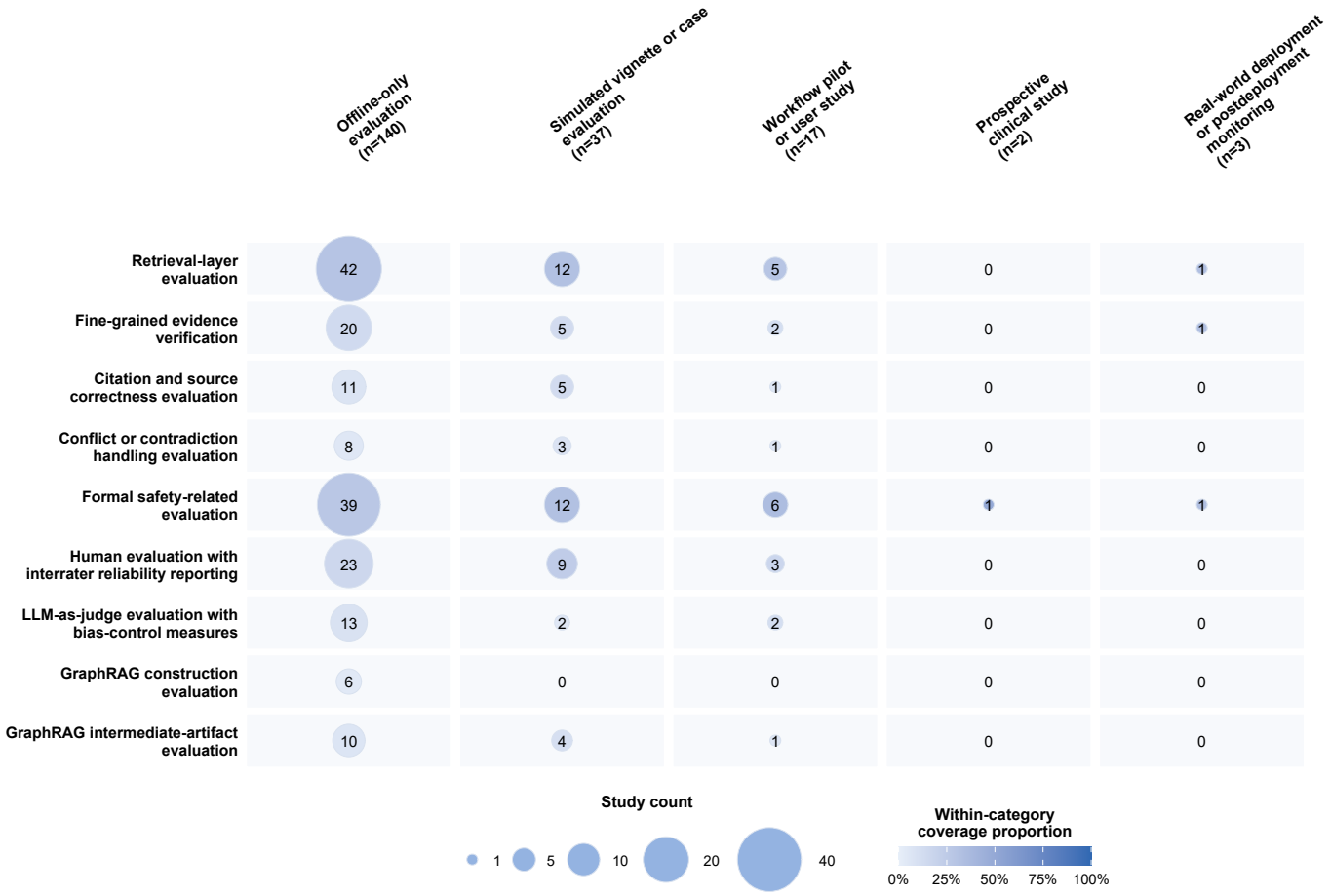
Exploratory Findings for GraphRAG-Specific Evaluation

In total, 27 studies met the prespecified minimum GraphRAG criterion and were included in GraphRAG-specific analyses (Table 3; Table S3 in Multimedia Appendix 2). All 27 GraphRAG studies reported end-to-end output evaluation. GraphRAG intermediate-artifact evaluation was reported in 11 of the 27 (40.7%) GraphRAG studies. Graph construction evaluation was reported in 6 (22.2%) studies. Evaluation of retrieved paths, subgraphs, graph-derived summaries, or provenance-linked graph artifacts remained uncommon. Given the still limited number of studies meeting the minimum GraphRAG criterion, these findings should be interpreted as exploratory.

Reporting, Governance, and Study-Level Audit Trail

Reporting-related details relevant to the interpretation of evaluation procedures varied across the included studies. Graph-specific reporting and governance indicators also varied across studies (Table 3; Table S3 in Multimedia Appendix 2). Study-level reporting of knowledge-source type, indexing unit, and retriever type was heterogeneous, and graph-specific evaluation of intermediate artifacts was uncommon even among studies meeting the minimum GraphRAG criterion. Real patient data use was reported in 49 (31.2%) studies. Among these studies, deidentification procedures were reported in 33 (67.3%) studies and institutional review board approval, exemption, or waiver was reported in 23 (46.9%) studies. Concise study-level core coding tables and the source data for Figure 3 are provided in Tables S1-S4 in Multimedia Appendix 2.

Figure 3. Evidence-and-gap map of evaluation domains by evaluation-setting category. The evidence-and-gap map cross-classifies descriptive evaluation-setting categories with key evaluation domains. Evaluation-setting categories include offline-only evaluation, simulated vignette or case evaluation, workflow pilot or user study, prospective clinical study, and real-world deployment or postdeployment monitoring. Evaluation domains include retrieval-layer evaluation, fine-grained evidence verification, citation and source correctness evaluation, conflict or contradiction handling, formal safety-related evaluation, human evaluation with interrater reliability reporting, LLM-as-judge evaluation with bias-control measures, GraphRAG construction evaluation, and GraphRAG intermediate-artifact evaluation. Circle area indicates the number of studies reporting the corresponding evaluation domain. Fill intensity indicates within-category coverage proportion, calculated as the number of studies in a given evaluation-setting category reporting the evaluation domain divided by the total number of studies in that category. Evaluation-setting categories and evaluation domains were coded as nonmutually exclusive; therefore, counts across rows or columns were not expected to sum to the total corpus. Source data for the map are provided in Table S4 in Multimedia Appendix 2. GraphRAG: graph-structured retrieval-augmented generation; LLM: large language model.



Evidence and Gap Map by Evaluation-Setting Category

Figure 3 presents an evidence-and-gap map (Table S4 in Multimedia Appendix 2) cross-classifying descriptive evaluation-setting categories with key evaluation domains. The map shows that evaluation coverage was concentrated in offline-only and simulated case-based evaluation categories. Across workflow-facing and prospective settings, coverage was sparse for fine-grained evidence verification, conflict or contradiction handling, formal safety-related evaluation, IRR reporting among studies with human evaluation, judge bias-control measures for LLM-as-judge evaluation, GraphRAG construction evaluation, and GraphRAG intermediate-artifact evaluation. This pattern indicates that the literature has developed more extensively for offline and simulated case-based evaluation than for workflow-facing, safety-oriented, governance-aware, or graph-artifact-specific evaluation.

Synthesis-Informed Evaluation Considerations by Intended Evaluation Setting

Based on the mapped evaluation gaps, Table 5 presents synthesis-informed evaluation considerations by intended evaluation setting. This table is intended as a practice-oriented synthesis rather than an empirically validated checklist, implementation guideline, or ordinal maturity scale. Because the included studies were concentrated in offline-only and simulated evaluations, and because real-world deployment or postdeployment monitoring was sparsely represented, considerations for workflow-facing, prospective, and deployment-level settings should be interpreted as future-oriented evaluation considerations informed by the observed gaps and the implementation-relevant issues identified in this review.

Table 5. Synthesis-informed evaluation considerations by intended evaluation setting. This table is a practice-oriented synthesis based on the observed evaluation gaps and implementation-relevant considerations. It is not an empirically validated reporting checklist, implementation guideline, or ordinal maturity scale. The elements should be interpreted as considerations that may be adapted to task risk, user group, knowledge source, and intended use context.

