Journal of the American Medical Informatics Association, 2025, **32(4)**, 605–615
https://doi.org/10.1093/jamia/ocaf008
Advance access publication 15 January 2025
Research and Applications

Research and Applications

Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines

Siru Liu, PhD*[1,2], Allison B. McCoy, PhD[1], Adam Wright, PhD[1,3]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212, United States, [2]Department of Computer Science, Vanderbilt University, Nashville, TN 37212, United States, [3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37212, United States

*Corresponding author: Siru Liu, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave #1475, Nashville, TN 37212, United States (siru.liu@vumc.org)

Abstract

Objective: The objectives of this study are to synthesize findings from recent research of retrieval-augmented generation (RAG) and large language models (LLMs) in biomedicine and provide clinical development guidelines to improve effectiveness.

Materials and Methods: We conducted a systematic literature review and a meta-analysis. The report was created in adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 analysis. Searches were performed in 3 databases (PubMed, Embase, PsycINFO) using terms related to "retrieval augmented generation" and "large language model," for articles published in 2023 and 2024. We selected studies that compared baseline LLM performance with RAG performance. We developed a random-effect meta-analysis model, using odds ratio as the effect size.

Results: Among 335 studies, 20 were included in this literature review. The pooled effect size was 1.35, with a 95% confidence interval of 1.19-1.53, indicating a statistically significant effect ($P = .001$). We reported clinical tasks, baseline LLMs, retrieval sources and strategies, as well as evaluation methods.

Discussion: Building on our literature review, we developed **G**uidelines for **U**nified **I**mplementation and **D**evelopment of **E**nhanced LLM Applications with **RAG** in Clinical Settings to inform clinical applications using RAG.

Conclusion: Overall, RAG implementation showed a 1.35 odds ratio increase in performance compared to baseline LLMs. Future research should focus on (1) system-level enhancement: the combination of RAG and agent, (2) knowledge-level enhancement: deep integration of knowledge into LLM, and (3) integration-level enhancement: integrating RAG systems within electronic health records.

Key words: large language model; retrieval augmented generation; systematic review; meta-analysis.

Introduction

Large language models (LLMs) have reported remarkable performance in question-answering, summarization, and text generation.[1] Given this, researchers have explored its potential in biomedical areas.[2] For example, several studies reported the ability of using LLM to answer patient messages,[3] to analyze alert logic in clinical decision support,[4] and to make discharge summaries more readable to patients.[5] However, several challenges remain.

LLMs are trained on fixed datasets, which restrict their knowledge to information available up to the training cut-off date. For example, GPT-4o's training data only includes information up to October 2023, making it unable to respond accurately to findings that emerged afterward. LLM training datasets are also generally broad and lack the specificity required for biomedical applications. Finally, not all sources used to train the LLMs are reliable and trustworthy. To address these limitations, researchers have performed fine-tuning and retrieval-augmented generation (RAG) techniques. Fine-tuning can adapt LLMs to specific domains, but it is resource-intensive and does not allow for real-time updates. In contrast, RAG maintains the original LLM architecture while incorporating relevant context directly into queries, offering more flexibility and control. In addition, RAG's unique advantage in biomedical applications lies in its ability to adapt to dynamic environments by delivering up-to-date information and efficiently integrating external knowledge sources with high interpretability.[6]

Another limitation of using LLMs directly is the risk of hallucination, where the model generates incorrect or fabricated information.[7] To mitigate such issues, researchers have proposed RAG as a solution that integrates up-to-date, relevant information, enhancing both the accuracy and reliability of LLM generated responses.[8,9] For example, when ChatGPT

Received: November 19, 2024; **Revised:** December 17, 2024; **Editorial Decision:** January 2, 2025; **Accepted:** January 3, 2025

was asked about medications for peripheral artery disease patients without increased bleeding risk, it initially omitted low-dose rivaroxaban. After integrating retrieved text from the 2024 American College of Cardiology / American Heart Association Guideline for the Management of Lower Extremity Peripheral Artery Disease,[10] the model correctly recommended rivaroxaban.

Several guidelines exist for evaluating Artificial Intelligence (AI) applications and LLMs in healthcare, including DECIDE-AI (Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence),[11] CLAIM (Checklist for Artificial Intelligence in Medical Imaging),[12] and CONSORT-AI (Consolidated Standards of Reporting Trials-AI).[13] In addition, Tam et al. introduced QUEST, a framework specifically for human evaluation of LLMs in healthcare.[14] However, these guidelines do not cover RAG applications in clinical settings, emphasizing the need for a more specific guideline.

Despite the promise of RAG in improving LLM performance in clinical settings, there is limited understanding of its overall effectiveness comparing with the baseline LLM, adoption in clinical domains, and optimal strategies for its development in biomedical applications. The aim of this study is to synthesize findings from recent research of RAG and LLM in biomedicine and provide clinical development guidelines to improve effectiveness as well as transparency in future research.

Materials and methods

Study design

We conducted a systematic literature review. The report was created in adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 analysis checklist (File S1).[15] We also performed a meta-analysis following the Cochrane Handbook guidelines.[16]

Screening papers

We searched in 3 databases (PubMed, Embase, and PsycINFO), using terms related to "retrieval augmented generation" and "large language model." PubMed and Embase were selected based on recommendations from the Cochrane Handbook, as they are widely recognized for their extensive coverage of biomedical literature.[17] PsycINFO was included to capture articles from the psychological and behavioral sciences. To maintain a high standard of quality and reliability, we focused on peer-reviewed articles and excluded preprints. The specific search terms used for each database are provided in File S2. Given that ChatGPT was released on November 30, 2022, we set the publication filter to search papers published in 2023 and 2024. The search was performed on December 12, 2024. The inclusion criteria were: (1) the study must compare baseline LLM performance with RAG performance and (2) the study must address a biomedical question. The exclusion criteria were: (1) literature reviews, editorial comments, or viewpoint papers, (2) studies focusing on LLMs in languages other than English, or (3) studies centered on a multi-agent system without a focus on RAG. SL screened titles and abstracts, then conducted a full-text review of papers meeting the criteria.

Data extraction

For each included study, we extracted the following information: author, title, publication year, journal, clinical task, and specialty. Regarding RAG techniques, we gathered details about the baseline LLM, retrieval sources, and strategies used in the pre-retrieval, retrieval, and post-retrieval stages. For evaluation, we extracted the evaluation method (human, automated, or a combination of both), the number of evaluators, the evaluation dataset, and the evaluation metrics.

