

Retrieval-Augmented Large Language Models for Clinical Decision Support: A Systematic Review of Hallucination Mitigation and Evidence Grounding

Sumit Barua

`sumit.barua@wmich.edu`

Western Michigan University <https://orcid.org/0009-0003-9775-7009>

Charles Barnabas Rodgers

Western Michigan University

Systematic Review

Keywords: Retrieval-augmented generation (RAG), Large language models, Clinical decision support systems, Hallucination mitigation, Evidence grounding, Patient safety, Clinical artificial intelligence, Systematic review

Posted Date: May 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9741159/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background: Large language models (LLMs) are increasingly explored for clinical decision support systems (CDSS), yet hallucinations—plausible but factually unsupported outputs—pose direct patient safety risks. Retrieval-augmented generation (RAG), which conditions LLM outputs on external evidence at inference time, has emerged as a proposed mitigation strategy. However, the extent to which RAG improves clinical safety and rigor in evaluation remains unclear.

Objective: To systematically examine architectural patterns of RAG in clinical CDSS, evaluate empirical evidence for hallucination mitigation, and assess whether current evaluation practices align with clinically meaningful safety standards.

Methods: Following PRISMA guidelines, we reviewed studies published between January 2023 and December 2025 evaluating retrieval-augmented LLM systems in clinical contexts. Inclusion required inference-time retrieval, empirical evaluation, and assessment of correctness, grounding, or safety. Data were extracted on architectural design, knowledge sources, evaluation methodology, and validation practices. Owing to heterogeneity in clinical tasks and reported metrics, findings were synthesized narratively.

Results: Thirty studies met the inclusion criteria. Four architectural patterns were identified: single-step retrieval, iterative or agentic retrieval, graph-based retrieval, and hybrid symbolic–neural systems. Across comparative evaluations, RAG approaches were associated with improved factual grounding and reduced unsupported claims relative to non-retrieval LLMs, particularly in guideline- and protocol-based tasks. However, evaluation practices were predominantly retrospective and metric-driven, with limited clinician involvement and nopro prospective

clinical outcome studies. Recurrent failure modes included retrieval errors, incomplete or selective integration of evidence, and residual hallucinations despite retrieved context.

Conclusions: RAG architectures represent a substantive advance in improving evidentiary grounding of clinical language models under controlled conditions. Nevertheless, the current evidence base reflects architectural innovation without parallel maturation in clinical validation. Standardized safety-oriented evaluation frameworks, prospective validation, and regulatory-aligned governance mechanisms are necessary before RAG-based systems can be considered ready for high-stakes clinical deployment.

1 Introduction

1.1 Clinical Safety Motivation

The integration of artificial intelligence into clinical workflows represents a consequential transformation in healthcare delivery. Large language models (LLMs) have demonstrated promising performance in medical question answering, documentation support, and structured reasoning tasks [1, 2]. Yet their

clinical adoption remains constrained by a critical limitation: the generation of plausible but factually unsupported content, commonly referred to as hallucination [3, 4].

In healthcare settings, hallucination constitutes a safety risk rather than a benign technical artifact. Incorrect medication dosing, inappropriate guideline interpretation, or unsupported diagnostic reasoning may influence clinical decision-making with potentially harmful consequences [5, 6]. Linguistically coherent but evidentially unsupported outputs are particularly concerning because they may evade superficial detection while undermining trust, traceability, and accountability requirements central to clinical practice [7, 8]. Empirical evaluations of baseline LLMs across radiology, guideline assessment, and medical QA domains have documented unsupported claims, citation inconsistencies, and reasoning errors [3, 7, 9]. These findings underscore the need for architectures that improve evidentiary grounding without sacrificing expressive capability.

1.2 Retrieval-Augmented Generation as a Mitigation Strategy

Retrieval-augmented generation (RAG) has emerged as a leading architectural approach for mitigating hallucinations in clinical language models. RAG systems incorporate external knowledge retrieval at inference time, conditioning model outputs on curated sources such as clinical practice guidelines, biomedical literature, drug databases, or institutional protocols [2, 10, 11]. By explicitly grounding responses in retrieved evidence, RAG aims to reduce unsupported generation while preserving natural language flexibility [12].

Conceptually, RAG offers several advantages. It enables linkage between generated outputs and authoritative sources, facilitating verification and citation [13]. It allows knowledge bases to be updated without retraining model parameters, an important property in rapidly evolving medical domains [14]. It also supports the incorporation of domain- or institution-specific knowledge not fully represented in general training corpora [11, 15].

However, retrieval augmentation does not eliminate hallucination risk. Retrieval systems may fail to identify relevant evidence, retrieve outdated or incomplete information, or surface conflicting sources [9, 16]. Even when relevant context is retrieved, models may inconsistently integrate or selectively interpret evidence [2, 14]. Multi-source synthesis, contradiction resolution, and alignment between parametric knowledge and retrieved content remain open challenges [17]. Thus, RAG redistributes rather than removes failure risk.

1.3 Limitations of Existing Reviews

Although prior reviews have examined LLM applications in healthcare or RAG architectures more broadly, systematic evaluation of RAG-specific hallucination mitigation in clinical decision support remains limited. Many surveys aggregate heterogeneous LLM applications without isolating retrieval-augmented designs or interrogating grounding mechanisms in depth [18, 19]. Others discuss RAG in general-purpose contexts where safety requirements differ substantially from those in clinical care [20]. Critically,

existing reviews rarely evaluate whether commonly reported performance metrics meaningfully capture clinical risk. Metrics such as ROUGE, BLEU, or semantic similarity may not correspond to diagnostic accuracy, treatment appropriateness, or patient safety outcomes [21]. Furthermore, systematic cross-study synthesis of evaluation rigor, clinician involvement, prospective validation, and deployment readiness remains limited [22, 23]. Recurring failure modes—including retrieval errors, conflict-resolution failures, and residual hallucinations despite grounding—have not

been comprehensively synthesized.

1.4 Objectives and Contributions

To address these gaps, this systematic review synthesizes empirical evidence on retrieval-augmented LLM systems designed for clinical decision support. We structure our analysis around six research questions:

1. **Architectural Patterns:** What retrieval architectures, knowledge sources, and integration strategies are employed in clinical RAG systems?
2. **Effectiveness Evidence:** What empirical evidence supports RAG’s effectiveness in improving factual grounding or reducing unsupported claims?
3. **Evaluation Methodologies:** How are RAG systems evaluated for hallucination mitigation and clinical safety?
4. **Failure Modes:** What recurrent failure patterns are reported, and what are their safety implications?
5. **Deployment Readiness:** What evidence exists regarding workflow integration, regulatory alignment, and real-world validation?
6. **Future Directions:** What methodological, technical, and governance priorities emerge for advancing safe clinical deployment?

Our contributions include:

1. A structured synthesis of architectural patterns and grounding mechanisms in clinical RAG systems
2. Critical evaluation of methodological rigor and safety alignment
3. Systematic identification of recurring failure modes
4. Development of an integrative conceptual framework linking architecture, evaluation methodology, and deployment constraints

To guide interpretation, we introduce a Clinical RAG Hallucination Mitigation Framework (Fig. 1), conceptualizing retrieval-augmented decision support as a multi-stage process spanning retrieval design, evidence integration, output generation, evaluation methodology, and deployment context. This framework highlights the persistent gap between NLP-oriented performance metrics and clinically meaningful harm assessment.

2 Methods

2.1 Review Design

This study was conducted as a systematic review with an embedded scoping analysis in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The systematic component focused on identifying, screening, and synthesizing empirical studies evaluating retrieval-augmented generation (RAG) systems in clinical decision support contexts.

Given the rapid evolution and architectural heterogeneity of clinical RAG systems, a scoping element was incorporated to characterize emerging design patterns, evaluation methodologies, failure modes, and deployment considerations that were not amenable to quantitative pooling. The review, therefore, integrates structured, systematic identification with qualitative thematic synthesis.