Evaluation setting and use context	Evaluation components to consider	Evidence expectations and common failure modes
Offline-only evaluation: retrospective benchmark or held-out test set without workflow embedding	Retrieval-layer metrics; end-to-end task metrics; at least one evidence-linkage check at response or claim level; explicit reporting of corpus, retriever, and generator	Reference evidence or relevance labels; retrieval unit and granularity; prompts and configuration sufficient to interpret results. Common failures include retrieval miss, unsupported synthesis, and benchmark overfitting.
Simulated vignette or case evaluation: simulated clinician- or patient-facing scenarios	Offline elements plus scenario-based human review; claim- or citation-level verification when feasible; contradiction handling; at least one task-relevant safety-related end point	Expert-authored or expert-curated cases; adjudication rule or clinician rubric; explicit handling of conflicting evidence. Common failures include unsafe advice, weak abstention, and brittle behavior under conflicting evidence.
Workflow pilot or user study: interaction in a workflow-resembling environment with intended users	Simulated-setting elements plus usability and workflow outcomes; time burden or efficiency; user trust and calibration; escalation or handoff assessment	Clearly described user group; protocolized task flow; predefined safety oversight during pilot use. Common failures include workflow disruption, overtrust, hidden latency, and poor handoff to clinicians.
Prospective clinical study: prospective assessment in live or near-live care processes	Workflow-pilot elements plus predefined safety monitoring; subgroup analysis; governance documentation; protocolized human oversight	Prospective protocol; monitoring triggers; incident definitions; ethics and governance reporting. Common failures include consequential harm, performance heterogeneity, and inadequate monitoring thresholds.
Real-world deployment or postdeployment monitoring: operational deployment or postimplementation surveillance	Ongoing drift surveillance; incident logging; feedback loops; periodic re-audit of retrieval and evidence-linkage performance; version tracking	Monitoring cadence; trigger thresholds; rollback or escalation pathways; clear governance ownership. Common failures include performance drift, silent failure, configuration drift, and surveillance without remediation.

Discussion

Principal Findings

This scoping review mapped how inference-time RAG and GraphRAG systems in health care were evaluated across retrieval, evidence linkage, end-to-end performance, safety, reporting, and evaluation-setting categories. Across studies published or posted from 2024 to May 14, 2026, evaluation practices expanded rapidly while definitions, units of analysis, and reporting conventions remained fragmented. Evaluation coverage was concentrated in offline-only and simulated evaluation settings and was sparser in workflow-facing, prospective, and deployment-level settings, especially for retrieval-layer evaluation, fine-grained evidence verification, contradiction handling, formal safety-related evaluation, bias-control measures for LLM-as-judge evaluation, GraphRAG intermediate artifacts, and deployment-level monitoring. These findings support descriptive mapping of evaluation practices and suggest that clinical readiness claims are more interpretable when supported by layer-specific evidence [18,21,22].

Comparison With Previous Work

Independent retrieval-layer evaluation was reported in only a minority of included studies, despite inference-time retrieval being a defining feature of all eligible systems. When retrieval is evaluated only through final outputs, output-level errors are more difficult to attribute to retrieval failures, evidence selection failures, or generation failures [12]. This limits failure-mode localization and weakens comparative interpretation across RAG pipelines. The practical consequence is that end-to-end improvements can be misattributed to retrieval changes when they may primarily reflect prompt design, decoding constraints, answer formatting, or evaluator sensitivity. Conversely, apparently weak performance can be driven by corpus segmentation or retrieval unit choices rather than by limitations in generation. Without retrieval-layer evaluation and clear specification of retrieval unit and granularity, claims about the causal role of retrieval in improving factual reliability remain difficult to substantiate. This interpretation is consistent with broader RAG evaluation literature, which emphasizes that hybrid systems require layer-aware assessment because retrieval relevance, generation quality, and answer faithfulness are related but nonidentical targets [18]. It also extends previous health care LLM reviews, which have documented fragmented practices but have focused less directly on component-level evaluation and failure-mode localization in retrieval-augmented systems [5,6].

The literature frequently reported evaluations labeled as grounding, faithfulness, citation and source correctness, or hallucination assessment, but these terms often referred to partially overlapping end points. Clear construct boundaries matter because these end points are not interchangeable and can yield conflicting conclusions when treated as substitutes [185]. Previous RAG and hallucination-evaluation literature has emphasized related constructs, but the health care studies mapped in this review often operationalized them with limited granularity or inconsistent terminology. Grounding concerns whether generated claims are supported by retrieved evidence available at inference time. Faithfulness concerns whether the response remains appropriately constrained by retrieved evidence, without introducing unsupported elaborations. Factuality concerns correctness with respect to an external reference standard that may include information

not present in retrieved context. Citation and source correctness concerns whether cited sources actually support the local claims they are attached to, including placement, specificity, and claim-to-source alignment. Recent scholarship supports this distinction, showing that citation presence or citation plausibility is not equivalent to citation and source correctness, because apparently correct citations may still reflect post hoc rationalization rather than genuine evidence use [186]. Hallucination rubrics vary widely and may mix grounding violations, factuality errors, omissions, and miscalibrated certainty depending on rubric design [20,185]. Fine-grained evidence verification is therefore important because response-level assessment can mask localized unsupported claims, whereas claim-level, citation-level, or sentence-level verification can make evidence-linkage failures more visible [187].

Graph-structured systems were evaluated primarily through end-to-end output outcomes, while evaluation of intermediate artifacts, such as graph construction validity, retrieved paths or subgraphs, graph-derived summaries, and provenance-related interpretability, remained less consistently reported. This limits the ability to substantiate the primary motivations for graph structure, particularly claims about improved traceability and interpretability. The practical consequence is that graph-structured approaches can be judged using end points that do not test their stated advantages, which weakens interpretation of when graph structure provides measurable benefits and which graph components contribute to improvements or failures. Because a limited but expanding subset of studies met the minimum GraphRAG criterion, these findings should be interpreted as exploratory rather than definitive. When interpretability or provenance is presented as a rationale for graph-structured RAG, intermediate-artifact evaluation may help interpret the added value of GraphRAG beyond conventional RAG pipelines [15].

Evaluation-setting context is central to interpreting health care evaluation evidence. Most studies evaluated systems in offline-only settings, with vignette- or case-based testing used as an intermediate step in some work. Offline evaluation is essential for controlled iteration, yet it can underrepresent workflow constraints, incomplete context, time pressure, and local practice variation, all of which directly shape reliability and clinician reliance. This observation aligns with broader health care LLM reviews finding limited high-realism prospective evidence [5,6]. Previous health care AI evaluation guidance has similarly emphasized staged evaluation before clinical implementation, including escalation, oversight, governance, and monitoring requirements as systems move closer to clinical use [23]. In this review, the evidence-and-gap map showed sparse workflow-facing, prospective, and deployment-level coverage for fine-grained evidence verification, contradiction handling, safety-related assessment, human-evaluation reliability reporting, bias-control measures for LLM-as-judge evaluation, and GraphRAG intermediate-artifact evaluation. This pattern supports aligning evaluation expectations with the intended use context. Early-stage systems may reasonably begin with retrieval-layer testing, end-to-end task metrics, and evidence-linkage checks, whereas workflow-facing systems warrant added attention to safety monitoring, contradiction handling, escalation pathways, user interaction, and governance.

Human evaluation was reported in over half of studies, often involving clinicians or domain experts. However, IRR was reported in only a minority of human-evaluated studies. Limited reporting of rater expertise, training, and reliability restricts interpretation of end points such as usefulness, appropriateness, and safety, which are constructs sensitive to rubric framing and rater background. Without reliability evidence, apparent differences between systems may reflect measurement instability rather than robust behavioral differences [188]. LLM-as-judge evaluation was used in a subset of studies, yet bias-control measures for LLM-as-judge evaluation were reported in fewer than half of those studies [189]. This pattern suggests automated evaluation is being adopted for scalability with variable reporting of safeguards. Validity threats include prompt sensitivity [190], judge drift, correlated model errors, and reward hacking [191,192]. Safety-related evaluation also requires explicit operationalization because grounded or factually accurate responses can still be unsafe when they omit warnings, express inappropriate certainty, ignore patient-specific constraints, or fail under conflicting evidence [1,193-196].

Implications for Evaluation and Reporting

Reporting of implementation-relevant study details varied across the included studies. In retrieval-augmented systems, incomplete reporting is particularly consequential because system behavior depends on corpus provenance and versioning, chunking policy, embedding and retrieval configurations, reranking settings, and citation policies [197]. A concrete consequence is that nominally similar systems may behave differently due to unreported corpus or indexing differences, while readers attribute differences to model choice or retrieval method. Improving reporting transparency therefore requires emphasizing RAG-specific determinants of behavior as first-class reporting items rather than as implementation details [8]. This observation strongly aligns with the movement toward LLM-specific reporting standards in biomedicine, such as TRIPOD-LLM [8], and emerging governance frameworks emphasizing continuous drift monitoring [198].