Meta-analysis

Effect size was defined as a metric quantifying the relationship between variables, including both direction and magnitude.[18] For each included study, we calculated the effect size between baseline LLM performance and RAG-enhanced LLM performance. The outcomes focused on the performance of generation results, such as accuracy and usefulness. Metrics related to the retrieval process, cost, or speed were not included as outcomes in the meta-analysis. For continuous outcomes, we used Cohen's d, standardized between-group mean difference (SMD), calculated as the difference in means divided by the pooled standard deviation. The standard error (SE) of SMD was calculated using the following formula (1), where n_1 and n_2 represent the sample sizes of each group.[19] For dichotomous measurements, we calculated the log-odds ratio, obtained by transforming the odds ratio (OR) with the natural logarithm, and the associated SE was calculated using formula (2), where a, b, c, and d represent the number of successful and failed events in the baseline LLM and RAG-enhanced LLM approaches. For studies reporting multiple outcomes, we used the overall outcome to calculate effect size. If no overall outcome was reported, we averaged the effect sizes of all reported outcomes. We excluded outcomes with a sample size of less than 30 to avoid small-sample bias.

$$\mathrm{SE}_{\mathrm{SMD}} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{\mathrm{SMD}^2}{2(n_1+n_2)}} \tag{1}$$

$$\mathrm{SE}_{\mathrm{logOR}} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \tag{2}$$

We developed a random-effect meta-analysis model, because of the variability in RAG architectures and evaluation datasets among the included studies. The random-effect model was used when individual study effects likely contained additional sources of variance beyond sampling error. Between-study heterogeneity was assessed using Higgins & Thompson's I^2 statistic, where 25% indicated low heterogeneity, 50% moderate, and 75% substantial.[20]

We conducted subgroup analyses to explore performance variations across different factors. First, we analyzed the influence of the baseline LLM, referring to the foundation model (eg, GPT-4 or Llama2) that provides the core architecture for the system. Second, we examined data retrieval strategies, categorizing them as simple or complex. Simple strategies included fixed-length chunking and basic similarity search, and we performed a subgroup analysis to compare these with complex retrieval strategies. Third, we analyzed differences based on evaluation methods, distinguishing between human evaluations, such as Likert scale ratings for helpfulness and accuracy, and automatic evaluation metrics, including ROUGE-1 and BLEU. Finally, we conducted a subgroup analysis based on the type of task, classifying studies

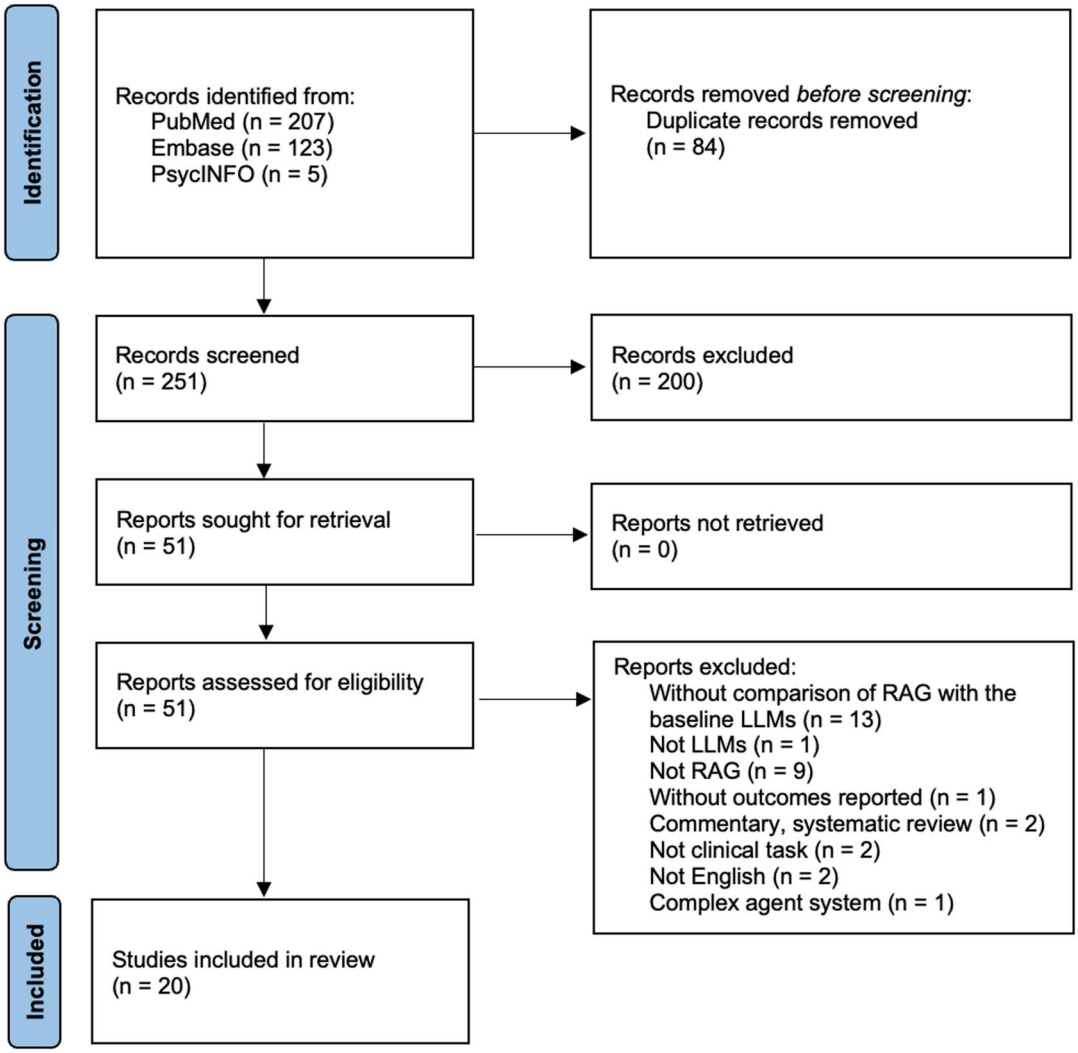

Figure 1. Flow diagram of included studies illustrating the systematic review process. Out of 335 records identified from PubMed, Embase, and PsycINFO, 84 duplicates were removed, leaving 251 records screened. After excluding 200 records, 51 reports were assessed for eligibility. 31 reports were excluded for reasons such as lack of RAG-LLM comparison, non-clinical focus, or commentary. Ultimately, 20 studies were included in the final review. LLM, large language model; RAG, retrieval-augmented generation.

into clinical decision-making and medical question-answering. These analyses provided insights into how variations in model architecture, retrieval strategies, evaluation methods, and task types affect system outcomes.

To visualize the meta-analysis outcomes, we generated a forest plot. This plot displayed the effect size, confidence interval for each study, as well as the pooled effect and predicted effect size. We evaluated the publication bias using a contour-enhanced funnel plot to investigate small-study effects. This scatter plot had the effect size on the x-axis and the inverted SE on the y-axis, with contours indicating P-values (<.1, .05, and .01).[21] Symmetry in the funnel plot suggested no publication bias, and asymmetry was quantified using Egger's regression test.[22] We used the "meta" package in R to conduct the meta-analysis and perform statistical analyses.