2.2 Search Strategy

2.2.1 Databases and Sources

A comprehensive literature search was conducted in two structured bibliographic databases:

- PubMed/MEDLINE
- Scopus

PubMed/MEDLINE was selected to capture biomedical and clinical literature, while Scopus was selected to provide broader interdisciplinary coverage of computer science, artificial intelligence, medical informatics, and conference-based publications relevant to retrieval-augmented generation systems.

2.2.2 Search Period and Date

Studies published between January 1, 2023, and December 31, 2025, were eligible. The final search was conducted on January 22, 2026. No additional updates were performed after this date.

2.2.3 Language Restrictions

Only studies published in English were included due to feasibility constraints and to ensure consistent evaluation reporting.

2.2.4 Search Terms

The search strategy combined four concept domains using Boolean operators:

1. LLM concepts: ("large language model*" OR "LLM" OR "GPT" OR "generative AI" OR "foundation model*" OR "transformer model*")
2. Retrieval concepts: ("retrieval augmented" OR "retrieval-augmented" OR "RAG" OR "knowledge retrieval" OR "evidence grounding" OR "document retrieval")

3. Clinical concepts: ("clinical decision support" OR "medical" OR "healthcare" OR "clinical" OR "diagnosis" OR "treatment recommendation" OR "patient care")
4. Hallucination/safety concepts: ("hallucination*" OR "factual accuracy" OR "factuality" OR "reliability" OR "safety" OR "evidence grounding")

A representative Scopus query was:

("large language model*" OR "LLM" OR "GPT" OR "generative AI") AND ("retrieval-augmented" OR "RAG" OR "evidence grounding") AND ("clinical decision support" OR "healthcare" OR "medical") AND ("hallucination*" OR "factual accuracy" OR "reliability").

The search syntax was adapted to meet database-specific requirements. Full search strings for each database are provided in the supplementary material.

2.3 Eligibility Criteria

2.3.1 Inclusion Criteria

Studies were included if they met all of the following:

1. Employed retrieval-augmented generation or equivalent inference-time retrieval grounding mechanisms.
2. Focused on clinical, medical, or healthcare decision-support contexts.
3. Reported empirical evaluation of factual accuracy, hallucination mitigation, evidence grounding, or clinical reliability.
4. Presented original research (experimental evaluation, comparative study, or system development with evaluation).
5. Published between January 2023 and December 2025.
6. Available in full text with sufficient methodological detail.

2.3.2 Exclusion Criteria

Studies were excluded if they:

1. Addressed non-medical domains.
2. Did not include an explicit retrieval augmentation component.
3. Were reviews, editorials, commentaries, or perspective articles without empirical evaluation.
4. Lacked evaluation of factuality, grounding, or safety-related outcomes.
5. Provided insufficient methodological detail for appraisal.

2.4 Study Selection Process

Study selection followed a three-stage process consistent with PRISMA guidelines.

Identification and Deduplication

A total of 463 records were identified through database searching across PubMed and Scopus. After removal of 121 duplicate records, 342 unique records remained for title and abstract screening.

Title and Abstract Screening

Titles and abstracts were screened against predefined eligibility criteria. Screening was conducted independently by two reviewers. Discrepancies were resolved through discussion and consensus. Forty-seven records

were excluded at this stage for predefined reasons (non-clinical domain, absence of a retrieval component, review articles, or lack of hallucination-related evaluation).

Full-Text Review

Both reviewers assessed 30 full-text articles for eligibility. All met the inclusion criteria and were included in the qualitative synthesis.

A PRISMA flow diagram is provided in Fig. 2.

2.5 Data Extraction

Data extraction was performed independently by two reviewers using a predefined structured extraction form. The form was piloted on five studies and refined prior to full extraction.

For each study, the following domains were recorded:

1. Study characteristics (year, venue, clinical domain, study design).
2. RAG architectural design (retrieval mechanism, knowledge source, integration strategy).
3. Hallucination mitigation techniques (citation generation, verification mechanisms, iterative retrieval).
4. Evaluation methodology (datasets, metrics, baseline comparisons, clinician involvement).
5. Reported outcomes (comparative performance, failure modes, limitations).
6. Clinical safety and deployment considerations (error severity assessment, deployment status, regulatory discussion).

Disagreements were resolved through discussion. No automation tools were used for data extraction.

2.6 Quality and Rigor Appraisal

Given the absence of a validated appraisal tool tailored specifically to retrieval-augmented LLM systems in clinical contexts, we employed a structured domain-based appraisal framework adapted from established principles in clinical AI evaluation literature.

Each study was assessed across five domains:

1. Methodological rigor
2. Evaluation robustness
3. Clinical relevance
4. Safety and risk consideration
5. Transparency and reporting completeness

Studies were categorized as high, medium, or low quality based on overall performance across domains. Quality classification informed interpretive weighting during narrative synthesis but was not used as an exclusion criterion.

2.7 Data Synthesis

Due to heterogeneity in clinical tasks, evaluation metrics, and outcome reporting, quantitative meta-analysis was not feasible. Findings were synthesized narratively, structured around predefined research questions. Architectural patterns, evaluation methodologies, and failure modes were identified through thematic grouping of extracted data.

Where quantitative outcomes were reported, results were described descriptively without statistical pooling.

3 Results

3.1 Study Characteristics

The literature search yielded 106 records across PubMed, Google Scholar, and Scopus. After deduplication, 77 unique records were screened, and 30 studies met the inclusion criteria following title/abstract screening and full-text assessment (Fig. 2). Included studies were published between 2023 and 2025, reflecting the rapid emergence of retrieval-augmented generation (RAG) for clinical decision support systems (CDSS) in the post-LLM era.

3.1.1 Publication and geographic profile

Most included studies were disseminated as preprints or technical reports, with a smaller proportion appearing in peer-reviewed journals or conference proceedings. Studies originated primarily from the United States, followed by multinational collaborations and single-country studies from Europe, Asia, and Australia.

3.1.2 Clinical domains and task types

The most common application category was general medical question answering, followed by specialty-specific CDSS, including radiology, nursing care, oncology, and other subspecialties. Across domains, primary task types included clinical question answering, diagnostic support, treatment recommendation, guideline concordance assessment, and clinical documentation or summarization.

3.1.3 Study design and evaluation scale

Most studies used comparative evaluations, typically contrasting RAG-augmented systems against non-retrieval LLM baselines or alternative retrieval configurations. A smaller subset reported system development with empirical evaluation but without formal baseline comparison. Evaluation scales varied substantially, ranging from expert-curated test sets to larger benchmark-style datasets. Prospective evaluation within real clinical workflows was not observed, and no study reported longitudinal clinical outcome assessment.

3.1.4 Base model families

GPT-3.5/GPT-4 were frequently used as base models, alongside open-source families (e.g., LLaMA variants), domain-adapted medical LLMs, and other proprietary systems. Several studies evaluated multiple base models to test whether retrieval-related gains generalized across model families.

3.1.5 Reporting Characteristics and Quantitative Outcome Availability

Table 1 summarizes reporting transparency across included studies. Reporting variables were extracted using a predefined, structured framework applied to published

abstracts and verified against the full texts for studies included in the comparative quantitative synthesis.

Table 1
Reporting transparency and quantitative outcome availability across included studies (n = 30).

Reporting Variable	Studies Reporting (n = 30)
Quantitative effectiveness metrics (numeric)	21
Hallucination-specific quantitative metrics	4
Sample size or evaluation denominator reported	15
Statistical significance testing (p-values or inferential tests)	11
Expert clinician evaluation described	5
Prospective real-world evaluation	0
Regulatory clearance reported	0

Most studies (21/30, 70%) reported at least one numeric effectiveness metric. However, only 11 studies (37%) reported inferential statistical testing, and fewer explicitly quantified hallucination rates (4/30,

13%). Sample size or evaluation denominators were reported in 15 studies (50%), and clinician-based expert evaluation was described in 5 studies (17%).

