Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving Retrieval-Augmented Generation: A Systematic Literature Review of Retrieval-Enhancement Techniques

Abdullah Karasan

akarasan@umbc.edu

University of Maryland https://orcid.org/0009-0002-1651-7437

Systematic Review

Keywords: Retrieval-augmented generation, Large language models, Systematic literature review, Information retrieval, Knowledge graphs, Re-ranking

Posted Date: July 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10368581/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.

Improving Retrieval-Augmented Generation: A Systematic Literature Review of Retrieval-Enhancement Techniques

Abdullah Karasan[1,2*]

[1]University of Maryland, Baltimore County, Baltimore, MD, USA.
[2]Leveragai, United States.

Corresponding author(s). E-mail(s): akarasan@umbc.edu;

Abstract

Large language models are highly fluent and effective across a wide range of tasks but remain vulnerable to hallucination, knowledge staleness, and limited provenance of generated outputs. Retrieval-augmented generation addresses these limitations by grounding generation in externally retrieved evidence and has become a rapidly growing research area. However, advances are dispersed across pipeline stages, evaluated using heterogeneous metrics, and spread across multiple research communities, leaving the field without a comprehensive synthesis.

This paper presents a systematic literature review of 173 studies on RAG retrieval enhancement published between 2024 and 2026, identified through searches across six major bibliographic databases. Studies are categorised by (i) their primary technical contribution, (ii) application domain, and (iii) the RAG pipeline stage they target, including pre-retrieval, retrieval, post-retrieval, generation-integrated approaches, and cross-stage modular or agentic architectures. Representative and highly cited methods are analysed in detail while ensuring coverage of all included studies.

The review identifies four key findings: Uneven research effort across pipeline stages, fragmented evaluation practices, limited reporting of computational cost, and a strong focus on healthcare applications. Based on these findings, it proposes a structured research agenda to advance RAG systems toward more comparable, efficient, and generalisable methodologies.

Keywords: Retrieval-augmented generation, Large language models, Systematic literature review, Information retrieval, Knowledge graphs, Re-ranking

1 Introduction

Large language models (LLMs) have become general-purpose engines for knowledge-intensive language tasks, from open-domain question answering to clinical decision support [17]. Their performance, however, rests on knowledge compressed into parameters that are fixed at training time, and this design yields three well-documented failure modes. Firstly, LLMs hallucinate, producing plausible but unsupported statements, secondly, their knowledge becomes stale as the world changes after the training cut-off, finally, they provide no provenance, offering answers that cannot be traced to a verifiable source. These weaknesses are most consequential precisely where LLMs are most attractive in high-stakes, fast-moving, or specialised domains such as medicine, law, and finance, where an unsupported or outdated claim carries real cost.

Retrieval-augmented generation (RAG) has emerged as the dominant remedy. Rather than relying solely on parametric memory, a RAG system retrieves relevant evidence from an external knowledge at inference time and conditions generation on that evidence, grounding outputs in a knowledge source that is updatable and inspectable [72]. This simple idea decouples what the model knows from what it has learned to say, allowing knowledge to be corrected or refreshed without retraining and allowing answers to cite their sources. The paradigm has since matured from the basic retrieve-then-read pipeline into a richer design space, conventionally framed as a developmental phase from Naive, through Advanced, to Modular RAG [33].

The volume of research aimed specifically at improving RAG has grown explosively, and that effort has concentrated overwhelmingly on the retrieval component, the stage that most directly determines whether the generator receives the right evidence. Researchers have proposed structure-aware chunking and indexing, query reformulation, hybrid sparse, dense and knowledge-graph retrieval, multimodal retrieval, reranking and context refinement, retrieval-augmented reasoning, and agentic pipelines that retrieve iteratively under a control loop.Yet this progress is fragmented. Contributions intervene at different stages of the pipeline, are evaluated with different metrics on different datasets, and are reported in venues ranging from archival journals to preprints, so that it is difficult to see which techniques represent genuine advances, how they relate to one another, or where the field's effort is concentrated and where it is thin.

In order to remedy these issues, we conduct a systematic literature review, following the PRISMA reporting standard [100] and established software-engineering review guidelines [57], of techniques proposed to enhance the retrieval component of RAG. From a transparent and reproducible search-and-screening process we assemble a corpus of 173 studies published between January 2024 and March 2026, classify each study along two axes, the technical contribution and, where applicable, the application domain, and synthesise them by the stage of the RAG pipeline at which they intervene. Representative, highly cited studies are analysed in depth, with their mechanisms, reported quantitative gains, and stated limitations examined, while the remaining studies are grouped by technical approach so that every included work is accounted for.