Results

Study selection

A total of 335 studies were identified from 3 databases: PubMed, Embase, and PsycINFO. After removing duplicates, 251 studies were screened. Of these, 20 studies were included in this literature review, all of which were published in 2024. One of the included studies was a conference paper.[23] The flow diagram depicting the study selection process is shown in Figure 1. For each included study, their author, title, publication year, journal, clinical task, specialty, and retrieval sources are listed in Table S1 of File S2.

Meta-analysis

The pooled effect size was 1.35, with a 95% confidence interval of 1.19-1.53, indicating a statistically significant effect ($P = .001$). All outcomes and associated SEs are listed in File S2. The I^2 value was 37%, indicating low to moderate heterogeneity among the studies. The prediction interval ranged from 1.01 to 1.8. The forest plot is shown in Figure 2. The contour-enhanced funnel plot is presented in File S2. In Egger's regression test, the intercept (β_0) was 1.1, with a 95% confidence interval of [0.56, 1.64] and a P-value of .001, indicating the presence of small-study effects and potential publication bias.

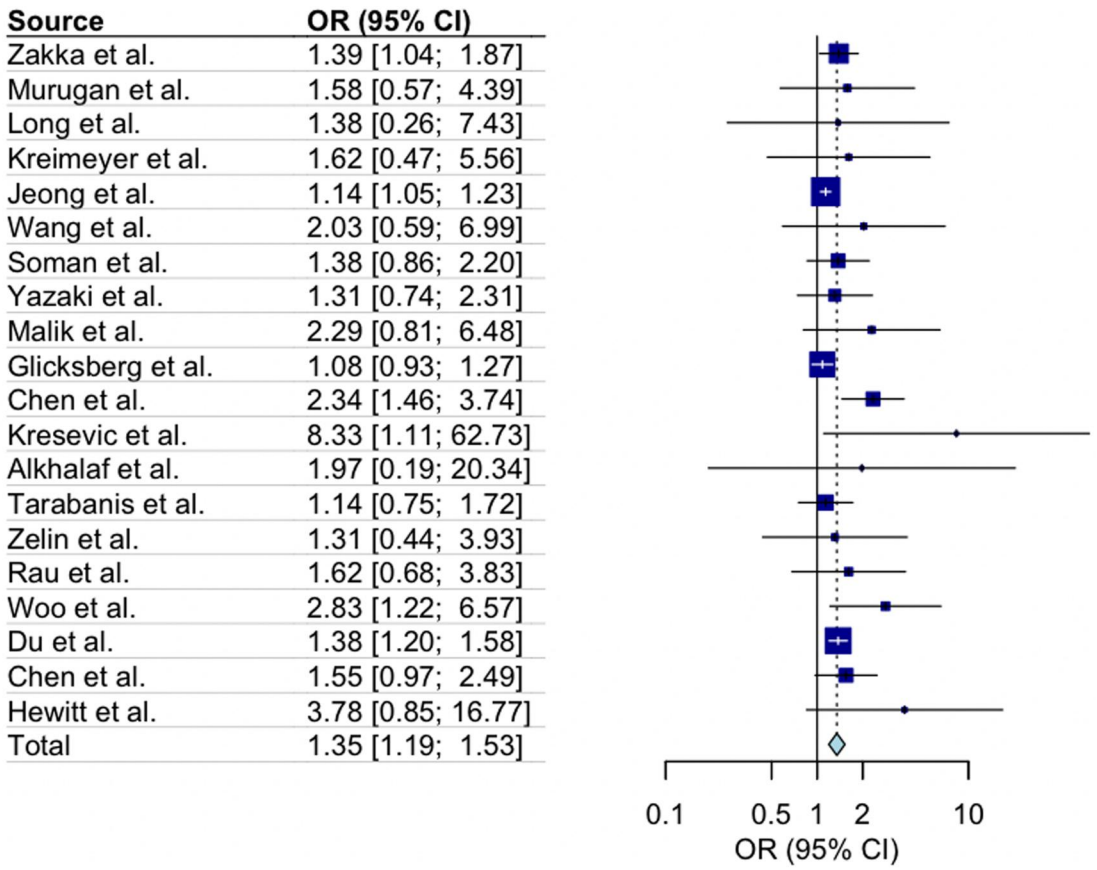

Figure 2. A forest plot showing the odds ratio (OR) of total impacts of the RAG-enhanced system compared with the baseline LLM system in clinical tasks. The left panel lists individual studies (e.g., Zakka et al., Murugan et al., Long et al.) along with their corresponding ORs and 95% confidence intervals (CIs). The right panel visualizes these ORs, with square markers proportional to study weights and horizontal lines representing CIs. The overall pooled OR is 1.35 (95% CI: 1.19–1.53), represented by a diamond at the bottom.

Table 1. Distribution of RAG applications by medical specialty and task type.

Medical Specialty	Frequency
Internal medicine	4
General medicine	3
Oncology	3
Emergency medicine	2
Gastroenterology	2
Otolaryngology	1
Hepatology	1
Rare diseases	1
Orthopedics	1
Neurology	1
Ophthalmology	1
Task	Frequency (%)
Clinical decision-making	13 (65%)
Medical question-answering	7 (35%)

Clinical applications of RAG

RAG techniques have been applied across a broad range of medical specialties, as shown in Table 1. These applications include clinical decision-making and medical question-answering. In clinical decision making, RAG has supported personalized treatment,[23,24] emergency triage,[25] and disease management.[26,27] For medical question-answering, RAG's capability has been explored to address complex treatment guidelines questions,[28] as well as queries focused on specific areas, such as head and neck surgery-related questions,[29] and patient questions regarding diabetes.[30] In the subgroup analysis, 13 studies focused on clinical decision-making (OR 1.46, 95% CI [1.16, 1.71]) and 7 studies focused on medical question-answering (OR 1.32, 95% CI [1.08, 1.63]), with no statistically significant difference observed between these 2 groups.

Baseline LLMs

The baseline LLMs varied across studies, with GPT-4 being the most common, used in 14 studies, (OR: 1.58, 95% CI: 1.21-2.04). GPT-3.5, used in 6 studies, showed an OR of 1.43 (95% CI: 1.06-1.93). Llama2 was applied in 5 studies (OR: 1.25, 95% CI: 1.08-1.44).