Despite frequent reporting of numeric performance metrics, outcome definitions, denominators, and evaluation protocols were highly heterogeneous. Metrics ranged from automated similarity scores (e.g., ROUGE, BLEU, BERTScore) to clinician-rated factuality, guideline concordance, and task-specific accuracy measures. Paired baseline–RAG comparisons were inconsistently reported, and effect sizes were not standardized across studies.

Due to substantial heterogeneity in outcome definitions, evaluation designs, and reporting granularity, statistical pooling or formal meta-analysis was not performed. Findings were synthesized narratively and structured for comparison across architectural features, evaluation contexts, and reported performance patterns.

3.2 RAG Architectural Patterns for Hallucination Mitigation (RQ1)

Across included studies, clinical retrieval-augmented generation (RAG) systems exhibited recurring architectural patterns that influence hallucination risk at three points:

(i) how evidence is retrieved, (ii) which sources are treated as authoritative, and (iii) how retrieved content is enforced during generation. These architectural choices were not uniformly validated through formal ablation studies; however, they were consistently described as mechanisms intended to reduce unsupported generation or improve evidence grounding.

3.2.1 Retrieval paradigms

Three broad retrieval paradigms were observed.

Dense retrieval (vector similarity using general or biomedical embeddings) was most common for capturing semantic synonymy and concept-level matching in medical language [2, 3, 8, 10, 24].

Sparse retrieval (e.g., BM25/TF–IDF) appeared in systems prioritizing interpretability, efficiency, and traceability in safety- or compliance-oriented settings [11, 15, 20].

Hybrid retrieval combined sparse term matching with dense semantic retrieval to improve robustness on complex queries where both exact terminology and conceptual relevance are needed [9, 16, 17, 25].

3.2.2 Knowledge source strategies

Knowledge sources are clustered into four practical categories.

Clinical practice guidelines were frequently used as foundations for decision support and guideline-concordant recommendations [2, 3, 6, 12], though multiple studies noted challenges with versioning,

organizational disagreement, and policy drift [4, 14].

Biomedical literature (e.g., PubMed-derived corpora) was often incorporated to integrate emerging evidence, introducing additional burdens related to study quality appraisal and reconciliation of conflicting findings [8, 26–28].

Structured medication resources (drug databases and formularies) were used for medication-centered tasks where precision and contraindication handling are safety-critical [1, 13, 14, 29].

Institutional protocols and local knowledge bases were employed to align outputs with site-specific practice patterns and operational constraints [7, 22, 30].

3.2.3 Retrieval–generation integration models

Integration strategies varied in the degree to which retrieved evidence constrained generation.

Context concatenation (top- k passages appended to the prompt) was the most commonly described integration pattern due to simplicity and portability [2, 8, 10], but remained vulnerable to context-length limitations and imperfect evidence utilization [16].

Iterative or agentic pipelines introduced multi-step retrieval and refinement processes, improving handling of multi-hop or ambiguous queries at the cost of latency and additional failure points [9, 17, 19].

Confidence- or rule-modulated integration approaches attempted to prioritize higher-relevance evidence or down-weight weak retrievals to reduce propagation of irrelevant context [16, 17, 25].

3.2.4 Hallucination-targeted mitigation features

Beyond core retrieval mechanics, studies have described additional safeguards to mitigate unsupported generation.

Citation requirements were frequently used to operationalize verifiability, although multiple reports cautioned that citation presence does not guarantee correct attribution or faithful reasoning [2, 26, 28].

Uncertainty signaling (confidence scores or abstention mechanisms) sought to reduce overconfident errors but raised calibration challenges [4, 5, 16].

Multi-source verification reduced reliance on single potentially erroneous passages, improving precision while sometimes reducing recall for rare conditions [13, 17, 25].

Structured output constraints (templates or schema-bound generation) were used to limit free-form fabrication and facilitate downstream verification, though occasionally at the expense of expressive flexibility [15, 18, 22].

Collectively, these patterns indicate that clinical RAG architectures do not eliminate the risk of hallucinations but instead redistribute it from internal model fabrication to retrieval quality, knowledge

base curation, and evidence integration fidelity.

Figure 3 Mechanistic representation of hallucination risk transition across retrieval, knowledge section, integration strategy, and response generation. Risk nodes illustrate common failure patterns reported in the included studies, demonstrating that retrieval augmentation redistributes, rather than eliminates, hallucination risk within the pipeline.

While Fig. 3 focuses on architectural risk transition within the retrieval-generation pipeline, Fig. 1 situates these technical risk points within broader evaluation, safety assessment, and deployment contexts.

3.3 Effectiveness of RAG in Hallucination Reduction (RQ2)

Twelve studies reported statistically tested comparative quantitative outcomes evaluating the impact of retrieval augmentation in clinical language model applications (Table 2). These studies spanned diverse clinical domains, including general medical question answering, preoperative risk assessment, prostate cancer screening, nursing consultation, radiology interpretation, ophthalmology decision support, and sepsis-related biomarker evaluation. Across these settings, retrieval-augmented generation (RAG) consistently demonstrated statistically significant improvements in factual alignment, task accuracy, clinician-rated quality, or hallucination reduction relative to non-retrieval baselines or human comparators.

Improvements were most pronounced in structured, guideline-concordant tasks. For example, RAG-based systems achieved 95.5% concordance in prostate-specific antigen screening compared to 62–74% among junior clinicians [6], and 96.4% accuracy in preoperative fitness assessment compared to 86.6% for human-generated responses [5]. Retrieval refinement strategies, including adaptive self-query mechanisms and optimized document chunking, yielded substantial gains in accuracy (87% vs. 50%) under controlled experimental conditions [17]. In specialty-focused applications, radiology and nursing systems demonstrated significant improvements in correctness and expert-rated performance, and one terminology-enhanced RAG model reported a 23.7% reduction in hallucination rate ($p < 0.001$) [28]. Agentic retrieval frameworks further reduced unsupported claims in radiology question answering while improving overall diagnostic alignment [9].

However, performance gains were not uniform across tasks. In more open-ended diagnostic reasoning scenarios, improvements were smaller or reflected non-inferiority rather than superiority relative to expert clinicians. These findings suggest that RAG is particularly effective in settings where high-quality, structured knowledge sources are available and where task objectives align closely with retrieval-based grounding.

Despite the frequency of reported numeric improvements (21/30 studies), inferential statistical testing was conducted in only 11 studies. Evaluation denominators varied widely, ranging from 30 to 3,234 queries or responses, and outcome definitions were heterogeneous, encompassing automated similarity metrics (e.g., ROUGE, F1), clinician-rated factuality, guideline concordance, satisfaction scores, and explicitly measured hallucination rates. Paired baseline–RAG reporting was inconsistently structured,

and effect sizes were not standardized across studies. Given this methodological heterogeneity, statistical pooling or meta-analysis was not performed.

Collectively, the evidence indicates that RAG improves factual grounding and reduces unsupported generation compared to non-retrieval LLM configurations. Nevertheless, retrieval augmentation does not eliminate residual hallucination risk. Instead, the locus of failure shifts from purely generative fabrication toward retrieval quality, knowledge base completeness, and integration fidelity. Effectiveness is therefore architecture-sensitive and task-dependent rather than universally guaranteed.

Table 2

Statistically tested comparative quantitative outcomes among included RAG studies (n = 12). Only studies reporting explicit inferential statistical testing are included.