Concretely, the review is organised around three research questions:

1. What techniques have been proposed to enhance the retrieval component of RAG, and at which stage of the pipeline do they intervene?
2. How are the resulting improvements in retrieval accuracy, precision, and answer quality measured and evaluated?
3. What limitations and open challenges does the current literature reveal, and what research directions do they motivate?

In answering these questions, the paper makes four contributions. First, it provides a PRISMA-based corpus of 173 recent retrieval-enhancement studies, together with a two-axis, i.e. technique and domain classification that accounts for the full corpus. Second, it organises the literature into a stage-structured taxonomy as pre-retrieval, retrieval, post-retrieval, and generation-integrated enhancement, plus cross-stage modular architectures superimposed on the Naive - Advanced - Modular framing, giving a single map of where recent effort is concentrated. Third, it offers an in-depth synthesis anchored in concrete quantitative results and paired, for each technique family, with an explicit critique of its costs and limitations, rather than a catalogue of summaries. Fourth, it distils four cross-cutting findings into a concrete research agenda for the next phase of RAG research.

To support transparency and reproducibility of the search and screening process described in this section, the methodology was implemented as an open-source Python library, `litreview` (available at https://github.com/abdullahkarasan/litreview). The library programmatically queries multiple bibliographic APIs, applies the inclusion/exclusion criteria defined in Section 2.3, deduplicates records, and exports structured metadata for downstream screening and synthesis. Using a packaged, version-controlled tool rather than ad hoc scripts allows the search to be re-run, audited, and extended by other researchers, in line with calls for greater methodological transparency in AI literature reviews.

The remainder of the paper is organised as follows. Section 2 details the systematic review methodology, including the research questions, search strategy, inclusion and exclusion criteria, and the thematic classification procedure. Section 3 reports the results and thematic synthesis, answering research question 1 and 2. Section 4 discusses the cross-cutting findings and develops the research agenda in response to research question 3.

2 Methodology

This study adopts a systematic literature review (SLR) methodology to investigate retrieval enhancement techniques in RAG systems. The review protocol follows the PRISMA 2020 reporting guidelines [100] and draws on the foundational SLR framework established by Kitchenham et al. [57] for software engineering research. The methodology is structured into five sequential phases: (1) formulation of research questions, (2) search strategy design, (3) definition and application of inclusion and exclusion criteria, (4) data extraction, and (5) synthesis and analysis. Each phase is documented to ensure transparency, reproducibility, and methodological rigour, in accordance with best practices observed in highly cited AI survey and review articles [44, 33, 17].

2.1 Research questions

The review is guided by above-stated three research questions, formulated prior to the systematic search to delimit the scope of investigation and focus the analysis on RAG retrieval improvement specifically. These questions follow the PICO-style decomposition recommended by Kitchenham et al. [57], adapted here as Population (RAG systems), Intervention (retrieval enhancement techniques), Comparison (baseline RAG or parametric-only approaches), and Outcome (performance, accuracy, precision, or quality improvements).

These questions were designed to capture the breadth of proposed methods, the evaluative and benchmarking practices employed in the field , and the outstanding gaps and unresolved issues that may guide future research.

2.2 Search strategy

2.2.1 Data sources

To maximise coverage and minimise database-specific selection bias, six academic data sources were queried: Semantic Scholar, arXiv, Scopus (Elsevier), OpenAlex, IEEE Xplore, and the ACM Digital Library. This multi-source strategy aligns with established practice in literature. For instance, Holzinger et al. [44] employed IEEE Xplore, Scopus, DBLP, PubMed, Web of Science, and Google Scholar for their deep learning meta-survey, while Brown et al. [17] queried ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect, and DBLP for their PRISMA-compliant RAG systematic review. Each source was selected to capture a different stratum of the literature: Semantic Scholar provides broad multidisciplinary coverage with rich citation metadata; arXiv captures the substantial volume of AI and NLP preprints that precede or never enter formal peer review; Scopus offers curated, peer-reviewed journal and conference coverage; OpenAlex provides open bibliometric data spanning over 250 million scholarly works; and IEEE Xplore and ACM Digital Library represent the two principal computing-specific digital libraries.

2.2.2 Search queries

Search queries were tailored to each database's query syntax and retrieval capabilities while preserving semantic equivalence across sources. The core search concept combined two mandatory dimensions: (a) a domain concept, represented by the term "retrieval augmented generation" and its morphological variants, and (b) an improvement orientation, represented by the terms "enhance," "improve," "accuracy," "precision," "performance," and "quality." Table 1 presents the full search strings used for each data source.