Retrieval sources

Retrieval sources were categorized as pre-stored documents and real-time online browsing. Regarding pre-stored documents, 6 studies used clinical guidelines, such as the Emergency Severity Index (ESI) Ver.3 Field Triage.[25] Five studies used academic articles from sources like PubMed abstracts or full texts, or document sets such as the *Radiographics Top 10 Reading List on Gastrointestinal Imaging*.[31] Three studies used specialized knowledge bases, including ChatENT, OncoKB, and RareDis Corpus, while one study employed a general biomedical knowledge graph (Scalable Precision Medicine Open Knowledge Engine [SPOKE]). SPOKE

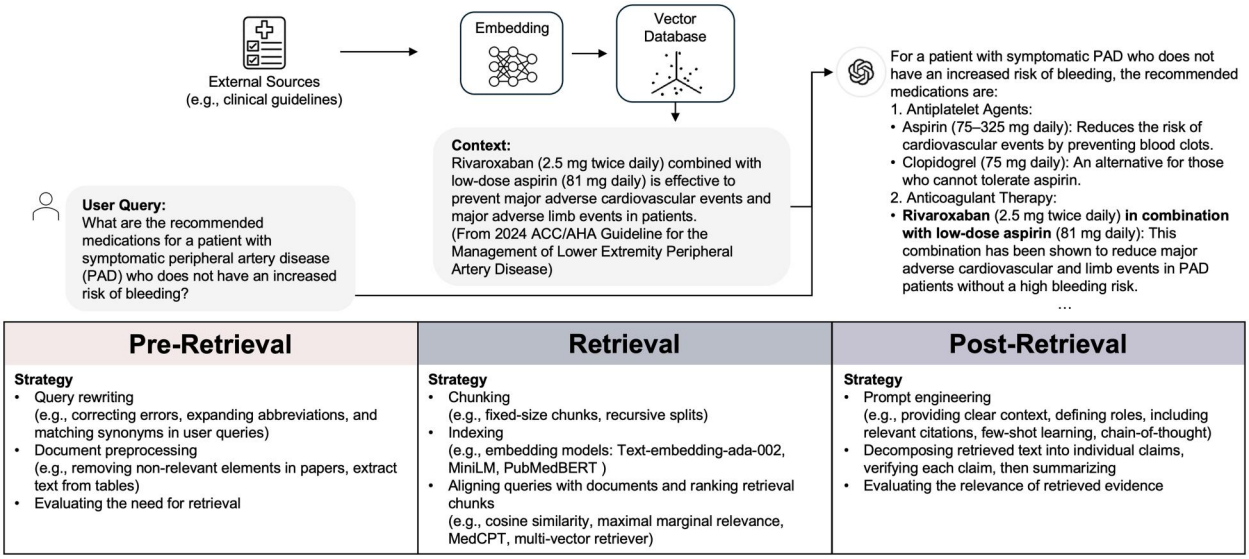

Figure 3. An example of using RAG in clinical applications, with identified strategies in 3 stages: pre-retrieval, retrieval, and post-retrieval. The user query seeks recommended medications for symptomatic peripheral artery disease without increased bleeding risk. The system retrieves evidence from clinical guidelines, processes it through embeddings and a vector database, and outputs a response, including rivaroxaban with low-dose aspirin, as recommended by the retrieved guideline information. In this example, GPT-4 suggested a dose of 75 mg Aspirin, but the common low-dose Aspirin is 81 mg.

integrates over 40 publicly available biomedical knowledge sources across separate domains, such as genes, proteins, drugs, compounds, and diseases, along with their known relationships.[32] Two studies used textbooks, such as *Harrison's Principles of Internal Medicine*, while 3 others utilized electronic health record (EHR) data. Additionally, Zakka et al. added over 500 markdown files from MDCalc to improve clinical calculation capabilities in LLM.[28] Two studies employed real-time online browsing to search academic sites, such as PubMed and UpToDate. The amount of retrieval resources varied across studies, ranging from a small dataset specific to 6 osteoarthritis guidelines to a large dataset of EHR data from 7 hospitals.

Retrieval strategies

Identified retrieval strategies were grouped based on the RAG stages: pre-retrieval, retrieval, and post-retrieval. Figure 3 presents an example of how RAG is applied and lists identified strategies within each stage.

In the pre-retrieval stage, 50% of studies ($n = 10$) reported strategies, such as query rewriting, document preprocessing, and assessing the necessity of retrieval. Zakka et al. simplified queries by rephrasing text into search terms that are better suited for website browsing,[28] while Wang et al. focused on techniques such as correcting errors, expanding abbreviations, and matching synonyms in user queries.[30] Soman et al. extracted disease entities in queries and retrieved corresponding nodes from a knowledge graph.[33] Document preprocessing involved removing non-textual elements from PMC papers (eg, figures, references, and author disclosures),[30] extracted tables from PDFs using pdfplumber, structured the content with pydantic for seamless integration.[25] In addition to query modification and document preprocessing, Jeong et al. fine-tuned a model to determine whether retrieval was necessary for a given query.[34]

During the data retrieval stage, 85% of studies ($n = 17$) reported strategies regarding indexing, aligning queries with documents, and ranking retrieval chunks. Chunking methods ranged from fixed-size chunks[35] to recursive splits.[36] Embedding models such as Text-embedding-ada-002,[24,28–30,36,37] MiniLM, and PubMedBERT[33] were commonly used to convert sentences into vectors. Cosine similarity was the primary metric for measuring query-document alignment. Two studies adopted Maximal Marginal Relevance for search and highlighted its improved performance over similarity-based methods.[24,35] A domain-specific retriever, MedCPT, was used in one study.[34] Another study used the multi-vector retriever that leveraged summarized document sections to identify the original content for final answer generation.[25] The retrieval cutoff parameters varied widely, with probability thresholds up to 0.83 and the number of retrieved chunks ranging from 3 to 90.[28,36,38] Vector databases like FAISS and Chroma were frequently reported, and LangChain was widely used for document processing and retrieval.[23,25,35,38] In the subgroup analysis, 12 studies used simple data retrieval strategies (OR 1.30, 95% CI [1.16, 1.45]), while 5 studies used complex data retrieval strategies (OR 1.30, 95% CI [1.07, 1.24]), with no statistically significant difference observed between the 2 approaches.

In the post-retrieval stage, 65% of studies ($n = 13$) implemented specific strategies to refine outputs. Murugan et al. tailored prompts by providing clear context, defining roles (eg, distinguishing between healthcare providers and patients to deliver appropriately detailed information), and incorporating relevant citations from retrieval sources such as the Clinical Pharmacogenetics Implementation Consortium guidelines and Food and Drug Administration (FDA) labeling.[24] Soman et al. utilized prompt engineering to integrate accurate knowledge sources and statistical evidence, such as *P*-values and *z*-scores, from the SPOKE knowledge graph into their outputs.[33] Wang et al. outlined a detailed process in the post-retrieval stage using prompt engineering, which involved decomposing retrieved text into individual claims, verifying each claim with external knowledge sources,

conducting safety checks by applying 24 predefined rules to ensure ethical and factual accuracy, and summarizing the results.[30] Glicksberg et al. developed an ensemble model that combined structured and unstructured data to predict hospital admission probabilities. These predicted probabilities, along with similar historical cases, were incorporated into the prompt to enhance the performance of LLM.[37] Chen et al. used Chain-of-Thought (CoT) prompting to improve LLM reasoning capabilities.[39] Kresevic et al. customized prompts to help the model interpret structured guidelines, combined with few-shot learning using 54 question-answer pairs.[27] Jeong et al. fine-tuned LLMs to assess the relevance of retrieved evidence, ensure all statements were evidence-based, and confirm that the response effectively addressed the query.[34]