Ref	Clinical Task	Comparator	Primary come(s)	Out-	Evaluation Scale	Statistical Test
[2]	Clinical	RAG vs baseline	Factuality,	complete-	314clinical	Significant

[5]	medicine QA (Almanac) Preoperative	LLM RAG vs human	ness, safety improvement Accuracy: 96.4% vs	questions; 8 clinicians + 2 practitioners 14 scenarios;	improvement ¹ p = 0.016
	fitness assessment	responses	86.6%	3,234 outputs	
[6]	PSA testing guideline concordance	RAG vs junior clinicians	Concordance: 95.5% vs 62.3% / 74.1%	44 scenarios	p < 0.001
[8]	Preoperative medicine case study	GPT-4 RAG vs GPT-4 vs human	80.1% → 91.4%; non-inferior to 86.3% human	1,260 responses	p = 0.610 (non-inferiority)
[10]	Nursing QA (NurRAG)	RAG vs conventional LLM	ROUGE-L; accuracy gains	1,000 expert-verified Q/A pairs	p < 0.001
[9]	Radiology QA (Agentic RAG)	Agentic RAG vs baseline LLMs	Accuracy 73% vs 64–68%; hallucination ↓9.4%	104 expert-curated questions; 24 LLMs	p < 0.001
[26]	Breast cancer nursing care	RAG-GPT vs GPT-4	Satisfaction 8.4 vs 5.4; accuracy 8.6 vs 5.6	200 clinical questions	p < 0.01
[16]	Adaptive self-query retrieval	Refined RAG vs baseline RAG	Accuracy 87% vs 50%; precision/recall/F1 ↑	30 postoperative queries	p < 0.01; p = 0.03 (ANOVA + Tukey)
[17]	Advanced chunking RAG	Enhanced vs fixed chunking	Accuracy 87% vs 50%; F1 ↑	30 postoperative queries	p = 0.001 (ANOVA +

					Tukey)
[27]	Ophthalmology	LLM+RAG vs	Correctness 54.5%	100 questions;	p
	QA	LLM alone	vs 20.6%; evidence	10 clinicians	< 0.001; p
			attribution ↑	= 0.03	
[28]	Terminology-enhanced RAG	Enhanced vs traditional RAG	Hallucination ↓23.7%; precision	Medical consultation dataset ¹	p < 0.001
			+18.4%; recall	(95% CI	
			+15.2%	reported)	
[30]	Sepsis decision support (MetaSepsisKnowHub)	RAG vs 4 baseline LLMs	Expert-rated quality scores (mean differences)	427 biomarkers; expert evaluation	p = .02; p < .001

Notes: Only studies explicitly reporting inferential statistical testing were included. Evaluation denominators are extracted from primary reports. Metrics are heterogeneous and cannot be pooled directly.

¹ Dataset size confirmed in full text; the abstract did not specify the exact denominator.

¹ Dataset size confirmed in full text; the abstract did not specify the exact denominator.

3.4 Evaluation Methodologies and Metrics (RQ3)

Evaluation approaches for clinical RAG systems were highly heterogeneous, ranging from benchmark testing and expert-curated scenarios to limited workflow-embedded studies.

3.4.1 Evaluation Contexts

Studies used three primary evaluation settings. First, standardized medical benchmarks enabled controlled comparison across models [5, 8, 13, 27], but several authors cautioned that benchmark performance did not necessarily reflect clinical reliability [4, 20].

Second, many studies employed expert-curated clinical scenarios to assess guideline adherence or specialty-specific reasoning [2, 6, 11, 12], thereby improving ecological validity but limiting cross-study comparability.

Third, a smaller subset evaluated systems within realistic clinical workflows or using real-world queries [7, 22, 23], though long-term outcome monitoring was not reported.

3.4.2 Evaluation Metrics

Four broad metric categories were observed:

1. Automated similarity metrics (e.g., ROUGE, BLEU) assessed response overlap with reference answers [8, 10, 27] but did not reliably indicate clinical correctness [20, 28].
2. Factual accuracy and hallucination assessment relied primarily on expert review or binary correctness judgments [2, 4–6, 14, 28].
3. Clinical appropriateness measures, including guideline concordance and diagnostic alignment, were used in specialty-focused studies [6, 12, 13, 23] but lacked standardization [20].
4. Retrieval quality metrics evaluated document relevance and precision [9, 16, 17, 28], though strong retrieval did not consistently guarantee faithful generation [16].

3.4.3 Structural Evaluation Gap

Across studies, a consistent misalignment emerged between technical performance metrics and clinical safety assessment. High similarity or benchmark scores did not uniformly correspond to clinically safe recommendations [20, 28]. Few investigations explicitly evaluated error severity, patient harm risk, or downstream clinical decision impact [2, 4, 20], and no long-term prospective safety monitoring was described.

These findings indicate that evaluation practices remain more aligned with NLP benchmarking paradigms than with established standards for clinical validation.

3.5 Failure Modes and Limitations (RQ4)

Although retrieval-augmented systems improved factual alignment relative to baseline LLMs, recurrent failure patterns persisted across retrieval, integration, reasoning, and safety domains.

3.5.1 Retrieval and Knowledge Base Limitations

Failures frequently originated in the retrieval stage. Irrelevant or weakly relevant documents were sometimes retrieved and incorporated into downstream generation, degrading response quality despite acceptable relevance scores [9, 16, 17]. Retrieval recall limitations were observed for rare conditions, ambiguous terminology, and multi-source synthesis tasks [4, 13, 25, 27].

Knowledge base curation further constrained performance. Several studies noted challenges related to outdated guideline versions, incomplete document coverage, or static corpora without drift monitoring [14, 20, 27]. No included study described automated surveillance mechanisms for guideline updates or knowledge base currency.

3.5.2 Evidence Integration and Residual Hallucination

Even when relevant documents were retrieved, models occasionally generated responses inconsistent with retrieved evidence or selectively cited supportive content while omitting qualifiers or conflicting information [2, 16, 28]. Citation presence, therefore, did not guarantee faithful grounding.

Residual hallucination persisted across multiple studies despite retrieval conditioning [4, 14, 28], indicating that RAG mitigates but does not eliminate unsupported generation.

3.5.3 Reasoning and Synthesis Constraints

Systems demonstrated difficulty resolving conflicting recommendations across retrieved sources or expressing calibrated uncertainty [4, 13, 17]. Context-length limitations and multi-hop reasoning demands further constrained performance in complex clinical scenarios [9, 16].

3.5.4 Safety-Relevant Vulnerabilities

Several reports described confident but incorrect recommendations, medication-related inaccuracies, or incomplete differential reasoning in safety-critical contexts [1, 4, 13, 14, 27]. These findings underscore that retrieval augmentation shifts the risk of hallucinations toward retrieval fidelity and integration quality rather than fully eliminating it.

Collectively, failure patterns suggest that performance gains from RAG are bounded by retrieval robustness, knowledge base quality, and the reliability of evidence integration.

3.6 Clinical Readiness and Deployment Feasibility (RQ5)

Despite architectural innovation and statistically tested performance improvements, evidence of real-world clinical deployment was absent across included studies. All evaluations were conducted in retrospective, simulated, or expert-reviewed settings without prospective clinical validation or regulatory-aligned implementation.

3.6.1 Translational Gap

Most systems were evaluated using curated datasets, controlled scenario testing, or structured expert adjudication [2, 8, 10]. While several studies demonstrated statistically significant performance gains [5, 6, 16], none reported prospective workflow-embedded trials, multi-site validation, or regulatory clearance.

These findings indicate that methodological rigor in benchmarking has progressed more rapidly than evidence of operational clinical readiness.

3.6.2 Workflow and Infrastructure Constraints

Practical deployment barriers were acknowledged but rarely empirically evaluated. Latency and computational overhead were identified as constraints for multi-step or agentic retrieval pipelines [9, 17, 19]. Integration with electronic health record systems and patient-specific contextualization was

discussed conceptually in some reports [7, 12, 22], but real-world integration outcomes were not described.