For databases that support field-specific operators, the improvement terms were constrained to the title field. In Scopus, for example, the query was formulated as `TITLE("enhance" OR "improve" OR "accuracy" OR "precision") AND TITLE("retrieval augmented generation")`. For arXiv, the `ti:` prefix was applied to constrain matching to paper titles. For Semantic Scholar and OpenAlex,

Table 1 Search queries used for each data source.

Data source	Search query
Semantic Scholar	`retrieval augmented generation enhance improve accuracy precision`
arXiv	`ti:"enhance" AND ti:"retrieval augmented generation" OR ti:"improve" AND ti:"retrieval augmented generation"`
Scopus	`TITLE("enhance" OR "improve" OR "accuracy" OR "precision") AND TITLE("retrieval augmented generation")`
OpenAlex	`"retrieval augmented generation" enhance improve precision accuracy`
IEEE Xplore	`("retrieval augmented generation") AND ("enhance" OR "improve" OR "accuracy" OR "precision")`
ACM DL	`retrieval augmented generation improve enhance accuracy`

which provide simplified query interfaces, plain keyword strings combining both dimensions were used. The complete search strings for all six databases are reported in Table 1, in accordance with PRISMA 2020 Item 7, which recommends presenting full search strategies for every source searched [100].

2.2.3 Temporal scope

The search was restricted to papers published between January 2024 and March 2026. This temporal window was chosen deliberately to capture the most recent and rapidly expanding wave of RAG research. Although the foundational RAG architecture was introduced by Lewis et al. [72] in 2020, the field underwent a qualitative acceleration following the mainstreaming of large language models in late 2022 and early 2023. By 2024, RAG had matured from a foundational concept into a rapidly diversifying subfield with hundreds of published enhancement proposals, making a focused contemporary window both practical and analytically tractable.

2.2.4 Automation-assisted search

Given the high volume of candidate literature, the search process was partially automated using a custom Python script that programmatically queried the application programming interfaces (APIs) of each data source, collected results with pagination (up to 200 results per query per source) to avoid truncation, and performed initial title-based filtering. This automation-assisted approach is consistent with the broader trend toward computational support in systematic reviews, as documented by van de Schoot et al. [148] and others who have advocated for machine learning and natural language processing tools in literature screening. The script retrieved structured metadata for each candidate paper, including title, author list, publication year, venue, citation count, digital object identifier (DOI), URL, and abstract.

The search and filtering pipeline described above is implemented in the `litreview` Python package, which provides a reusable, configurable interface for API querying, pagination, deduplication, and criteria-based filtering.

2.2.5 Supplementary snowballing

To mitigate database indexing lag and ensure engagement with top-ranked journals, a supplementary backward and forward snowballing pass [156] was conducted over 2024-2026 proceedings of KDD, SIGIR, and TKDE. Studies identified this way are discussed as comparative baselines but are *not* counted in the 173-study corpus, whose composition is fixed by the protocol above. This preserves the reproducibility of the reported distributions while allowing the synthesis to engage with methodologically adjacent work that the title-based filter (IC1) excludes by construction.

2.3 Inclusion and exclusion criteria

Following Kitchenham et al.'s [57] recommendation that selection criteria should be derived directly from the research questions and piloted prior to full application, the inclusion and exclusion criteria presented in Table 2 were defined.

Table 2 Inclusion and exclusion criteria.

Type	Criterion
IC1	The paper's title must contain at least one improvement-oriented term ("enhance," "improve," "accuracy," "precision," "performance," or "quality").
IC2	The paper must substantively address retrieval-augmented generation, verified by the presence of "retrieval-augmented generation," "retrieval augmented generation," or a combination of "RAG" with "retriev-" or "generation" in the title or abstract.
IC3	The paper must be published or deposited between January 2024 and March 2026.
IC4	The paper must be written in English.
EC1	Papers that mention RAG only peripherally (e.g., in a related-work section) without proposing or evaluating a retrieval enhancement.
EC2	Duplicate records across databases, identified through title normalisation and merged programmatically.
EC3	Non-research documents such as editorials, workshop calls, tutorials without original contributions, or slide decks.
EC4	Papers with normalised titles shorter than ten characters, indicating malformed or incomplete records.

The title-based filtering strategy (IC1) warrants explicit justification. Requiring improvement-related terms in the title constitutes a more restrictive approach than

abstract- or full-text-level filtering and entails a deliberate trade-off: it significantly increases precision, ensuring that every included paper has retrieval enhancement as a central rather than incidental concern, at the cost of potentially excluding papers that propose genuine improvements but employ different terminology in their titles. This trade-off was judged acceptable for two reasons. First, the primary objective of this review is to survey papers that explicitly position themselves as contributing to RAG retrieval improvement. Second, the exceptionally high volume of RAG literature in the target period made a precision-oriented filter necessary to maintain analytical tractability.