Based on the descriptive synthesis and the evidence-and-gap map, these findings point to several practice-oriented implications for future evaluation and reporting. Retrieval and generation may be more interpretable when evaluated as separable layers, particularly when retrieval is claimed to improve performance. Grounding, faithfulness, and citation and source correctness are more interpretable when explicitly defined and assessed at a stated verification unit. Contradiction handling, uncertainty communication, and formal safety-related evaluation are particularly relevant when systems are positioned for clinical or workflow-facing use. For human evaluation and LLM-as-judge evaluation, reporting safeguards that support measurement validity may improve interpretability, including rater expertise,

IRR, evaluator identity, prompts or rubrics, calibration, and sensitivity analyses. Graph-structured systems should consider GraphRAG construction evaluation and intermediate-artifact evaluation when interpretability, provenance, or graph-based evidence organization is presented as a key contribution. These priorities are consistent with emerging reporting guidance for LLM studies, RAG failure-mode analyses, health care AI evaluation frameworks, and literature on LLM-as-judge evaluation and medical safety evaluation [1,11,22,23,189-197].

The synthesis-informed evaluation considerations proposed in this review should therefore be interpreted as a practice-oriented synthesis to support more transparent and setting-appropriate evaluation, rather than as a mandatory checklist or an ordinal maturity scale. These considerations are intended to help align evaluation intensity with intended use context, while allowing task- and context-specific adaptation [11,22,23,197,198].

Limitations

This scoping review has limitations. First, the review focused on contemporary studies from 2024 onward, with searches conducted through May 14, 2026; earlier foundational work may therefore be underrepresented. The review was limited to English-language records, which may have excluded relevant studies reported in other languages. Second, consistent with scoping review methodology, the synthesis maps reported practices rather than estimating pooled effects, and formal risk-of-bias appraisal lay outside the review scope [24,25]. Third, findings depend on reporting completeness; procedures performed but left undescribed may be underestimated, and classification necessarily depended on study descriptions; information relevant to a field but not explicitly described was coded as not reported, and items structurally irrelevant to a study design or evaluation approach were coded as not applicable. Fourth, heterogeneity in tasks, corpora, including private and EHR-derived resources, and evaluation end points limited quantitative synthesis and supported the descriptive nature of the evidence-and-gap map. Finally, the identification of GraphRAG systems relied on prespecified operational criteria; studies may be classified differently under alternative definitions that use broader or narrower criteria for graph-structured inference-time retrieval and evidence assembly.

Conclusions

This scoping review contributes to health care generative AI evaluation by mapping how RAG and GraphRAG systems are evaluated across retrieval, evidence linkage, safety, reporting, GraphRAG-specific artifacts, and evaluation-setting categories. Unlike previous reviews that broadly survey health care LLM applications or general NLP benchmarks, this study examines how evaluation is defined, operationalized, and reported within RAG and GraphRAG systems. The evidence-and-gap map shows areas of concentrated coverage and clinically relevant gaps, while the synthesis-informed evaluation considerations translate these gaps into practice-oriented evaluation considerations. These findings may inform more transparent, layer-specific, and safety-oriented evaluation planning for systems being considered for workflow-facing implementation [11,22,23,193,198].

Acknowledgments

No generative artificial intelligence tools were used in the preparation of this manuscript.

Funding

This research was supported by the Key Program of the National Natural Science Foundation of China (72034005) and the Specialised Organised Research Project of the Chinese Institutes for Medical Research (CX23YZ02). The funders had no role in the design, conduct, analysis, interpretation, or reporting of this scoping review.

Data Availability

The study-level core coding tables used to generate the manuscript tables and figures are provided in Multimedia Appendix 2. Additional audit materials are retained by the authors and can be made available by the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: YZ, YM, YW
Data curation: YZ, YM, RG, HW
Formal analysis: YZ
Funding acquisition: YW
Investigation: YZ, YM, RG, HW
Methodology: YZ, YM, YL, YW
Project administration: YW
Resources: YL
Software: YZ, YM
Supervision: YW
Validation: YL, RG, YW
Visualization: YL, HW
Writing – original draft: YZ
Writing – review & editing: YM, RG, HW, YW

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete database-specific search strategies.
[DOCX File (Microsoft Word File), 35 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Study-level core coding tables and source data for the evidence-and-gap map.
[DOCX File (Microsoft Word File), 143 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-ScR checklist.
[DOCX File (Microsoft Word File), 45 KB-Checklist 1]

Checklist 2

PRISMA-S checklist.
[DOCX File (Microsoft Word File), 30 KB-Checklist 2]

References

1. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
2. Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: a scoping review. J Int Med Res. Jul 2025;53(7):3000605251347556. [doi: 10.1177/03000605251347556] [Medline: 40615349]
3. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
4. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. May 17, 2024;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
5. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. Mar 2026;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
6. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. Mar 7, 2025;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
7. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
8. Zhu Z, Zhang Y, Zhuang X, et al. Can we trust AI doctors? A survey of medical hallucination in large language and large vision-language models. Presented at: Findings of the Association for Computational Linguistics; Jul 27 to Aug 1, 2025:6748-6769; Vienna, Austria. [doi: 10.18653/v1/2025.findings-acl.350]
9. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a bowtie analysis. Journal of Safety Science and Resilience. Dec 2024;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
10. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res. Jul 29, 2025;27:e69678. [doi: 10.2196/69678] [Medline: 40772775]
11. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
12. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]
13. Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.398]
14. Edge D, Trinh H, Cheng N, et al. From local to global: a GraphRAG approach to query-focused summarization. arXiv. Preprint posted online on Apr 24, 2024. [doi: 10.48550/arXiv.2404.16130]

15. Peng B, Zhu Y, Liu Y, et al. Graph retrieval-augmented generation: a survey. arXiv. Preprint posted online on Aug 15, 2024. [doi: 10.48550/arXiv.2408.08921]
16. Gupta S, Ranjan R, Singh SN. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions. arXiv. Preprint posted online on Oct 3, 2024. [doi: 10.48550/arXiv.2410.12837]
17. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intelligent Medicine. May 2025;5(2):151-163. [doi: 10.1016/j.imed.2025.03.002]
18. Yu H, Gan A, Zhang K, Tong S, Liu Q. Evaluation of retrieval-augmented generation: a survey. Presented at: CCF Conference on Big Data; Aug 9-11, 2024; Qingdao, China. [doi: 10.1007/978-981-96-1024-2_8]
19. Gan A, Yu H, Zhang K, et al. Retrieval augmented generation evaluation in the era of large language models: a comprehensive survey. arXiv. Preprint posted online on Apr 21, 2025. [doi: 10.48550/arXiv.2504.14891]
20. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 16-21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.20]
21. Yang R, Wong MYH, Li H, et al. Retrieval-augmented generation in medicine: a scoping review of technical implementations, clinical applications, and ethical considerations. arXiv. Preprint posted online on Nov 8, 2025. [doi: 10.48550/arXiv.2511.05901]
22. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform. Oct 2021;28(1):e100444. [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
23. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. May 18, 2022;377:e070904. [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
24. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
25. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. Oct 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. Jan 26, 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
27. Han H, Wang Y, Shomer H, et al. Retrieval-augmented generation with graphs (GraphRAG). arXiv. Preprint posted online on Dec 31, 2024. [doi: 10.48550/arXiv.2501.00309]
28. Karim A, Uzuner O. MasonNLP at MEDIQA-WV 2025: multimodal retrieval-augmented generation with large language models for medical VQA. Presented at: Proceedings of the 7th Clinical Natural Language Processing Workshop; Oct 30, 2025. URL: https://aclanthology.org/2025.clinicalnlp-1.10 [Accessed 2026-07-05]
29. Wada A, Tanaka Y, Nishizawa M, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Digit Med. Jul 2, 2025;8(1):395. [doi: 10.1038/s41746-025-01802-z] [Medline: 40604147]
30. Fink A, Nattenmüller J, Rau S, et al. Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: a proof-of-concept study. Eur Radiol. Aug 2025;35(8):5091-5098. [doi: 10.1007/s00330-025-11445-z] [Medline: 39953150]
31. Kelly A, Noctor E, Ryan L, van de Ven P. The effectiveness of a custom AI chatbot for type 2 diabetes mellitus health literacy: development and evaluation study. J Med Internet Res. May 5, 2025;27:e70131. [doi: 10.2196/70131] [Medline: 40324160]
32. Su AY, Knebel A, Xu AY, et al. Evaluation of retrieval-augmented generation and large language models in clinical guidelines for degenerative spine conditions. Eur Spine J. Mar 2026;35(3):1301-1310. [doi: 10.1007/s00586-025-08994-8] [Medline: 40619525]
33. Ho CM, Guan S, Mok PKL, et al. Development and validation of a large language model-powered chatbot for neurosurgery: mixed methods study on enhancing perioperative patient education. J Med Internet Res. Jul 15, 2025;27:e74299. [doi: 10.2196/74299] [Medline: 40663377]
34. Gan C, Yang D, Hu B, et al. POLYRAG: integrating polyviews into retrieval-augmented generation for medical applications. arXiv. Preprint posted online on Apr 21, 2025. [doi: 10.48550/arXiv.2504.14917]
35. Tarabanis C, Khurshid S, Karamanou A, et al. Cardiology knowledge assessment of retrieval-augmented open versus proprietary large language models. PLOS Digit Health. Mar 2026;5(3):e0001029. [doi: 10.1371/journal.pdig.0001029] [Medline: 41818295]