Standardized usability assessment, clinician interaction design, and longitudinal monitoring frameworks were largely absent from empirical reports [20].

3.6.3 Regulatory and Governance Considerations

Explicit discussion of regulatory classification, medical device pathways, or compliance frameworks was limited [11, 20]. No study reported prospective multi-site validation consistent with regulatory expectations, nor did any describe post-deployment safety monitoring systems.

Issues of accountability, liability, and error-reporting governance were acknowledged in conceptual discussions [4], but not operationalized within empirical evaluations.

3.6.4 Human–AI Collaboration and Trust

Several studies suggested that transparent citation and evidence presentation improved perceived reliability [2, 28]. However, systematic measurement of clinician trust, calibration of uncertainty, and long-term adoption outcomes was limited [4, 16]. Overconfidence in incorrect responses was identified as a persistent safety concern [4]. Overall, the current evidence positions clinical RAG systems at a stage of retrospective validation. While technically promising, substantial gaps remain in prospective validation, workflow integration, regulatory alignment, and longitudinal governance

structures.

3.7 Evidence Maturity Profile

Across the included studies, evaluation rigor and deployment readiness were unevenly distributed. Although architectural refinement and statistically significant performance gains have been reported in several studies, most evaluations have remained confined to retrospective benchmarking, curated expert review, or controlled experimental settings. Prospective validation within live clinical workflows, multi-site

replication, discussion of regulatory pathways, and longitudinal safety monitoring were infrequently described.

To synthesize these structural patterns, we introduce a conceptual evidence maturity framework positioning clinical RAG systems along two independent dimensions:

(i) evaluation rigor and (ii) deployment readiness. Evaluation rigor spans from automated similarity metrics and internal benchmarks to expert adjudication and prospective validation. Deployment readiness ranges from prototype or proof-of-concept systems to workflow-embedded and regulatorily aligned implementations.

This framework is interpretive rather than prescriptive. It does not rank individual systems nor imply quantitative scoring. Instead, it provides a structural lens for understanding where the current body of evidence is concentrated and where translational gaps remain. Consistent with findings from RQ2–RQ5, the majority of systems cluster within retrospective, experimentally controlled settings, with comparatively limited representation in prospective or operational clinical environments. To synthesize these structural patterns without imposing quantitative scoring, we propose a conceptual evidence maturity framework that situates included clinical RAG systems along two dimensions: evaluation rigor and deployment readiness (Fig. 4).

4 Discussion

4.1 Principal Findings

This review synthesizes evidence from 30 studies evaluating retrieval-augmented generation (RAG) for hallucination mitigation in clinical decision support. The literature demonstrates rapid architectural experimentation and consistent improvements in factual alignment relative to non-retrieval baselines [2, 5, 8, 9]. Several studies reported statistically significant gains in accuracy, guideline concordance, or hallucination reduction when retrieval from curated medical sources was incorporated [4, 14, 28].

However, these gains were heterogeneous across tasks and domains, and residual hallucination persisted even in high-performing systems, particularly in complex or ambiguous clinical scenarios [4, 14, 28]. Thus, current evidence supports mitigation—not elimination—of unsupported generation.

Architectural diversity was substantial. Studies implemented dense guideline-grounded retrieval [2], hybrid retrieval strategies [9, 16], agentic multi-step reasoning [9, 17], and structured or knowledge-graph integrations [20, 25]. No single configuration emerged as universally superior; performance appeared context-dependent and sensitive to task structure and knowledge base quality.

A defining pattern was the imbalance between technical benchmarking and translational validation. While improvements were frequently demonstrated in retrospective settings, no included study reported prospective clinical validation or regulatory-aligned deployment. This structural maturity gap suggests that architectural advancement has outpaced the advancement of validation science.

4.2 Evaluation Rigor and Clinical Risk

A recurring theme was misalignment between reported evaluation metrics and clinically meaningful risk assessment.

Automated NLP metrics (e.g., ROUGE, BLEU, embedding similarity) were commonly used as primary indicators [8, 10, 27]. Yet high similarity scores do not necessarily reflect factual correctness or clinical appropriateness [20, 28]. Benchmark performance may therefore overestimate real-world safety.

Error characterization was typically binary, with limited stratification by severity or potential patient harm [2, 4]. Few studies have incorporated structured harm-grading frameworks [4, 20]. Given the asymmetric consequences of clinical errors, the absence of severity-sensitive evaluation constrains the interpretation of reported hallucination rates.

Moreover, benchmark datasets do not fully capture the real-world clinical ambiguity. Studies evaluating expert-curated or real clinical queries reported more variable performance than that observed in standardized tasks [4, 7, 23]. Importantly, no study linked RAG system use to downstream patient outcomes.

Collectively, improvements in grounding should not be equated with demonstrated clinical safety.

4.3 Implications for Clinical Deployment

Current evidence supports cautious, scoped use cases rather than autonomous decision-making.

RAG systems appear most suitable for bounded, knowledge-grounded tasks such as guideline interpretation or protocol clarification [6, 11, 12]. Open-ended diagnostic reasoning, rare disease identification, and multi-source conflict resolution remain challenging [4, 13, 27].

Given persistent residual hallucination, human–AI collaborative models are more consistent with the current evidence base than fully automated clinical decision systems [2, 6, 28]. Deployment readiness depends not only on model architecture but also on curated and continuously updated knowledge bases [14, 27], workflow integration infrastructure [7, 22], and governance mechanisms for monitoring and accountability [20].

Regulatory pathways and post-deployment surveillance frameworks were rarely addressed [11, 20]. In the absence of prospective validation, claims of clinical readiness remain premature.

4.4 Limitations of the Review

This synthesis is constrained by heterogeneity in reporting practices and evaluation metrics, precluding formal meta-analysis. Findings reflect structured qualitative synthesis rather than pooled effect sizes.

The literature base is dominated by retrospective studies and includes preprints [2, 8, 12]. Real-world performance may differ from reported results. Rapid methodological evolution may further limit temporal generalizability. Finally, publication bias and selective reporting remain possible despite the inclusion of preprints.

4.5 Open Challenges and Research Priorities (RQ6)

Several priorities emerge.

First, standardized clinical evaluation frameworks incorporating severity-stratified error analysis, calibration assessment, and prospective validation are needed [2, 4, 20]. Second, retrieval robustness

remains a technical bottleneck, particularly for rare conditions and conflicting evidence [3, 16, 17, 24].

Pre-generation retrieval quality

assessment may mitigate downstream errors.

Third, improving the integration of faithful evidence and calibrated uncertainty communication remains critical [4, 14, 28].

Fourth, handling clinical complexity—including multimodal reasoning and patient-specific personalization—requires scalable architectures beyond text-only retrieval [12, 18, 29].

Finally, deployment science—including regulatory alignment, quality management systems, and post-deployment monitoring—must advance in parallel with algorithmic innovation [11, 20, 22].

5 Conclusions

This systematic review synthesized empirical evidence from 30 studies evaluating retrieval-augmented generation (RAG) for hallucination mitigation in clinical decision support. Across diverse tasks, RAG systems demonstrated improved factual grounding relative to baseline large language models. However, hallucination was mitigated rather than eliminated, and recurrent failure modes—particularly retrieval errors and incomplete evidence synthesis—persisted.

A central finding is the structural gap between architectural advancement and the maturity of clinical validation. Most systems have been evaluated retrospectively using automated or benchmark-based metrics, with limited clinician-centered safety assessment and no prospective outcome studies. Accordingly, current evidence supports RAG as a grounding-enhancement mechanism under controlled conditions, not as proof of clinical readiness.

RAG represents an important step toward safer clinical language modeling, but translating it into high-stakes healthcare environments will require standardized, severity-aware evaluation frameworks, prospective validation, and regulatory-aligned governance. Progress in hallucination mitigation must be accompanied by parallel advancement in clinical validation science.