2.4 Screening and selection process

The selection process proceeded in three phases. In the first phase, the automated search collected all candidate papers from each database, applying pagination to ensure comprehensive retrieval. In the second phase, the title filter (IC1) was applied programmatically: each paper's title was checked for the presence of at least one required improvement term, and non-matching papers were discarded. In the third phase, a relevance verification check confirmed that remaining papers addressed RAG substantively (IC2), filtering out papers that use the acronym "RAG" in an unrelated context (e.g., recombination-activating genes in immunology). After all three phases, deduplication (EC2) was performed across sources: paper titles were normalised by lowercasing and removing special characters, and records sharing a matching normalised key were merged into a single entry, retaining the version with the higher citation count and combining source metadata.

This multi-phase screening workflow aligns with the structure recommended by PRISMA 2020 and employed in recent RAG-specific reviews. Brown et al. [17], for example, used citation thresholds (≥ 15 citations for 2025 publications, ≥ 30 for earlier ones) as a primary gate in their PRISMA-compliant review. Our approach differs in that it prioritises topical precision over citation-based impact, but the structural logic, an automated first pass followed by criteria-based refinement, is analogous.

2.5 Data extraction

For each included paper, the following data fields were extracted: title, author list, publication year, venue (journal or conference), citation count at the time of retrieval, research category, source database(s) in which the paper appeared, DOI, URL, and a truncated abstract (first 500 characters). Citation counts were sourced primarily from Semantic Scholar, supplemented by Scopus and OpenAlex where available. These counts were used for descriptive profiling and ranking rather than as formal quality indicators, given that citation accumulation is time-sensitive and inherently favours older publications.

Extracted data were stored in both structured spreadsheet format (with auto-filtered columns for sorting and faceted browsing) and JSON format to facilitate subsequent computational analysis. Summary sheets aggregating paper counts by source and by thematic category were generated automatically.

2.6 Data synthesis

The extracted data were synthesised using a combination of descriptive statistical analysis and thematic categorisation. Descriptive analysis encompassed publication trends over time, distribution of papers across venues and data sources, and citation count distributions. Thematic categorisation grouped papers by their primary contribution type—such as query reformulation, chunk optimisation, re-ranking mechanisms, hybrid retrieval architectures, multi-step retrieval pipelines, knowledge-graph-augmented retrieval, or cross-encoder-based refinement—following an open coding procedure in which categories emerged inductively from the abstracts and were iteratively refined through multiple passes over the corpus.

For each thematic category, representative papers were selected and analysed in depth to illustrate the technical approach, evaluation methodology, reported performance gains, and stated limitations. Cross-category patterns were identified and discussed to highlight convergent trends, recurring architectural choices, and shared evaluation gaps across the RAG retrieval improvement literature.

2.7 Limitations of the methodology

Several methodological limitations should be acknowledged. First, the restriction to six databases, while broad by the standards of the field, may miss relevant work published in venues not indexed by any of these sources, such as certain regional conferences or non-English-language publications. Second, the title-based filtering strategy (IC1), as discussed in Section 2.3, trades recall for precision and may therefore exclude studies that frame their contributions using alternative terminology. Third, citation counts are inherently time-sensitive and favour older publications; given the 2024 - March 2026 temporal window, many recent papers will have accumulated low citation counts regardless of their scientific significance. Fourth, the reliance on API-based automated search means that results are subject to each platform's indexing lag, query interpretation logic, and rate-limiting constraints, which may introduce non-determinism across search executions performed at different times. Finally, the single-reviewer screening process, while partially offset by the deterministic and reproducible nature of the automated filters, does not provide the inter-rater reliability that a multi-reviewer screening process would offer. Future work could address this limitation by introducing a second independent reviewer for a random subset of papers and computing Cohen's kappa to assess agreement.

3 Results and thematic synthesis

This section reports and synthesises the results of the review. It profiles 173 studies (Section 3.1), establishes the organising taxonomy (Section 3.2), and then answers first research question by surveying enhancement techniques stage by stage with technical depth, including the governing formalism and, where the literature is large enough to warrant it, representative quantitative results (Sects. 3.3–3.8). Evaluation practice is treated in Section 3.9, domain deployments in Section 3.10, and recurring limitations in

Table 3 Distribution of the 173 studies across primary contribution categories.