36. de Jesus DDR, de Souza Júnior AP, de Albergaria ET, et al. Enhanced LLM-supported instructions for medication use through retrieval-augmented generation. Comput Biol Med. Nov 2025;198(Pt A):111135. [doi: 10.1016/j.compbiomed.2025.111135] [Medline: 41077040]
37. Baur D, Ansorg J, Heyde CE, Voelker A. Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: mixed-methods study. JMIR AI. Oct 23, 2025;4:e75262. [doi: 10.2196/75262] [Medline: 41134117]
38. Steybe D, Poxleitner P, Aljohani S, et al. Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. J Craniomaxillofac Surg. Apr 2025;53(4):355-360. [doi: 10.1016/j.jcms.2024.12.009] [Medline: 39799075]
39. Yu D, Wang Y, Jin S, et al. YpathRAG: a retrieval-augmented generation framework and benchmark for pathology. arXiv. Preprint posted online on Oct 7, 2025. [doi: 10.48550/arXiv.2510.08603]
40. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res. Nov 8, 2024;26:e58041. [doi: 10.2196/58041] [Medline: 39046096]
41. Li D, Liang J, Li W, Wang X, Cao L, Yu K. CliCARE: grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. AAAI. 2026;40(37):31554-31562. [doi: 10.1609/aaai.v40i37.40421]
42. Busch F, Kaibel L, Nguyen H, et al. Evaluation of a retrieval-augmented generation-powered chatbot for Pre-CT informed consent: a prospective comparative study. J Imaging Inform Med. Dec 2025;38(6):4312-4323. [doi: 10.1007/s10278-025-01483-w] [Medline: 40119020]
43. Kim H, Sohn J, Gilson A, et al. Rethinking retrieval-augmented generation for medicine: a large-scale, systematic expert evaluation and practical insights. arXiv. Preprint posted online on Nov 10, 2025. [doi: 10.48550/arXiv.2511.06738]
44. Sohn J, Park Y, Yoon C, et al. Rationale-guided retrieval augmented generation for medical question answering. Presented at: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics; Apr 29 to May 4, 2025; Albuquerque, New Mexico. [doi: 10.18653/v1/2025.naacl-long.635]
45. Wu J, Zhu J, Qi Y, et al. Medical graph RAG: evidence-based medical large language model via graph retrieval-augmented generation. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025:28443-28467; Vienna, Austria. [doi: 10.18653/v1/2025.acl-long.1381]
46. Ou J, Huang T, Zhao Y, Yu Z, Lu P, Ying R. Experience retrieval-augmentation with electronic health records enables accurate discharge QA. arXiv. Preprint posted online on Mar 23, 2025. [doi: 10.48550/arXiv.2503.17933]
47. Soman K, Rose PW, Morris JH, et al. Biomedical knowledge graph-optimized prompt generation for large language models. Bioinformatics. Sep 2, 2024;40(9):btae560. [doi: 10.1093/bioinformatics/btae560] [Medline: 39288310]
48. Wołk K. Evaluating retrieval-augmented generation variants for clinical decision support: hallucination mitigation and secure on-premises deployment. Electronics (Basel). 2025;14(21):4227. [doi: 10.3390/electronics14214227]
49. Park J, Yoon B, Kim S, Choi K. RA-RRG: multimodal retrieval-augmented radiology report generation with key phrase extraction. Presented at: Findings of the Association for Computational Linguistics; Jul 2-7, 2026; San Diego, California, United States. [doi: 10.18653/v1/2026.findings-acl.247]
50. Garza L, Kotal A, Grasso MA, Umucu E. Retrieval-augmented framework for LLM-based clinical decision support. arXiv. Preprint posted online on Oct 1, 2025. [doi: 10.48550/arXiv.2510.01363]
51. Stuhlmann L, Saxer MA, Fürst J. Efficient and reproducible biomedical question answering using retrieval augmented generation. Presented at: 2025 IEEE Swiss Conference on Data Science (SDS); Jun 26-27, 2025; Zürich, Switzerland. [doi: 10.1109/SDS66131.2025.00029]
52. Xiong L, Zeng Q, Luo W, Liu R. Nursing retrieval-augmented generation: retrieval augmented generation for nursing question answering with large language models. Int J Nurs Sci. Nov 2025;12(6):516-523. [doi: 10.1016/j.ijnss.2025.10.005] [Medline: 41367595]
53. Sun L, Zhao JJ, Han W, Xiong C. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. Presented at: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics; Apr 29 to May 4, 2025; Albuquerque, New Mexico. [doi: 10.18653/v1/2025.naacl-long.28]
54. Sesen MB, Au Yeung J, Asgari E. Development and validation of Retrieval Augmented Generation (RAG) and GraphRAG for complex clinical cases. medRxiv. Preprint posted online on Nov 27, 2025. [doi: 10.1101/2025.11.25.25341010]
55. Moureau MK, Davis B, Hong CX. Development and evaluation of an augmented artificial intelligence model for urogynecology queries. Int Urogynecol J. May 2026;37(5):1423-1428. [doi: 10.1007/s00192-025-06446-x] [Medline: 41364238]

56. Lewis M, Thio S, Roberts A, et al. Grounding large language models in clinical evidence: a retrieval-augmented generation system for querying UK NICE clinical guidelines. arXiv. Preprint posted online on Oct 3, 2025. [doi: 10.48550/arXiv.2510.02967]
57. Welsh M, Lopez-Rippe J, Alkhulaifat D, et al. Custom-tailored radiology research via retrieval-augmented generation: a secure institutionally deployed large language model system. Inventions. 2025;10(4):55. [doi: 10.3390/inventions10040055]
58. 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. Aug 2024;156:104662. [doi: 10.1016/j.jbi.2024.104662] [Medline: 38880236]
59. 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. Presented at: Findings of the Association for Computational Linguistics; Nov 4-9, 2025:12682-12701; Suzhou, China. URL: https://aclanthology.org/2025.findings-emnlp.679/ [Accessed 2026-07-06] [doi: 10.18653/v1/2025.findings-emnlp.679]
60. Son N, Kang I, Kim I, Lee K, Nam S, Lee D. Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system. Healthc Inform Res. Jul 2025;31(3):218-225. [doi: 10.4258/hir.2025.31.3.218] [Medline: 40840929]
61. Abdullayev N, Kottlors J, Habibov H, et al. European guideline informed RAG-based GPT-4 decision support tool in tumor board meetings for breast cancer treatment. Eur J Surg Oncol. Nov 2025;51(11):110384. [doi: 10.1016/j.ejso.2025.110384] [Medline: 40845749]
62. Valan P, Venugopal P. Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility-accuracy-readability study of the DIAN AI assistant. Front Artif Intell. 2025;8:1640994. [doi: 10.3389/frai.2025.1640994] [Medline: 41058908]
63. Xia P, Zhu K, Li H, et al. RULE: reliable multimodal RAG for factuality in medical vision language models. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, Florida, USA. [doi: 10.18653/v1/2024.emnlp-main.62]
64. DiGiacomo P, Wang H, Fang J, Leng Y, Brode WM, Ding Y. Demo: guide-RAG: evidence-driven corpus curation for retrieval-augmented generation in long COVID. arXiv. Preprint posted online on Oct 17, 2025. [doi: 10.48550/arXiv.2510.15782]
65. Chen R, Zhang S, Zheng Y, Yu Q, Wang C. Enhancing treatment decision-making for low back pain: a novel framework integrating large language models with retrieval-augmented generation technology. Front Med (Lausanne). 2025;12:1599241. [doi: 10.3389/fmed.2025.1599241] [Medline: 40438365]
66. Wind S, Sopa J, Truhn D, et al. Multi-step retrieval and reasoning improves radiology question answering with large language models. NPJ Digit Med. Dec 22, 2025;8(1):790. [doi: 10.1038/s41746-025-02250-5] [Medline: 41429891]
67. Kuo SM, Tai SK, Lin HY, Chen RC. Automated clinical trial data analysis and report generation by integrating retrieval-augmented generation (RAG) and large language model (LLM) technologies. AI. 2025;6(8):188. [doi: 10.3390/ai6080188]
68. Liang S, Zhang L, Zhu H, Wang W, He Y, Zhou D. RGAR: recurrence generation-augmented retrieval for factual-aware medical question answering. Presented at: Findings of the Association for Computational Linguistics; Nov 4-9, 2025; Suzhou, China. [doi: 10.18653/v1/2025.findings-emnlp.214]
69. Liu S, An LCI, Mihalcea R. Patient-centered RAG for oncology visit aid following the Ottawa Decision Guide. arXiv. Preprint posted online on Jul 5, 2025. [doi: 10.48550/arXiv.2507.04026]
70. Myers S, Dligach D, Miller TA, et al. Evaluating retrieval-augmented generation vs. long-context input for clinical reasoning over EHRs. arXiv. Aug 20, 2025. [doi: 10.48550/arXiv.2508.14817]
71. Hassan T, Karim MF, Jeelani H, Behnam E, Green R, Syed FJ. Optimizing medical question-answering systems: a comparative study of fine-tuned and zero-shot large language models with RAG framework. arXiv. Preprint posted online on Dec 5, 2025. [doi: 10.48550/arXiv.2512.05863]
72. Sekar T, Kushal, Shankar S, Mohammed S, Fiaidhi J. Investigations on using evidence-based GraphRag pipeline using LLM tailored for USMLE style questions. medRxiv. Preprint posted online on May 5, 2025. [doi: 10.1101/2025.05.03.25325604]
73. Parameswaran V, Bernard J, Bernard A, et al. Evaluating large language models and retrieval-augmented generation enhancement for delivering guideline-adherent nutrition information for cardiovascular disease prevention: cross-sectional study. J Med Internet Res. Oct 7, 2025;27:e78625. [doi: 10.2196/78625] [Medline: 41057043]
74. Zhang W, Guo J, Zhang H, et al. Patho-AgenticRAG: towards multimodal agentic retrieval-augmented generation for pathology VLMs via reinforcement learning. AAAI. 2026;40(35):29921-29929. [doi: 10.1609/aaai.v40i35.40239]