References

1. Ozmen BB, Singh N, Shah K et al (2025) MicroRAG: Development of a Novel Artificial Intelligence Retrieval-Augmented Generation Model for Microsurgery Clinical Decision Support. *Microsurgery* 45(8):e70138. <https://doi.org/10.1002/micr.70138>
2. Hiesinger W, Zakka C, Chaurasia A, Shad R, Dalal AR, Moor M et al (2024) Almanac: Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* 1(2):Aloa2300068. <https://doi.org/10.1056/Aloa2300068>

3. Li W, Zhang H, Zhang H, Li Z, Dong Z, Chen Y et al Refine Medical Diagnosis Using Generation Augmented Retrieval and Clinical Practice Guidelines. *IEEE Journal of Biomedical and Health Informatics*. 2025;PP. Early access. <https://doi.org/10.1109/JBHI.2025.3641931>
4. Rakha I, Abbas N (2025) Large Language Models in Clinical Advice: Direct Generation and Retrieval Augmented Generation vs Expert Advice. In: 15th International Conference on Computer Science, Engineering and Applications (CSEA 2025)
5. Ke Y, Jin L, Elangovan K et al (2025) Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit Med* 8(1):187 Published: 05 April 2025. <https://doi.org/10.1038/s41746-025-01519-z>
6. Tung JYM, Le QA, Yao J, Huang Y, Lim DYZ, Sng GGR et al (2025) Retrieval-Augmented-Generation Large Language Models Outperform Junior Clinicians in Guideline-Concordant PSA Testing. *Res Square*. <https://doi.org/10.21203/rs.3.rs-5760861/v1>
7. Hu X, Zhou L, Li H, Fang Y (2025) A Self-Correcting Agentic Graph RAG for Clinical Decision Support in Hepatology. *Front Med*. <https://doi.org/10.3389/fmed.2025.1716327>
8. Ke Y, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH et al (2024) Development and Testing of Retrieval Augmented Generation in Large Language Models: A Case Study Report. *arXiv preprint*. <https://doi.org/10.48550/arxiv.2402.01733>
9. Wind S, Sopa J, Truhn D, Lotfinia M, Nguyen T, Bressemer KK et al (2025) Agentic Large Language Models Improve Retrieval-Based Radiology Question Answering. *arXiv preprint*. <https://doi.org/10.48550/arxiv.2508.00743>
10. Xiong L, Zeng Q, Luo W, Liu R (2025) Nursing Retrieval-Augmented Generation: Retrieval Augmented Generation for Nursing Question Answering with Large Language Models. *Int J Nurs Sci* 12(6):516–523. <https://doi.org/10.1016/j.ijnss.2025.10.005>
11. Sriram A, Maheswaran N, Sundan B, Krishnamoorthy S (2025) Context-Aware Retrieval-Augmented Generation for Artificial Intelligence in Urology. *Cureus*. <https://doi.org/10.7759/cureus.88167>
12. Liu CC, Yao HR, Chang DC, Frieder O (2025) TreatRAG: A Framework for Personalized Treatment Recommendation. In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. New York, NY, USA: Association for Computing Machinery; pp. 690–695. Available from: <https://doi.org/10.1145/3705328.3748022>
13. Vaid A, Lampert J, Lee J, Sawant A, Apakama D, Sakhuja A et al (2024) Natural Language Programming in Medicine: Administering Evidence Based Clinical Workflows with Autonomous Agents Powered by Generative Large Language Models. *arXiv preprint*. <https://doi.org/10.48550/arxiv.2401.02851>
14. Priola MP (2024) Addressing Hallucinations with RAG and NMISS in Italian Healthcare LLM Chatbots. *arXiv preprint*. <https://doi.org/10.48550/arxiv.2412.04235>
15. Ansari MS, Khan MSA, Revankar S, Varma A, Mokhade AS (2025) Lightweight Clinical Decision Support System using QLoRA-Fine-Tuned LLMs and Retrieval-Augmented Generation. *arXiv preprint* <https://doi.org/10.48550/arxiv.2505.03406>

16. Prabha S, Gomez-Cabello CA, Haider SA, Genovese A, Trabilisy M, Wood NG et al (2025) Enhancing Clinical Decision Support with Adaptive Iterative Self-Query Retrieval. *Bioengineering*. <https://doi.org/10.3390/bioengineering12080895>
17. Gomez-Cabello CA, Prabha S, Haider SA, Genovese A, Collaco JM, Wood NG et al (2025) Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation. *Bioengineering*. <https://doi.org/10.3390/bioengineering12111194>
18. Raza A, Salem M, Kwon S, Hussain M, Gu Y, Al-Antari M (2025) Multimodal Knowledge-Infused VLM for Respiratory Disease Prediction and Clinical Report Generation. *IEEE J Biomedical Health Inf* <https://doi.org/10.1109/JBHI.2025.3631264>
19. Salehi S, Singh R, Horst K, Hathaway Q, Erickson B (2025) Agentic AI and Large Language Models in Radiology: Opportunities and Hallucination Challenges. *Bioengineering*. <https://doi.org/10.3390/bioengineering12121303>
20. Nanua S, Steward C, Neely R, Datto M, Youens K (2025) Retrieval-Augmented Generation for Interpreting Clinical Laboratory Regulations. *J Pathol Inf*. <https://doi.org/10.1016/j.jpi.2025.100520>
21. Wang Z (2025) Retrieval-Augmented Question Answering System Based on Large Language Models and Knowledge Graphs. *Appl Comput Eng* 178:122–131. <https://doi.org/10.54254/2755-2721/2025.po25800>
22. Abd-Elhamid AI, Khalili A, Arabi M, Sattar AS, Hosseini A, Serag A (2025) MERA: Medical Electronic Records Assistant. *Mach Learn Knowl Extr*. <https://doi.org/10.3390/make7030073>
23. Algosaiibi A, Sait ARW (2025) Improving GPT-Driven Medical Question Answering Using SPARQL–Retrieval-Augmented Generation Techniques. *Electronics*. <https://doi.org/10.3390/electronics14173488>
24. Lu Y, Zhao X, Wang J, ClinicalRAG et al (2024) Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In: Li S, Li M, Zhang MJ, Choi E, Geva M, Hase P, editors. *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*. Bangkok, Thailand: Association for Computational Linguistics; pp. 64–68. Available from: <https://aclanthology.org/2024.knowllm-1.6/>
25. Puhakka O Usability of Local Large Language Models and Retrieval Augmented Generation in Health Care. Available from: <https://oulurepo.oulu.fi/handle/10024/55664>
26. Xu Ry, Hong Y, Zhang F, Xu H (2024) Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. *Sci Rep*. <https://doi.org/10.1038/s41598-024-81052-3>
27. Gilson A, Ai X, Arunachalam T, Chen Z, Cheong KX, Dave A et al (2024) Enhancing Large Language Models with Domain-specific Retrieval Augment Generation: A Case Study on Long-form Consumer Health Question Answering in Ophthalmology. *arXiv preprint arXiv:240913902*
28. Guan H (2025 November) Medical Terminology Definition-Enhanced Retrieval-Augmented Generation for Hallucination Mitigation in Medical Question Answering. *J Sci Innov Social Impact* 1(1):222–240 Accessed: 2026-01-18

29. Bazzi Mohamed Salim E, Anass T, Ider Abdelouahed A (2025) Advancing Multilingual Retrieval-Augmented Generation for Reliable Medication Counseling. IEEE Access 13:215550–215564. <https://doi.org/10.1109/ACCESS.2025.3646941>
30. Zhang C, Yang H, Liu X, Wu R, Zong H, Wu E et al (2025) A Knowledge-Enhanced Platform (MetaSepsisKnowHub) for Retrieval Augmented Generation–Based Sepsis Heterogeneity and Personalized Management: Development Study. J Med Internet Res 27:e67201. <https://doi.org/10.2196/67201>