Evaluation

Nine studies used human evaluation, 8 relied on automated evaluation (eg, similarity comparisons between generated sentences and original answers), and 3 used a mix of both. Outcomes from human evaluation showed an overall OR of 1.65 (95% CI: 1.36-2.03), while automatic evaluation resulted in an OR of 1.20 (95% CI: 1.1-1.41). The differences between the 2 were statistically significant ($P < .01$). There were 4 human evaluators on average, with the range spanning from 1 to 10. Most human evaluators were physicians from relevant specialties according to the study focus. In one case, 3 diabetic patients were involved in evaluating the understandability of diabetes-related patient queries.[30]

Twelve studies used self-curated datasets focused on research tasks. Examples included the ClinicalQA benchmark, which comprised 314 open-ended questions about treatment guidelines and clinical calculations generated by physicians,[28] and 43 diabetes-related questions sourced from the National Institute of Diabetes and Digestive and Kidney Diseases website.[30] Simulated cases from medical examinations were also utilized.[25] Three studies used EHR data.[35,37,40] Six studies used public benchmark datasets, such US board exam practice questions, MedMCQA[29,34] and longform question-answering benchmarks (eg, LiveQA, MedicationQA).[34] The self-curated datasets averaged 76 questions, ranging from 7 to 314. The length of public benchmark datasets varied significantly, from 102 questions in the LiveQA dataset[28] to 194 000 questions in the MedMCQA dataset.[34]

Most studies reported evaluation metrics for the final response generation, while 4 (25%) also included specific metrics to evaluate the retrieval process. For instance, 1 study measured recall in context retrieval,[24] another evaluated retrieval accuracy,[33] and a fine-tuned LLM was developed to assess the relevance of retrieved information to the user's query.[34] Additionally, 1 study evaluated the accuracy of using LLMs to extract text from figures and tables during document preprocessing.[27] The final evaluation metrics focused on the generated responses, consistent with those used in LLM-only systems. These metrics could be categorized as accuracy, completeness, user perception, safety, hallucination, citation, bias, and language. Accuracy was the most frequently reported metric, covering Likert scale ratings, match rates, correct treatment percentages,[9] AUC, AUPRC, and F1 scores, as well as text similarity metrics like ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), BLEU, METEOR, and BERTScore,[21] which compared LLM-generated responses to expert-provided answers. Completeness metrics assessed whether responses included all necessary information, typically using Likert scales. User perception captured subjective feedback from both healthcare providers and patients on understandability, helpfulness, and whether responses met user intent, usually using Likert scales. Safety metrics focused both on user-related and system-related aspects. These metrics assessed potential harm, adversarial safety, and risk management,[24] ensuring that outputs were free of harmful content or risks. Scientific validity and adherence to evidence were also evaluated.[29] One study used adversarial prompting, defined as intentionally adding harmful directives to a prompt, to evaluate the safety of the RAG system.[28] Hallucinations were primarily identified through manual review, with definitions varying across studies. Some studies defined hallucinations as nonfactual information, while one study added 2 other types of hallucinations: input-conflicting (content deviating from user-provided input) and contextual-conflicting (content conflicting with previously generated information).[27,41] Citation metrics measured the accuracy of provided references, with valid references considered those that pointed to established publications, guidelines, or research. Bias and language were evaluated for clarity and neutrality, ensuring responses were unbiased and empathetic to patient concerns.[24]

Discussion

This study presents a systematic review of current research on RAG for clinical tasks. Overall, RAG implementation increased outcomes by 1.35 times compared to baseline LLM. We analyzed clinical tasks, baseline LLMs, retrieval sources and strategies, as well as evaluation methods. Despite the potential benefits of RAG systems, there remains room for improvement. Building on our literature review, we developed GUIDE-RAG (**G**uidelines for **U**nified **I**mplementation and **D**evelopment of **E**nhanced LLM Applications with **RAG** in Clinical Settings) for future clinical applications using RAG (Figure 4).

GUIDE-RAG:

1) Define clear clinical tasks and evaluation datasets.
   Future research should clearly define clinical tasks and questions to maximize the effectiveness of RAGs. Ambiguity in questions can hinder performance, particularly in less powerful LLMs, making it challenging to achieve significant improvements in responses generation, even with improved knowledge selection.[42] For example, one study in the review constructed the evaluation dataset using 30 case reports on rare diseases from PubMed.[38] The authors did not report human performance on the self-build dataset. The questions themselves might have been inherently challenging or ambiguous. As expected, the reported performance showed modest improvement, with an OR of 1.31.
2) Identify appropriate external resources for specific clinical tasks.
   The first step in developing a RAG-based clinical system is to identify external resources that fill the knowledge gaps of the baseline LLM in relation to specific clinical tasks. The external knowledge should complement the LLM's existing capabilities to effectively address task

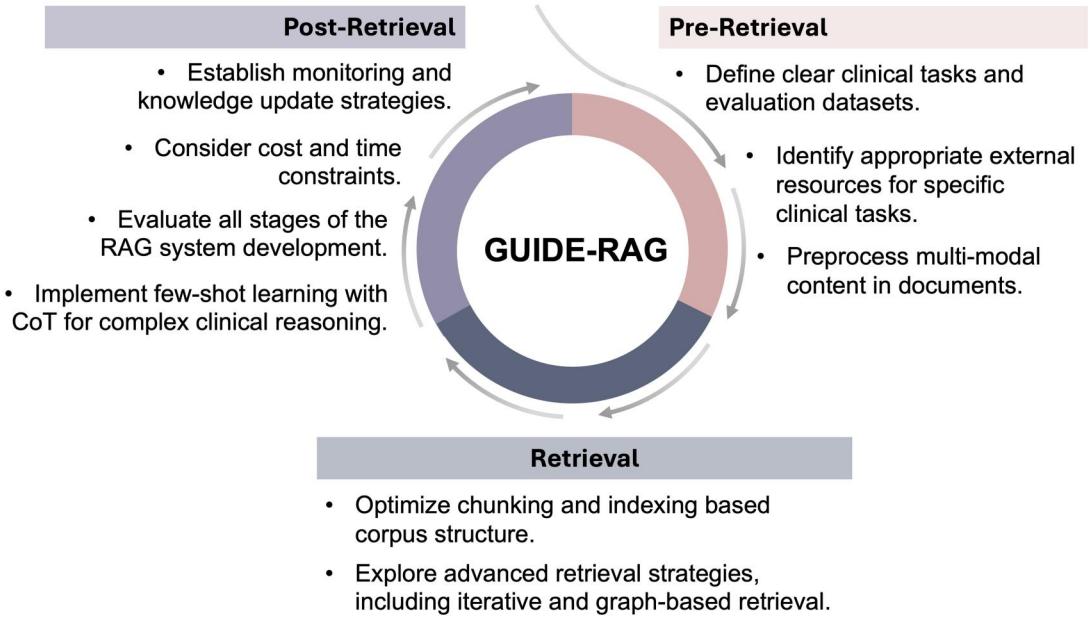

Figure 4. Overview of GUIDE-RAG: This framework streamlines RAG in clinical applications through three iterative stages. In the pre-retrieval stage, it focuses on defining tasks, identifying relevant resources, and preprocessing content. The retrieval stage enhances data retrieval with optimized chunking, indexing, and advanced strategies such as graph-based retrieval. The post-retrieval stage emphasizes system evaluation, monitoring, knowledge updates, and implementing few-shot learning for complex clinical reasoning, ensuring robust and adaptive performance. CoT, chain-of-thought.