75. Xu X, Liu S, Zhu L, et al. Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education. Int J Med Inform. Nov 2025;203:106006. [doi: 10.1016/j.ijmedinf.2025.106006] [Medline: 40479778]
76. 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 on Apr 30, 2025. [doi: 10.48550/arXiv.2504.21252]
77. Zhao X, Liu S, Yang SY, Miao C. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. Presented at: WWW '25; Apr 28 to May 2, 2025; Sydney, NSW, Australia. [doi: 10.1145/3696410.3714782]
78. Ge X, Murtaza S, Cortez A, Alemzadeh H. Expert-guided prompting and retrieval-augmented generation for emergency medical service question answering. AAAI. 2025;40(36):30798-30806. [doi: 10.1609/aaai.v40i36.40337]
79. Mali Y, Zeng Z, Heo K, et al. A chatbot for the management of bipolar disorder: using retrieval-augmented generation with an open-weight large language model to answer clinical questions based on the CANMAT and ISBD 2018 guidelines for bipolar disorder. medRxiv. Preprint posted online on Jan 7, 2026. [doi: 10.64898/2025.11.30.25341311]
80. Yu Y, Li L, Li Y. Augmenting large language models and retrieval-augmented generation with an evidence-based medicine-enabled agent system. medRxiv. Preprint posted online on Oct 20, 2025. [doi: 10.1101/2025.10.17.25338266]
81. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. Apr 5, 2025;8(1):187. [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
82. Ji Y, Zhang H, Wang Y. Bias evaluation and mitigation in retrieval-augmented medical question-answering systems. arXiv. Preprint posted online on Mar 19, 2025. [doi: 10.48550/arXiv.2503.15454]
83. Huang Z, Xue K, Fan Y, et al. Tool calling: enhancing medication consultation via retrieval-augmented large language models. arXiv. Apr 27, 2024. [doi: 10.48550/arXiv.2404.17897]
84. Wang Z, Gao J, Danek B, et al. InformGen: an AI copilot for accurate and compliant clinical research consent document generation. arXiv. Preprint posted online on Apr 1, 2025. [doi: 10.48550/arXiv.2504.00934]
85. Wang Z, Khatibi E, Rahmani AM. MedCoT-RAG: causal chain-of-thought RAG for medical question answering. Presented at: 2025 IEEE 21st International Conference on Body Sensor Networks (BSN); Nov 3-5, 2025; Los Angeles, CA, USA. [doi: 10.1109/BSN66969.2025.11337389]
86. Xu R, Hong Y, Zhang F, Xu H. Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. Sci Rep. Dec 28, 2024;14(1):30794. [doi: 10.1038/s41598-024-81052-3] [Medline: 39730573]
87. Luo MJ, Pang J, Bi S, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. Sep 1, 2024;142(9):798-805. [doi: 10.1001/jamaophthalmol.2024.2513] [Medline: 39023885]
88. Zhang G, Xu Z, Jin Q, et al. Leveraging long context in retrieval augmented language models for medical question answering. NPJ Digit Med. May 2, 2025;8(1):239. [doi: 10.1038/s41746-025-01651-w] [Medline: 40316710]
89. Low YS, Jackson ML, Hyde RJ, et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. Digit Health. 2025;11:20552076251348850. [doi: 10.1177/20552076251348850] [Medline: 40510193]
90. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology. Nov 1, 2024;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
91. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI. Feb 2024;1(2). [doi: 10.1056/aioa2300068] [Medline: 38343631]
92. Tozuka R, Johno H, Amakawa A, et al. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. Jpn J Radiol. Apr 2025;43(4):706-712. [doi: 10.1007/s11604-024-01705-1] [Medline: 39585559]
93. Hewitt KJ, Wiest IC, Carrero ZI, et al. Large language models as a diagnostic support tool in neuropathology. J Pathol Clin Res. Nov 2024;10(6):e70009. [doi: 10.1002/2056-4538.70009] [Medline: 39505569]
94. Wang D, Ye J, Li J, et al. Enhancing large language models for improved accuracy and safety in medical question answering: comparative study. JMIR Med Educ. Dec 2, 2025;11:e70190. [doi: 10.2196/70190] [Medline: 41329953]
95. Xu Y, Jia H, Wang M, et al. Enhancing clinical documentation with voice processing and large language models: a study on the LAOS system. NPJ Digit Med. Nov 28, 2025;8(1):798. [doi: 10.1038/s41746-025-02170-4] [Medline: 41315671]
96. Tung JYM, Le Q, Yao J, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: comparative study with junior clinicians. J Med Internet Res. Nov 19, 2025;27:e78393. [doi: 10.2196/78393] [Medline: 41259800]