Figures

Clinical RAG Hallucination Mitigation Framework

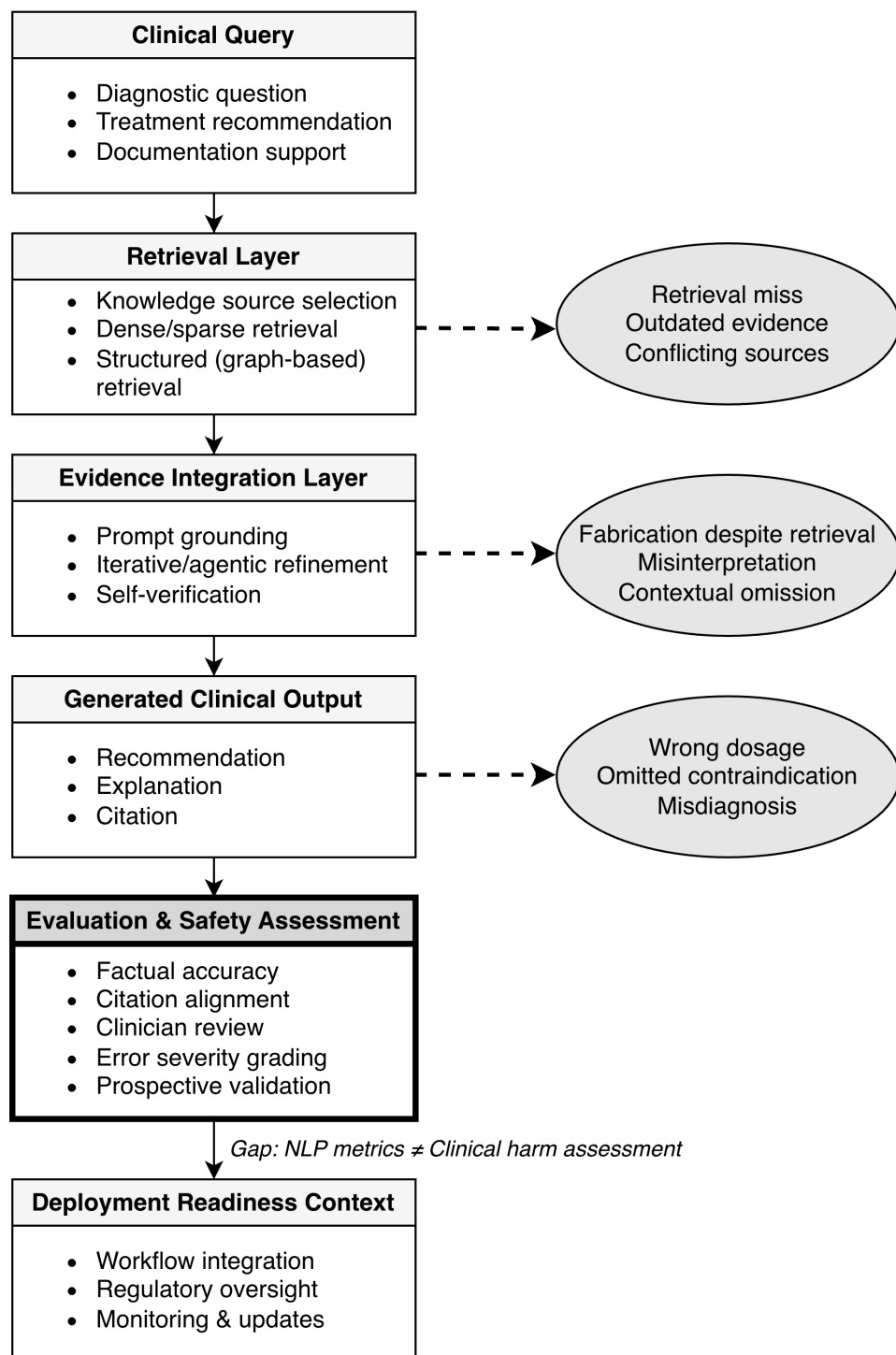


Figure 1

Systems-level conceptualization of retrieval-augmented clinical decision support pipelines, illustrating retrieval, evidence integration, generation, evaluation, and deployment stages alongside representative hallucination failure modes and safety assessment considerations.

PRISMA 2020 Flow Diagram

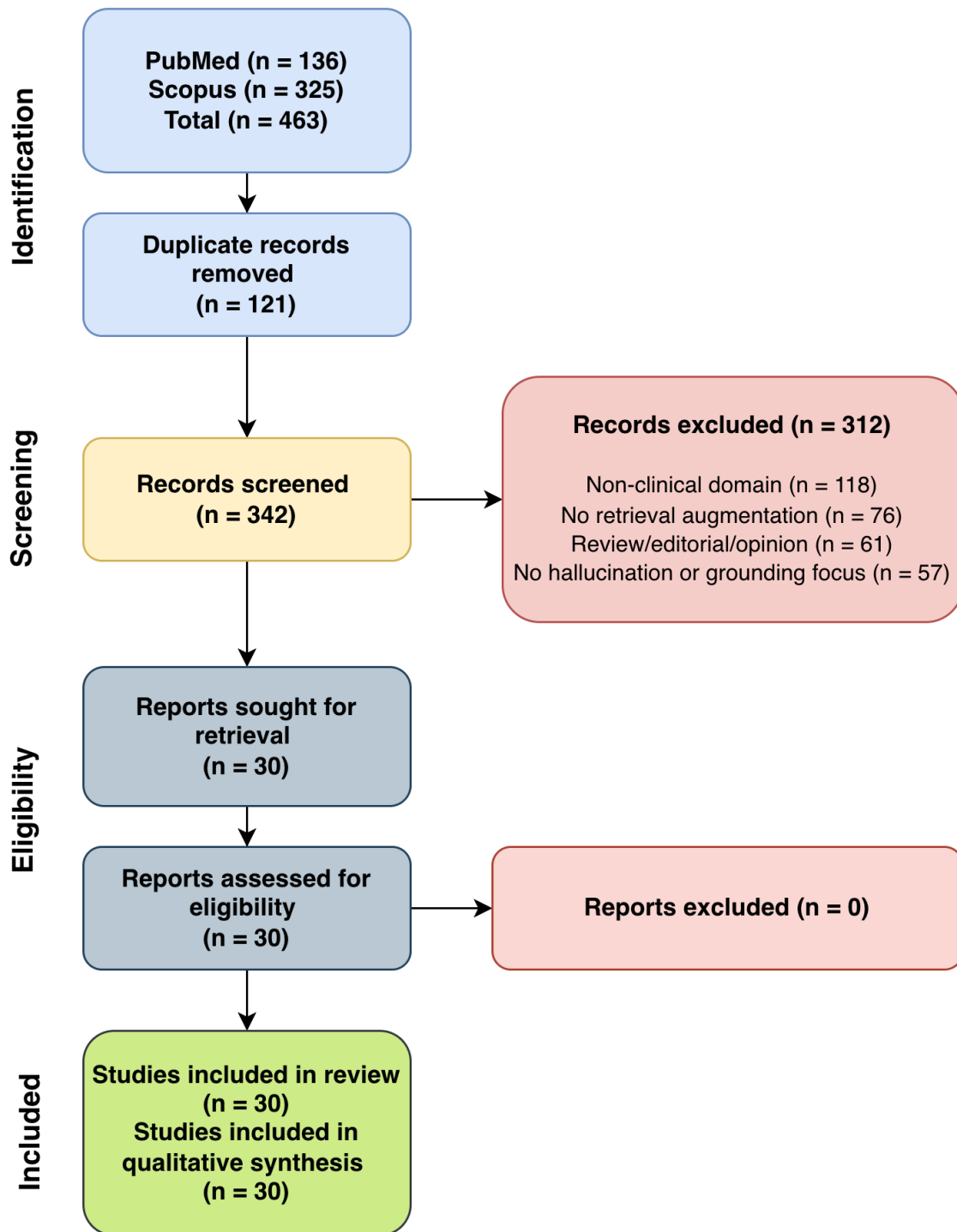


Figure 2

PRISMA 2020 flow diagram summarizing study identification, deduplication, screening, eligibility assessment, and final inclusion of studies evaluating retrieval-augmented large language models for clinical decision support.