requirements. For instance, in question-answering tasks related to broad medical exams for physicians, clinical guidelines (eg, StatPearls) and textbooks proved more useful than PubMed abstracts as external sources.[43] Another example from our review involved a task focused on medical question-answering in internal medicine. The study used a single source—Harrison's Principles of Internal Medicine—as the knowledge retrieval source, and the reported improvement was marginal (OR: 1.14).[36] Expanding the knowledge base to include additional resources, such as clinical guidelines, could potentially enhance the performance of the RAG system for such tasks.

3) Preprocess multi-modal content in documents.
Clinical guidelines and medical literature often contain complex information presented through flowcharts, graphs, and tables. Accurately parsing this multi-modal content is essential for effective retrieval. Relying solely on LLMs for text extraction may be insufficient; a preliminary study found that GPT-4 Turbo had only a 16% accuracy rate in extracting table data.[27] Comprehensive document preprocessing should systematically extract relevant information from text, tables, and figures to ensure accuracy and clarity. Only 3 studies in our review explicitly mentioned extracting text from tables or figures during the pre-retrieval process.[25,27,44]

4) Optimize chunking and indexing based corpus structure.
The structure of clinical knowledge corpora should be carefully considered during chunking and indexing. Fixed-length chunking can introduce noise by fragmenting related information, which can reduce retrieval accuracy. Researchers should optimize the chunking granularity based on a thorough review of the clinical knowledge corpus, to ensure the completeness of retrieved information. An alternative approach is dynamic chunking, which adjusts chunk boundaries based on semantic similarity changes.[45] Other approaches include recursive chunking, which hierarchically divides text into smaller chunks using delimiters like headings, subheadings, paragraphs, and sentences. Sliding window chunking enables layered retrieval by overlapping chunks of text, allowing the system to capture and merge contextually related information across different segments.[46] Context enriched chucking enhances retrieval by incorporating concise summaries within each segment to provide additional context for downstream tasks.[47] In indexing, while dense indexing (converting text to vectors) is widely used, it may miss global information. The structure of a clinical knowledge corpora such as some headings, keywords, can be used as sparse indexing and further combined with dense indexing. This hybrid approach that combines dense and sparse indexing can improve retrieval performance by capturing both global and local information.[48,49]

5) Explore advanced retrieval strategies, including iterative and graph-based retrieval.
Iterative retrieval improves accuracy by refining results through multiple rounds. Parameters such as the number of retrieved chunks or cutoff thresholds should be optimized based on specific clinical questions, as retrieval needs can vary—some questions may not require external knowledge at all. Researchers should evaluate retrieval requirements in advance and adapt retrieval parameters accordingly. Graph-based retrieval, which structures entities and relationships into a graph, can improve information synthesis from multiple sources. For example, GraphRAG identified entities and relationships from documents and developed a graph using LLM. Then, they used clustering algorithm to

offer global information based on user query, offering better performance than naïve RAG on the traditional vector databases.[50]

6) Implement few-shot learning with CoT for complex clinical reasoning.
   Few-shot learning has been shown to enhance LLMs' reasoning capabilities by teaching specific reasoning that may not have been included in their original training. Similarly, CoT techniques can improve complex reasoning in clinical tasks.[51,52] Researchers should generate high-quality examples and incorporate CoT strategies into the final query to refine specialized reasoning.
7) Evaluate all stages of the RAG system development.
   Most current studies focus only on final performance, overlooking the importance of evaluating each stage of development. It is crucial to formally assess and report performance at the pre-retrieval, retrieval, and post-retrieval stages. Evaluating the knowledge boundaries of the baseline LLM, potential conflicts between the LLM and external knowledge, and the accuracy and coverage of retrieved information helps ensure replicability and transparency. This level of evaluation enables other researchers to understand why a RAG system works (or does not) and facilitates reproducibility.
8) Consider cost and time constraints.
   Advanced retrieval strategies can improve performance but often increase processing time and computational costs. For example, graph-based RAG requires substantial resources for developing knowledge graphs, and responses from global summaries may take longer than with naïve RAG methods.[50] Another example is to fine-tune LLMs to evaluate the needs and the quality of retrieval.[34] In terms of computational cost, this process is expensive, especially when scaling the method to larger datasets or deploying it in a real-time system. Also, a set of extra processes will make the whole speed slow. The long response time might have a nonignorable negative impact in situations that need a quick answer, especially common in clinical settings. Researchers should balance performance improvements with time and cost considerations. Only 1 study in our review, which focused on gastrointestinal radiology diagnosis based on imaging descriptions, compared the cost and response time between LLMs and LLMs with RAG.[31] The mean response time was 29.8 s for LLM with RAG vs 15.7s for LLM alone, with costs of $0.15 and $0.02 per case, respectively. Another study used EHR to predict cognitive decline only reported cost, with LLM: $4.49; RAG: $12.51. Another study that used EHR data to predict cognitive decline reported costs of $4.49 for LLM alone and $12.51 for LLM with RAG.[53]
9) Establish monitoring and knowledge update strategies.
   An important concept in AI applications in healthcare, algorithmovigilance, which defined as "scientific methods and activities relating to the evaluation, monitoring, understanding, and prevention of adverse effects of algorithms in health care," [54] should also be considered in the RAG applications. Researchers need to develop long-term monitoring strategies for the RAG system performance, especially in clinical applications. In addition, current studies use fixed external datasets. Researchers should update external knowledge sources as latest information becomes available. Clear strategies for updating knowledge should be defined, specifying when and how updates will occur.