97. Zhang C, Yang H, Liu X, et al. A knowledge-enhanced platform (MetaSepsisKnowHub) for retrieval augmented generation-based sepsis heterogeneity and personalized management: development study. J Med Internet Res. Jun 6, 2025;27:e67201. [doi: 10.2196/67201] [Medline: 40478618]
98. Fanelli F, Saleh M, Santamaria P, Zhurakivska K, Nibali L, Troiano G. Development and comparative evaluation of a reinstructed GPT-4o model specialized in periodontology. J Clin Periodontol. May 2025;52(5):707-716. [doi: 10.1111/jcpe.14101] [Medline: 39723544]
99. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med. Mar 4, 2025;8(1):137. [doi: 10.1038/s41746-025-01536-y] [Medline: 40038423]
100. Yang Q, Zuo H, Su R, et al. Dual retrieving and ranking medical large language model with retrieval augmented generation. Sci Rep. May 24, 2025;15(1):18062. [doi: 10.1038/s41598-025-00724-w] [Medline: 40413225]
101. Vach M, Gliem M, Weiss D, et al. Evaluating retrieval augmented generation-enhanced large language models for question answering on German neurovascular guidelines. Clin Neuroradiol. Mar 2026;36(1):119-127. [doi: 10.1007/s00062-025-01562-z] [Medline: 40892297]
102. Masanneck L, Epping PZ, Meuth SG, Pawlitzki M. Evaluating web retrieval-assisted large language models with and without whitelisting for evidence-based neurology: comparative study. J Med Internet Res. Oct 29, 2025;27:e79379. [doi: 10.2196/79379] [Medline: 41159599]
103. Fukui Y, Kawata Y, Kobashi K, Nagatani Y, Iguchi H. Evaluation of a retrieval-augmented generation system using a Japanese institutional nuclear medicine manual and large language model-automated scoring. Radiol Phys Technol. Sep 2025;18(3):861-876. [doi: 10.1007/s12194-025-00941-y] [Medline: 40683982]
104. Sha H, Gong F, Liu B, Liu R, Wang H, Wu T. Leveraging retrieval-augmented large language models for dietary recommendations with traditional chinese medicine's medicine food homology: algorithm development and validation. JMIR Med Inform. Aug 21, 2025;13:e75279. [doi: 10.2196/75279] [Medline: 40840437]
105. Owoyemi J, Abubakar S, Owoyemi A, et al. Open-source retrieval augmented generation framework for retrieving accurate medication insights from formularies for African healthcare workers. medRxiv. Preprint posted online on Feb 21, 2025. [doi: 10.1101/2025.02.20.25322640]
106. Tata V, Bouchamaoui Z, Bhaskara NV. OrthoGraphRAG: enhancing clinical decision making with multi-level knowledge graphs. ICML 2025 GenBio Workshop Poster (OpenReview). Jun 2025. URL: https://openreview.net/pdf?id=ht8uX6Pj0d [Accessed 2026-07-06]
107. Tayebi Arasteh S, Lotfinia M, Bressem K, et al. RadioRAG: online retrieval-augmented generation for radiology question answering. Radiol Artif Intell. Jul 2025;7(4):e240476. [doi: 10.1148/ryai.240476] [Medline: 40530957]
108. Das S, Ge Y, Guo Y, et al. Two-layer retrieval-augmented generation framework for low-resource medical question answering using Reddit data: proof-of-concept study. J Med Internet Res. Jan 6, 2025;27:e66220. [doi: 10.2196/66220] [Medline: 39761554]
109. Li H, Huang J, Ji M, Yang Y, An R. Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study. J Med Internet Res. Apr 30, 2025;27:e66098. [doi: 10.2196/66098] [Medline: 40306628]
110. Shin M, Song J, Kim MG, Yu HW, Choe EK, Chai YJ. Thyro-GenAI: a chatbot using retrieval-augmented generative models for personalized thyroid disease management. J Clin Med. Apr 3, 2025;14(7):2450. [doi: 10.3390/jcm14072450] [Medline: 40217905]
111. Aguzzi G, Magnini M, Farahmand A, Ferretti S, Pengo MF, Montagna S. RAG-enhanced open SLMs for hypertension management chatbots. J Med Syst. Nov 13, 2025;49(1):159. [doi: 10.1007/s10916-025-02297-7] [Medline: 41231304]
112. Zhou Q, Liu C, Duan Y, et al. GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. Front Med (Lausanne). 2024;11:1392555. [doi: 10.3389/fmed.2024.1392555] [Medline: 38841582]
113. Aminan M, Darnell SS, Delsoz M, et al. GlaucoRAG: a retrieval-augmented large language model for expert-level glaucoma assessment. medRxiv. Preprint posted online on Jul 7, 2025. [doi: 10.1101/2025.07.03.25330805] [Medline: 40672509]
114. Hsu HL, Dao CT, Wang L, et al. MedPlan: a two-stage RAG-based system for personalized medical plan generation. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track); Jul 28-30, 2025:1072-1082; Vienna, Austria. [doi: 10.18653/v1/2025.acl-industry.76]
115. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med. Apr 23, 2024;7(1):102. [doi: 10.1038/s41746-024-01091-y] [Medline: 38654102]
116. Kang B, Kim J, Yun TR, Kim CE. Prompt-RAG: pioneering vector embedding-free retrieval-augmented generation in niche domains, exemplified by korean medicine. arXiv. Preprint posted online on Jan 20, 2024. [doi: 10.48550/arXiv.2401.11246]
117. AlSammarraie A, Al-Saifi A, Kamhia H, Aboagla M, Househ M. Development and evaluation of an agentic LLM based RAG framework for evidence-based patient education. BMJ Health Care Inform. Jul 25, 2025;32(1):e101570. [doi: 10.1136/bmjhci-2025-101570] [Medline: 40713064]

118. Nandy G, Gupta S, Mohammad Afzali F, Peeples E, Pilon B, Tsai CH. Balancing health information-seeking through retrieval-augmented generation-based LLM chatbot. Presented at: UMAP '25; Jun 16-19, 2025; New York City, USA. [doi: 10.1145/3708319.3733709]
119. Hetz MJ, Carl N, Haggenmüller S, et al. Superhuman performance on urology board questions using an explainable language model enhanced with European Association of Urology guidelines. ESMO Real World Data Digit Oncol. Dec 2024;6:100078. [doi: 10.1016/j.esmorw.2024.100078] [Medline: 41646097]
120. 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. Dec 18, 2024;7(1):364. [doi: 10.1038/s41746-024-01391-3] [Medline: 39695316]
121. Carl N, Hetz MJ, Wies C, et al. Enhancing clinicians' trust in large language models via transparent source attribution: a randomized controlled evaluation in uro-oncology. Eur J Cancer. Jan 17, 2026;233:116168. [doi: 10.1016/j.ejca.2025.116168] [Medline: 41401634]
122. Tytler K. Adoption, usability and perceived clinical value of a UK AI clinical reference platform: a mixed-methods formative evaluation of real-world usage and a 1,223-respondent user survey. arXiv. Preprint posted online on Sep 25, 2025. [doi: 10.48550/arXiv.2509.21188]
123. Wu Y, Chen X, Zhang W, et al. ChatMyopia: an AI agent for myopia-related consultation in primary eye care settings. iScience. Nov 2025;28(11):113768. [doi: 10.1016/j.isci.2025.113768] [Medline: 41244561]
124. Kim S. MedBioRAG: semantic search and retrieval-augmented generation with large language models for medical and biological QA. arXiv. Preprint posted online on Dec 10, 2025. [doi: 10.48550/arXiv.2512.10996]
125. Hasan M, Hossain MA, Sayem FH, et al. CLIN-LLM: a safety-constrained hybrid framework for clinical diagnosis and treatment generation. arXiv. Preprint posted online on Oct 26, 2025. [doi: 10.48550/arXiv.2510.22609]
126. Long Y, Yang C, Tang G, et al. KidneyTalk-open: no-code deployment of a private large language model with medical documentation-enhanced knowledge database for kidney disease. arXiv. Preprint posted online on Mar 6, 2025. [doi: 10.48550/arXiv.2503.04153]
127. Zhang J, Song J, Tu W. From evidence-based medicine to knowledge graph: retrieval-augmented generation for sports rehabilitation and a domain benchmark. arXiv. Preprint posted online on Jan 1, 2026. [doi: 10.48550/arXiv.2601.00216]
128. Ryan J, Gumilang AI, Wiliam R, Suhartono D. Self-MedRAG: a self-reflective hybrid retrieval-augmented generation framework for reliable medical question answering. arXiv. Preprint posted online on Jan 8, 2026. [doi: 10.48550/arXiv.2601.04531]
129. Yang W, Xue H, Peng Q, Hu H, Huang Q, Zhang T. Making medical vision-language models think causally across modalities with retrieval-augmented cross-modal reasoning. Presented at: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 3-8, 2026; Barcelona, Spain. [doi: 10.1109/ICASSP55912.2026.11462104]
130. Lorenzo L, Montana-Mendez M, Figueiras S, Boubeta B, Bernardo-Castineira C. Evaluation of oncotimia: an LLM based system for supporting tumour boards. arXiv. Preprint posted online on Jan 27, 2026. [doi: 10.48550/arXiv.2601.19899]
131. Li Z, Xu J, Hu Z, et al. MedCoRAG: interpretable hepatology diagnosis via hybrid evidence retrieval and multispecialty consensus. arXiv. Preprint posted online on Mar 5, 2026. [doi: 10.48550/arXiv.2603.05129]
132. Samanta HS. Grounded multimodal retrieval-augmented drafting of radiology impressions using case-based similarity search. arXiv. Preprint posted online on Mar 18, 2026. [doi: 10.48550/arXiv.2603.17765]
133. Chan RWC, Lin S, Ma YN, Chen H, Jiang L, Fan W. PriHA: a RAG-enhanced LLM framework for primary healthcare assistant in Hong Kong. arXiv. Preprint posted online on Apr 10, 2026. [doi: 10.48550/arXiv.2604.14215]
134. Chen X, Shi B, Le C, et al. Iterative multimodal retrieval-augmented generation for medical question answering. arXiv. Preprint posted online on Apr 30, 2026. [doi: 10.48550/arXiv.2604.27724]
135. Khosa T, Daramola O. Development and preliminary evaluation of a domain-specific large language model for tuberculosis care in South Africa. arXiv. Preprint posted online on Mar 28, 2026. [doi: 10.48550/arXiv.2604.19776]
136. Akbar AR, Wales-McGrath S, Levya A, et al. PathoScribe: transforming pathology data into a living library with a unified LLM-driven framework for semantic retrieval and clinical integration. arXiv. Preprint posted online on Mar 8, 2026. [doi: 10.48550/arXiv.2603.08935]
137. Li J, Chang Y, Liu Y, et al. TCM-DiffRAG: personalized syndrome differentiation reasoning method for traditional Chinese medicine based on knowledge graph and chain of thought. Front Med (Lausanne). Apr 21, 2026;13:1804478. [doi: 10.3389/fmed.2026.1804478]
138. Li H, Shi J, Chen X, et al. Large language model aided Birt-Hogg-Dube syndrome diagnosis with multimodal retrieval-augmented generation. arXiv. Preprint posted online on Nov 25, 2025. [doi: 10.48550/arXiv.2511.19834]
139. Chen Z, Liao Y, Zhu Z, et al. HeteroRAG: a heterogeneous retrieval-augmented generation framework for medical vision language tasks. Presented at: Findings of the Association for Computational Linguistics: ACL 2026; Jul 2-7, 2026; San Diego, California, United States. [doi: 10.18653/v1/2026.findings-acl.176]