Architectural Risk Redistribution in Clinical RAG Systems

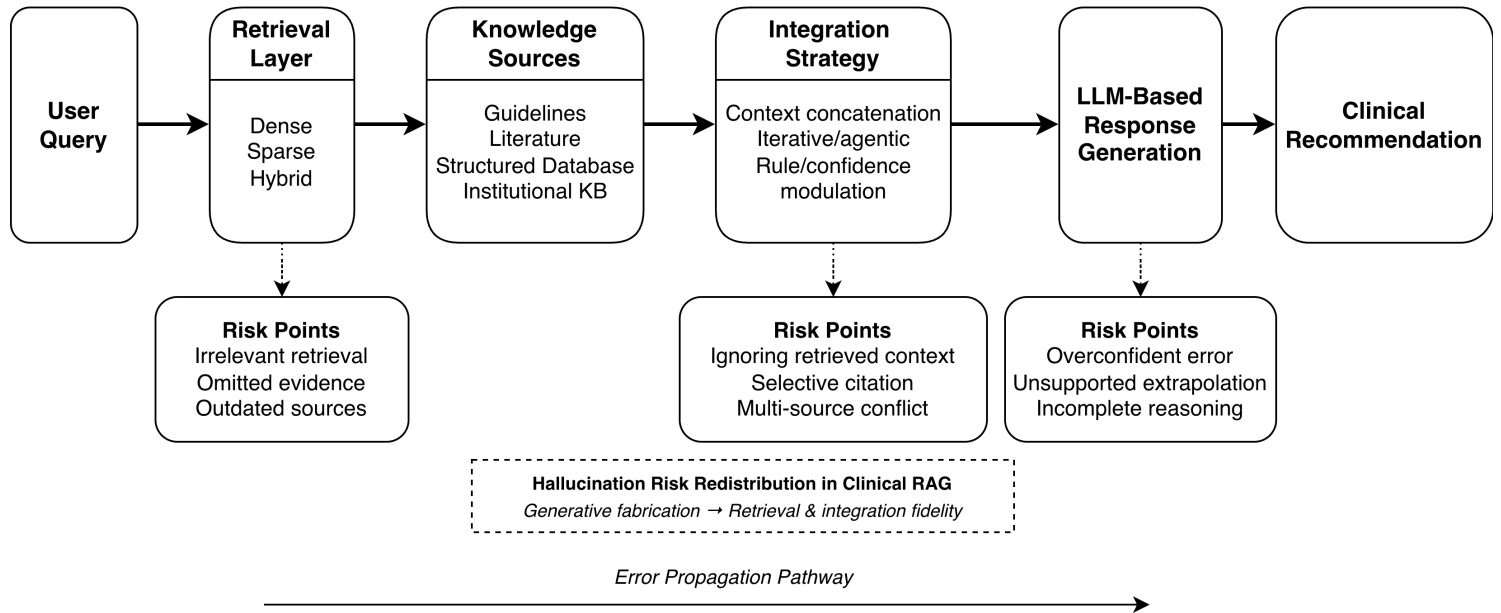


Figure 3

Hallucination risk redistribution across retrieval, knowledge sourcing, integration, and response-generation stages in clinical RAG systems.

Evidence Maturity Framework for Clinical RAG Systems

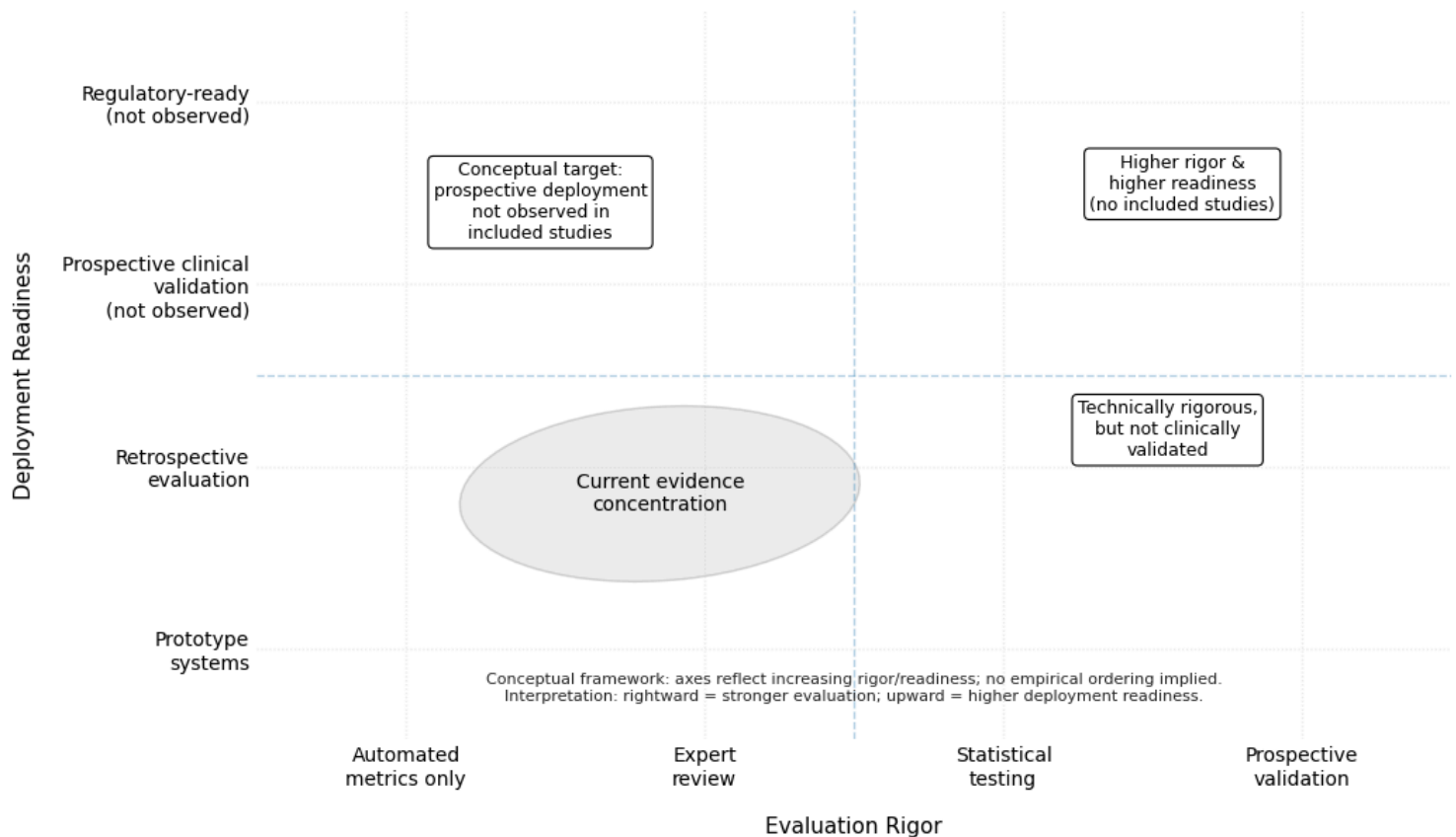


Figure 4

Conceptual evidence maturity framework illustrating evaluation rigor and deployment readiness patterns in clinical RAG systems identified in the included studies.