For future studies, the first direction could be the system-level enhancement, the combination of RAG and LLM-powered agents. LLM-powered agents are AI systems that use LLMs with complex reasoning and planning capabilities, memory management, interactive capabilities with the environment, and actions to execute tasks.[55,56] Recent research points to the emerging trend of combination of RAG and LLM-powered agents, where agents can assist in planning and decision making for complex tasks, rather than simple retrieval.[57] For example, clinicians and patients have diverse information access needs, some needing to analyze text from a knowledge base, others needing to incorporate structured data from an EHR. RAG will eventually only become one of the methods for agents to access information. Moreover, future research could focus on the usage of internal and external functions and tools, long-term and short-term memory module, self-learning module. For example, a study developed an agent to answer questions related to rare diseases by expanding beyond RAG with additional tool functions, such as querying phenotypes and performing web searches. This approach improved the overall correctness from 0.48 to 0.75 compared to the GPT-4 baseline LLM.[58]

The second future direction could focus on the knowledge-level enhancement: deep integration of external knowledge into LLM. LLM exhibits the knowledge boundaries. RAG approaches retrieving external knowledge and then integrates it into LLMs in the forms of prompts for the final generation to enhance the capabilities of LLMs in perceiving knowledge boundaries.[59] However, the integration of external knowledge into LLM reasoning is typically limited to providing the retrieved data as additional context for the LLM's query during generation. This approach keeps retrieval and generation loosely connected, and the LLM's output can still be influenced by its inherent knowledge boundaries or by noise in the retrieved text, leading to incorrect answers. Additionally, when the external knowledge source is EHR data, this enhancement becomes even more important. Current EHR data is organized in a "problem-oriented medical record" (POMR) format, which collects and displays information in a structured manner.[60] LLMs excel in free-form contexts, and their ability to perform clinical tasks depends on access to unstructured text that provides a comprehensive view of the patient. Achieving this within the structured POMR format in modern EHR systems poses a significant challenge.[61] Therefore, investigating how to realize the deep integration of external knowledge with LLM reasoning is an important direction for future research in clinical applications.

The final direction is the integration-level enhancement, focusing on integrating RAG systems within EHRs. Current research has primarily focused on development and testing outside of EHR systems. To seamlessly provide support for healthcare providers and patients, future efforts should prioritize embedding RAG systems into EHR interfaces. This requires collaboration with EHR vendors to ensure the necessary infrastructure is available. Researchers also can facilitate this integration using data exchange frameworks, such as SMART on FHIR.[62]

Limitations

This study was limited to peer-reviewed publications available in biomedical databases (eg, PubMed, Embase), excluding preprint articles from repositories like ArXiv. Additionally, only studies in English language were included, which might have excluded relevant studies in other languages. We did not include sources such as IEEE Xplore or Google Scholar, which might have additional relevant studies. However, our focus was on biomedicine, and we prioritized databases specifically tailored to biomedical research to maintain the relevance and quality of the included studies. Furthermore, we used free-text searches in the databases, which activated automatic mapping to Medical Subject Headings (MeSH) and Emtree terms, improving retrieval accuracy. However, the limitations of automatic term mapping cannot be ignored, as it may introduce variability if the underlying algorithms change. To address this, we have documented all identified papers from our search. The title, publication year, PMID, PUI, and database source for each study are provided in File S3.

Conclusion

We conducted a systematic literature review of studies exploring the use of RAG and LLM in clinical tasks. RAG implementation showed a 1.35 odds ratio increase in performance compared to baseline LLMs. To improve performance and transparency in future studies, we developed guidelines for improving clinical RAG applications based on current research findings. Future research could focus on these 3 directions: (1) system-level enhancement: the combination of RAG and agent, (2) knowledge-level enhancements: deep integration of knowledge into LLM, and (3) integration-level enhancements: integrating RAG systems within EHRs.

Author contributions

Siru Liu (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft), Allison B. McCoy (Conceptualization, Writing – review & editing), Adam Wright (Conceptualization, Writing – review & editing).

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by National Institutes of Health grants: R00LM014097-02 and R01LM013995-01.

Conflicts of interest

The authors do not have conflicts of interest related to this study.

Data availability

The characteristics and outcomes for each included study were reported in the File S2.

References

1. Raiaan MAK, Mukta MSH, Fatema K, et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*. 2024;12:26839-26874. https://doi.org/10.1109/ACCESS.2024.3365742
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930-1940. https://doi.org/10.1038/s41591-023-02448-8
3. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589-596. https://doi.org/10.1001/jamainternmed.2023.1838
4. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30:1237-1245. https://doi.org/10.1093/jamia/ocad072
5. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw Open*. 2024;7:e240357. https://doi.org/10.1001/jamanetworkopen.2024.0357
6. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2312.10997
7. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. January 22, 2024, preprint: not peer reviewed. https://arxiv.org/abs/2401.11817
8. Shuster K, Poff S, Chen M, et al. Retrieval augmentation reduces hallucination in conversation. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*. Association for Computational Linguistics (ACL); 2021:3784-3803.
9. Ayala O, Bechard P. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2024;228-38. https://doi.org/10.18653/v1/2024.naacl-industry.19
10. Gornik HL, Aronow HD, Goodney PP, et al. 2024 ACC/AHA/AACVPR/APMA/ABC/SCAI/SVM/SVN/SVS/SIR/VESS guideline for the management of lower extremity peripheral artery disease: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. 2024;149:e1313-e1410. https://doi.org/10.1161/CIR.0000000000001251
11. 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*. 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904
12. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029. https://doi.org/10.1148/ryai.2020200029
13. Martindale APL, Ng B, Ngai V, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. *Nat Commun*. 2024;15:6376-6311. https://doi.org/10.1038/s41467-024-45355-3
14. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7:258. https://doi.org/10.1038/s41746-024-01258-7
15. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015: elaboration and explanation. *BMJ*. 2015;350:g7647. https://doi.org/10.1136/bmj.g7647
16. Higgins JPT, Thomas J, Chandler J, et al. (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Cochrane, 2024. Available from www.training.cochrane.org/handbook. Date accessed December 11, 2024.

17. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: searching for and selecting studies. *Cochrane Handbook for Systematic Reviews of Interventions Version, Vol.* 6. Cochrane, 2024. https://training.cochrane.org/handbook/current/chapter-04
18. Chapter 3 Effect Sizes | Doing Meta-Analysis in R. Accessed October 13, 2024. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/effects.html
19. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. John Wiley & Sons; 2011.
20. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558. https://doi.org/10.1002/sim.1186
21. Peters JL, Sutton AJ, Jones DR, et al. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol*. 2008;61:991-996. https://doi.org/10.1016/j.jclinepi.2007.11.010
22. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test measures of funnel plot asymmetry. *BMJ*. 1997;315:629-634. https://doi.org/10.1136/bmj.315.7109.629
23. Kreimeyer K, Canzoniero JV, Fatteh M, et al. Using retrieval-augmented generation to capture molecularly-driven treatment relationships for precision oncology. *Stud Health Technol Inform*. 2024;316:983-987. https://doi.org/10.3233/SHTI240575
24. Murugan M, Yuan B, Venner E, et al. Empowering personalized pharmacogenomics with generative AI solutions. *J Am Med Inform Assoc*. 2024;31:1356-1366. https://doi.org/10.1093/jamia/ocae039
25. Yazaki M, Maki S, Furuya T, et al. Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model. *Prehosp Emerg Care*. 2024;1-7. https://doi.org/10.1080/10903127.2024.2374400
26. Malik S, Kharel H, Dahiya DS, et al. Assessing ChatGPT4 with and without retrieval-augmented generation in anticoagulation management for gastrointestinal procedures. *Ann Gastroenterol*. 2024;37:514-526. https://doi.org/10.20524/aog.2024.0907
27. Kresevic S, Giuffrè M, Ajcevic M, et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med*. 2024;7:102-109. https://doi.org/10.1038/s41746-024-01091-y
28. Zakka C, Shad R, Chaurasia A, et al. Almanac–retrieval-augmented language models for clinical medicine. *NEJM AI*. 2024;1(2):10.1056/aioa2300068. https://doi.org/10.1056/aioa2300068
29. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology–head and neck surgery. *Otolaryngol Head Neck Surg*. 2024;171:1042-1051. https://doi.org/10.1002/ohn.864
30. 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*. 2024;26:e58041. https://doi.org/10.2196/58041
31. Rau S, Rau A, Nattenmüller J, et al. A retrieval-augmented chatbot based on GPT-4 provides appropriate differential diagnosis in gastrointestinal radiology: a proof of concept study. *Eur Radiol Exp*. 2024;8:60. https://doi.org/10.1186/s41747-024-00457-x
32. Morris JH, Soman K, Akbas RE, et al. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*. 2023;39(2):btad080. https://doi.org/10.1093/BIOINFORMATICS/BTAD080
33. Soman K, Rose PW, Morris JH, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Commun ACM*. 2023;66:7-7. https://doi.org/10.1145/3606337
34. Jeong M, Sohn J, Sung M, et al. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*. 2024;40:i119-i129. https://doi.org/10.1093/bioinformatics/btae238
35. Alkhalaf M, Yu P, Yin M, et al. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J Biomed Inform*. 2024;156:104662. https://doi.org/10.1016/j.jbi.2024.104662
36. Tarabanis C, Zahid S, Mamalis M, et al. Performance of publicly available large language models on internal medicine board-style questions. *PLOS Digit Health*. 2024;3:e0000604. https://doi.org/10.1371/journal.pdig.0000604
37. Glicksberg BS, Timsina P, Patel D, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *J Am Med Inform Assoc*. 2024;31:1921-1928. https://doi.org/10.1093/jamia/ocae103
38. Zelin C, Chung WK, Jeanne M, et al. Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT. *J Biomed Inform*. 2024;157:104702. https://doi.org/10.1016/j.jbi.2024.104702
39. Chen X, Wang L, You MK, et al. Evaluating and enhancing large language models' performance in domain-specific medicine: development and usability study with DocOA. *J Med Internet Res*. 2024;26:e58158. https://doi.org/10.2196/58158
40. Du X, Novoa-Laurentiev J, Plasaek JM, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *medRxiv*. 2024;109:105401. https://doi.org/10.1101/2024.04.03.24305298
41. Zhang Y, Li Y, Cui L, et al. Siren's song in the AI Ocean: a survey on hallucination in large language models. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2309.01219
42. Li X, Ouyang JA. Systematic investigation of knowledge retrieval and selection for retrieval augmented generation. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2410.13258
43. Xiong G, Jin Q, Lu Z, et al. Benchmarking retrieval-augmented generation for medicine. *Findings of the Association for Computational Linguistics: ACL 2024*, 6233-6251. Bangkok, Thailand: Association for Computational Linguistics.
44. Hewitt KJ, Wiest IC, Carrero ZI, et al. Large language models as a diagnostic support tool in neuropathology. *J Pathol Clin Res*. 2024;10:e70009. https://doi.org/10.1002/2056-4538.70009
45. Allahverdiyev R, Taha M, Akalin A, et al. ChunkRAG: novel LLM-chunk filtering method for RAG systems. October 25, 2024, preprint: not peer reviewed. https://arxiv.org/abs/2410.19572
46. Cai B, Zhang FL, Wang C. Research on chunking algorithms of data de-duplication. *Advances in Intelligent Systems and Computing, Vol.* 181. 2013:1019-1025. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31698-2_144
47. Optimizing RAG with Advanced Chunking Techniques. Accessed December 15, 2024. https://antematter.io/blogs/optimizing-rag-advanced-chunking-techniques-study
48. Chen J, Xiao S, Zhang P, et al. M3-Embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Findings of the Association for Computational Linguistics ACL 2024*. 2024:2318-2335. Association for Computational Linguistics. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.137
49. Sawarkar K, Mangal A, Solanki SR. Blended RAG: improving RAG (Retriever-Augmented Generation) accuracy with semantic search and hybrid query-based retrievers. *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 155-161. San Jose, CA, USA: IEEE. https://doi.org/10.1109/MIPR62202.2024.00031
50. Edge D, Trinh H, Cheng N, et al. From local to global: a graph RAG approach to query-focused summarization. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2404.16130
51. Wu Z, Hasan A, Wu J, et al. KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-Though (CoT) prompting strategies for medical error detection and correction. *Proceedings of the 6th Clinical Natural Language Processing Workshop*. 2024:353-359. Association for Computational Linguistics. 10.18653/v1/2024.clinicalnlp-1.33
52. Kwon T, Tzu-Iunn Ong K, Kang D, et al. Large language models are clinical reasoners: reasoning-aware diagnosis framework with

prompt-generated rationales. *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence; 2024:18417-18425.
53. Du X, Novoa-Laurentiev J, Plasek JM, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *EBioMedicine*. 2024;109:105401. https://doi.org/10.1016/j.ebiom.2024.105401
54. Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence–driven health care for effectiveness and equity. *JAMA Netw Open*. 2021;4:e214622. https://doi.org/10.1001/jamanetworkopen.2021.4622
55. Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey. September 14, 2023, preprint: not peer reviewed.
56. Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. *Front Comput Sci*. 2024;18:1-26. https://doi.org/10.1007/S11704-024-40231-1/METRICS
57. Li X, Wang S, Zeng S, et al. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*. 2024;1:9. https://doi.org/10.1007/s44336-024-00009-2
58. Yang J, Shu L, Duan H, et al. RDguru: a conversational intelligent agent for rare diseases. *IEEE J Biomed Health Inform*. Published online September 19, 2024. https://doi.org/10.1109/JBHI.2024.3464555
59. Ren R, Wang Y, Qu Y, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2307.11019
60. Weed LL. Medical records that guide and teach. *N Engl J Med*. 1968;278:593-600. https://doi.org/10.1056/NEJM196803142781105/ASSET/9EE62BDC-88EB-469C-BCDC-DB379C2CAE47/ASSETS/IMAGES/MEDIUM/NEJM196803142781105_F2.GIF
61. McCoy LG, Manrai AK, Rodman A. Large language models and the degradation of the medical record. *N Engl J Med*. 2024;391:1561-1564. https://doi.org/10.1056/NEJMP2405999
62. Mandel JC, Kreda DA, Mandl KD, et al. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc*. 2016;23:899-908. https://doi.org/10.1093/jamia/ocv189

Journal of the American Medical Informatics Association, 2025, 32, 605–615
https://doi.org/10.1093/jamia/ocaf008
Research and Applications