140. Li W, Zhang H, Zhang H, et al. Refine medical diagnosis using generation augmented retrieval and clinical practice guidelines. IEEE J Biomed Health Inform. Dec 9, 2025;PP:1-14. [doi: 10.1109/JBHI.2025.3641931] [Medline: 41364573]
141. Ansari MS, Khan MSA, Revankar S, Varma A, Mokhade AS. Lightweight clinical decision support system using QLoRA-fine-tuned LLMs and retrieval-augmented generation. arXiv. Preprint posted online on May 6, 2025. [doi: 10.48550/arXiv.2505.03406]
142. Xia P, Zhu K, Li H, et al. MMed-RAG: versatile multimodal RAG system for medical vision language models. arXiv. Preprint posted online on Oct 16, 2024. [doi: 10.48550/arXiv.2410.13085]
143. Ning Y, Sun Y, Luo L, Wang Y, Pan Y, Lin H. MedTrust-RAG: evidence verification and trust alignment for biomedical question answering. Presented at: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 15-18, 2025; Wuhan, China. [doi: 10.1109/BIBM66473.2025.11356290]
144. Kazemzadeh H, Dizaji KM, Tavakoli SR, et al. DrugRAG: enhancing pharmacy LLM performance through a novel retrieval-augmented generation pipeline. arXiv. Preprint posted online on Dec 16, 2025. [doi: 10.48550/arXiv.2512.14896]
145. Chen Y, Li F, Song X, et al. Exploring the role of knowledge graph-based RAG in Japanese medical question answering with small-scale LLMs. arXiv. Preprint posted online on Apr 15, 2025. [doi: 10.48550/arXiv.2504.10982]
146. Han Z, Ge J, Li C. Knowledge-guided large language model for automatic pediatric dental record understanding and safe antibiotic recommendation. arXiv. Preprint posted online on Dec 9, 2025. [doi: 10.48550/arXiv.2512.09127]
147. Keerthana G, Gupta M. CLI-RAG: a retrieval-augmented framework for clinically structured and context aware text generation with LLMs. arXiv. Preprint posted online on Jul 9, 2025. [doi: 10.48550/arXiv.2507.06715]
148. Guo Z, Lai A, Ive J, et al. Development and evaluation of HopeBot: an LLM-based chatbot for structured and interactive PHQ-9 depression screening. arXiv. Preprint posted online on Jul 8, 2025. [doi: 10.48550/arXiv.2507.05984]
149. Zhu Y, Guo J, Jiang J, Gu P, Shu X, Chen D. Explainable interictal epileptiform discharge detection method based on scalp EEG and retrieval-augmented generation. arXiv. Preprint posted online on Feb 15, 2026. [doi: 10.48550/arXiv.2602.14170]
150. Mo G, Raman NJ, Chai M, et al. PeerCoPilot: a language model-powered assistant for behavioral health organizations. Presented at: AAAI'26: AAAI Conference on Artificial Intelligence; Jan 20-27, 2026; Singapore. [doi: 10.1609/aaai.v40i47.41476]
151. Shahnawaz A, Shafique A, Wang D, Mustafa M. Designing around stigma: human-centered LLMs for menstrual health. Presented at: CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems; Apr 13-17, 2026; Barcelona, Spain. URL: https://dl.acm.org/doi/proceedings/10.1145/3772318 [Accessed 2026-07-06] [doi: 10.1145/3772318.3791318]
152. Ahalpara TJ. Tell Me: an LLM-powered mental well-being assistant with RAG, synthetic dialogue generation, and agentic planning. arXiv. Preprint posted online on Nov 18, 2025. [doi: 10.48550/arXiv.2511.14445]
153. Boumans R, Cramer L, van de Poll S, Vermeulen H. A feasibility study on usability and trust among population groups of a medical avatar supported by large language models with retrieval augmented generation. arXiv. Preprint posted online on Oct 17, 2025. [doi: 10.48550/arXiv.2510.15531]
154. Parmanto B, Aryoyudanta B, Soekinto TW, et al. A reliable and accessible caregiving language model (CaLM) to support tools for caregivers: development and evaluation study. JMIR Form Res. Jul 31, 2024;8:e54633. [doi: 10.2196/54633] [Medline: 39083337]
155. Hang CN, Yu PD, Tan CW. TrumorGPT: graph-based retrieval-augmented large language model for fact-checking. IEEE Trans Artif Intell. 2025;6(11):3148-3162. [doi: 10.1109/TAI.2025.3567369]
156. Gu D, Gao Y, Zhou Y, Zhou M, Metaxas D. RadAlign: advancing radiology report generation with vision-language concept alignment. Presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025; Sep 23-27, 2025; Daejeon, South Korea. [doi: 10.1007/978-3-032-04981-0_46]
157. Yi Z, Xiao T, Albert MV. A multimodal multi-agent framework for radiology report generation. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2505.09787]
158. Bang B, Yoon J, Chang DJ, Park S, Lee YO. Retrieval augmented large language model system for comprehensive drug contraindications. Health Inf Sci Syst. Jan 11, 2026;14(1):26. [doi: 10.1007/s13755-025-00420-z] [Medline: 41531551]
159. Ting LPY, Zhao C, Zeng YH, Lim YJ, Chuang KT, Liu H. Leaps beyond the seen: reinforced reasoning augmented generation for clinical notes. arXiv. Preprint posted online on Jun 3, 2025. [doi: 10.48550/arXiv.2506.05386]
160. Johno H, Johno Y, Amakawa A, et al. Enhancing pancreatic cancer staging with large language models: the role of retrieval-augmented generation. Radiol Phys Technol. Jun 2026;19(2):593-603. [doi: 10.1007/s12194-026-01026-0] [Medline: 41784908]
161. He J, Guo Y, Lam LK, et al. OpenTCM: a GraphRAG-empowered LLM-based system for traditional chinese medicine knowledge retrieval and diagnosis. arXiv. Preprint posted online on Apr 28, 2025. [doi: 10.48550/arXiv.2504.20118]
162. Zhao YF, Bove A, Thompson D, et al. Generative AI Is not ready for clinical use in patient education for lower back pain patients, even with retrieval-augmented generation. AMIA Jt Summits Transl Sci Proc. 2025;2025:644-653. [Medline: 40502233]

163. Madrid-García A, Benavent D, Plasencia-Rodríguez C, Rosales-Rosado Z, Merino-Barbancho B, Freites-Núñez D. Optimising the clinical application of rheumatology guidelines using large language models: a retrieval-augmented generation framework integrating EULAR and ACR recommendations. EULAR Rheumatol Open. Oct 2025;1(3):228-236. [doi: 10.1016/j.ero.2025.08.001] [Medline: 42368617]
164. Saidu F, Wall J. Retrieval-augmented large language model for clinical decision support with a medical knowledge graph. Electronics (Basel). 2026;15(3):555. [doi: 10.3390/electronics15030555]
165. Felde S, Buchkremer R, Chehab G, et al. Low-energy small language models with retrieval-augmented generation can surpass large-model performance in rheumatology. Front Med (Lausanne). May 8, 2026;13:1817215. [doi: 10.3389/fmed.2026.1817215] [Medline: 42180760]
166. Jeon Y, Youn MS, Kang S, et al. Hierarchical RAG enhances a pharmacogenomic AI assistant in guideline related queries. Comput Biol Med. Jan 1, 2026;200:111323. [doi: 10.1016/j.compbiomed.2025.111323] [Medline: 41319469]
167. Kang D, Zhao K, Cheng D, Yuan L, Sun W, Jin K. Evaluation of large language models and retrieval-augmented generation for clinical reasoning in pediatric myopia: a 50-case real-world study. Sci Rep. May 7, 2026;16(1):21059. [doi: 10.1038/s41598-026-51205-7] [Medline: 42098248]
168. Komenda A, Makowski M, Can E, et al. Development and evaluation of a retrieval-augmented generation system for radiology guidelines. J Imaging Inform Med. Feb 12, 2026. [doi: 10.1007/s10278-025-01835-6] [Medline: 41680573]
169. Wang Y, Luan Y, Cheng S, et al. A multi-layer retrieval-augmented large language model framework for enhancing hypertension education. Hypertens Res. Apr 2026;49(4):1428-1440. [doi: 10.1038/s41440-025-02481-9] [Medline: 41501362]
170. Zhang S, Phan E, Velmovitsky P, Pham Q, Sanner S. Retrieval-augmented generation for medical question answering on a heart failure dataset: performance analysis. JMIR Form Res. Feb 26, 2026;10:e84932. [doi: 10.2196/84932] [Medline: 41747226]
171. Baseri Saadi S, Ver Berne J, Cavalcante Fontenele R, Claes P, Jacobs R. JADE: jawbone lesion diagnosis and decision supporting system. Dentomaxillofac Radiol. Jul 1, 2026;55(5):497-507. [doi: 10.1093/dmfr/twag017] [Medline: 41872013]
172. Song JK, Youk DB, Kim H, Hwang SH. Multimodal knowledge graph-guided RAG-LLM for clinical decision support in pediatric leukemia. Cancer Res Treat. Apr 21, 2026. [doi: 10.4143/crt.2026.0047] [Medline: 42025216]
173. Kabak Y, Erturkmen GBL, Gencturk M. FHIR-RAG-MEDS: integrating HL7 FHIR with retrieval-augmented large language models for enhanced medical decision support. arXiv. Preprint posted online on Sep 9, 2025. [doi: 10.48550/arXiv.2509.07706]
174. Ma J, Yue X, Chen Y, Shi J, Wei Z, Jia Z. MGK-RAG: multi-granularity knowledge guided retrieval-augmented generation for radiology report. Presented at: WWW '26; Apr 13-17, 2026; Dubai, United Arab Emirates. [doi: 10.1145/3774904.3792924]
175. Yang S, Kim KM, Kim M. CPR-RAG: clinical prior-regularized retrieval for anatomy-aware 3D CT report generation. Presented at: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 2-7, 2026; San Diego, California, United States. [doi: 10.18653/v1/2026.acl-long.1411]
176. Garcia-Font M, Dufey-Portilla N, Durán-Sindreu F, et al. Evaluating retrieval-augmented large language models on external cervical resorption: a comparative study of Gemini and NotebookLM. J Endod. Feb 2026;52(2):300-306. [doi: 10.1016/j.joen.2025.10.016] [Medline: 41207474]
177. Wong HS, Wong TK. Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage: retrospective evaluation study. JMIR Med Inform. Jan 26, 2026;14:e82026. [doi: 10.2196/82026] [Medline: 41587455]
178. Thio S, Lewis M, Denaxas S, Dobson RJB. Unlocking electronic health records: a hybrid graph RAG approach to safe clinical AI for patient QA. Front Digit Health. 2026;8:1780700. [doi: 10.3389/fdgth.2026.1780700] [Medline: 41890309]
179. Zaki HA, Brea A, Parvataneni K, et al. Retrieval-augmented language models for patient-centered periprocedural anticoagulation in interventional radiology. Cardiovasc Intervent Radiol. Jun 2026;49(6):1177-1187. [doi: 10.1007/s00270-026-04428-0] [Medline: 41917168]
180. Liu H, Hu Y, Li D, et al. LLM-driven collaborative framework for knowledge-enhanced cancer pain assessment and management. NPJ Digit Med. Jan 19, 2026;9(1):180. [doi: 10.1038/s41746-026-02362-6] [Medline: 41554973]
181. Kim D, Yoo S, Jeong O. MedSumGraph: enhancing GraphRAG for medical QA with summarization and optimized prompts. Artif Intell Med. Feb 2026;172:103311. [doi: 10.1016/j.artmed.2025.103311] [Medline: 41319399]
182. Lopez I, Swaminathan A, Vedula K, et al. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. Jan 19, 2025;8(1):45. [doi: 10.1038/s41746-024-01377-1] [Medline: 39828800]
183. Nanua S, Steward R, Neely B, Datto M, Youens K. Retrieval-augmented generation for interpreting clinical laboratory regulations using large language models. J Pathol Inform. Nov 2025;19:100520. [doi: 10.1016/j.jpi.2025.100520] [Medline: 41244595]
184. Xie W, Song X, Lu Z, et al. Retrieval-augmented large language models for depression screening and suicide risk stratification. BMC Psychiatry. Mar 31, 2026;26(1):386. [doi: 10.1186/s12888-026-07988-0] [Medline: 41917861]

185. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. Mar 31, 2025;43(2):1-55. [doi: 10.1145/3703155]
186. Wallat J, Heuss M, de Rijke M, Anand A. Correctness is not faithfulness in retrieval augmented generation attributions. Presented at: ICTIR '25: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR); Jul 18, 2025:22-32; Padua, Italy. [doi: 10.1145/3731120.3744592]
187. Min S, Krishna K, Lyu X, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023:12076-12100; Singapore. [doi: 10.18653/v1/2023.emnlp-main.741]
188. Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA. All that's 'human' is not gold: evaluating human evaluation of generated text. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Aug 1-6, 2021:7282-7296; Online. [doi: 10.18653/v1/2021.acl-long.565]
189. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Presented at: 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks; Dec 10-16, 2023; New Orleans, Louisiana, USA. URL: https://www.proceedings.com/content/075/075280-2020open.pdf [Accessed 2026-07-29] [doi: 10.52202/075280-2020]
190. Razavi A, Soltangheis M, Arabzadeh N, Salamat S, Zihayat M, Bagheri E. Benchmarking prompt sensitivity in large language models. Presented at: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025; Apr 6-10, 2025; Lucca, Italy. [doi: 10.1007/978-3-031-88714-7_29]
191. Zhao Y, Liu H, Yu D, Kung S, Chen M, Mi H, et al. One token to fool LLM-as-a-judge. arXiv. Preprint posted online on Jul 11, 2025. [doi: 10.48550/arXiv.2507.08794]
192. Li H, Dong Q, Chen J, et al. LLMs-as-judges: a comprehensive survey on LLM-based evaluation methods. arXiv. Preprint posted online on Dec 7, 2024. [doi: 10.48550/arXiv.2412.05579]
193. Wang H, Prasad A, Stengel-Eskin E, Bansal M. Retrieval-augmented generation with conflicting evidence. arXiv. Preprint posted online on Apr 17, 2025. [doi: 10.48550/arXiv.2504.13079]
194. Han T, Kumar A, Agarwal C, Lakkaraju H. MedSafetyBench: evaluating and improving the medical safety of large language models. Presented at: NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems; Dec 10-15, 2024; Vancouver, BC, Canada. [doi: 10.52202/079017-1054]
195. Savage T, Wang J, Gallo R, et al. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. medRxiv. Preprint posted online on Jun 10, 2024. [doi: 10.1101/2024.06.06.24308399]
196. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. Oct 2010;19 Suppl 3(Suppl 3):i68-74. [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
197. Barnett S, Kurniawan S, Thudumu S, Brannelly Z, Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system. Presented at: CAIN 2024; Apr 14-15, 2024; Lisbon, Portugal. [doi: 10.1145/3644815.3644945]
198. van der Vorst JP, Smit JM, van de Sande D, et al. Importance of model governance in clinical AI models: case study on the relevance of data drift detection. BMJ Digit Health. Jul 2025;1(1):e000046. [doi: 10.1136/bmjdhai-2025-000046]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
GraphRAG: graph-structured retrieval-augmented generation
IRR: interrater reliability
LLM: large language model
NLP: natural language processing
OSF: Open Science Framework
PCC: Population, Concept, and Context
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAG: retrieval-augmented generation

Edited by Stefano Brini; peer-reviewed by Dillon Chrimes, Mark Meleka, Mladen Borovic; submitted 20.Dec.2025; final revised version received 22.May.2026; accepted 01.Jun.2026; published 03.Aug.2026

Please cite as:
Zhao Y, Miao Y, Guo R, Luo Y, Wang H, Wu Y

Evaluation Methods for Inference-Time Retrieval-Augmented and Graph Retrieval-Augmented Large Language Models in Health Care: Scoping Review
J Med Internet Res 2026;28:e90046
URL: https://www.jmir.org/2026/1/e90046
doi: 10.2196/90046