Primary contribution category	n	Pipeline stage
Indexing & chunk optimisation	9	Pre-retrieval
Query reformulation	5	Pre-retrieval
Hybrid & dense retrieval	11	Retrieval
Knowledge-graph retrieval	25	Retrieval
Multimodal retrieval	13	Retrieval
Re-ranking & context refinement	6	Post-retrieval
Retrieval-augmented reasoning	10	Generation
Fine-tuning & parametric integration	10	Generation
Faithfulness & factual grounding	11	Generation
Modular, multi-step & agentic	12	Cross-stage
Efficiency & serving	3	Cross-stage
Evaluation & robustness	23	Cross-stage
Domain deployments	35	Application
Total	**173**	

Section 3.11. Every included study is cited within its category; representative anchors are analysed in depth, and the remaining studies are grouped by technical approach.

3.1 Overview of the corpus

The corpus comprises 173 studies published between January 2024 and March 2026 search date: 77 from 2024, 92 from 2025, and 4 already indexed for 2026. Roughly one third (58/173) are arXiv preprints; the remainder appeared in peer-reviewed venues spanning other data sources shown in Table 1. Citation counts are heavily skewed, with a few anchors, Blended RAG [117] (149), a domain-specific factual-accuracy study [75] (117), SCORE [170] (84), and KRAGEN [88] (77), and a long low-citation tail of recent work; counts are used here only for descriptive profiling and anchor selection. Table 3 gives the distribution across primary contribution categories; the largest are knowledge-graph retrieval (25), domain deployments (35), and evaluation (23).

3.2 A taxonomy of RAG retrieval enhancement

We organise the synthesis using the developmental framing of RAG as an evolution from Naive, through Advanced, to Modular RAG [33], superimposed on the canonical pipeline decomposition into pre-retrieval, retrieval, post-retrieval, and generation stages. Naive RAG is the basic retrieve-then-read pipeline; Advanced RAG comprises stage-local optimisations; Modular RAG restructures the pipeline through iteration, routing, memory, or agency. Mapping the corpus onto this scheme (Table 4) shows that the large majority of studies are Advanced RAG while a smaller, growing set pursues the Modular RAG agenda. The stage at which a technique intervenes therefore structures Sections. 3.3–3.8.

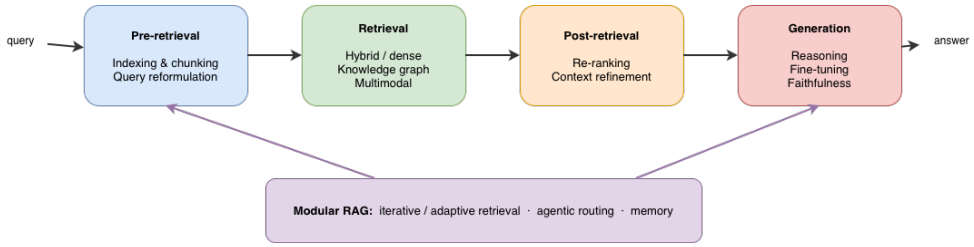


Fig. 1 The RAG pipeline. Advanced RAG optimises individual stages (pre-retrieval, retrieval, post-retrieval, generation); Modular RAG adds cross-stage feedback through iterative retrieval, agentic routing, and memory.

Table 4 Mapping of categories onto the Naive–Advanced–Modular framing.

Paradigm	Categories represented in the corpus
Advanced RAG (stage optimisation)	Chunk optimisation; query reformulation; hybrid/dense, knowledge-graph and multimodal retrieval; re-ranking and refinement; reasoning; fine-tuning; faithfulness
Modular RAG (pipeline restructuring)	Multi-step/iterative retrieval; agentic and memory-augmented pipelines; adaptive routing; efficient distributed serving

3.3 Pre-retrieval enhancement

Pre-retrieval techniques improve what enters the index and how intent is expressed, before any search runs; because every later stage operates on these units, pre-retrieval choices bound achievable retrieval quality. Two sub-themes emerge: indexing and chunk optimisation (Section 3.3.1) and query reformulation (Section 3.3.2).

3.3.1 Indexing and chunk optimisation

A corpus is partitioned into *chunks* that are the atomic units of retrieval; a query can retrieve a chunk but never a sub-span. Chunking thus fixes retrieval granularity. The baseline is fixed-size chunking: a document of N tokens is split into windows of length s with overlap o, giving

$$n_c = \left\lceil \frac{N - o}{s - o} \right\rceil \tag{1}$$

chunks. While this approach remains corpus-agnostic, it is inherently structure-blind, inducing two competing failure modes: fragmentation, where a fixed boundary splits a coherent fact across multiple chunks, and embedding dilution, where the mean-pooling of token embeddings by dense retrievers obscures fine-grained semantic signals.

$$e(c) = \frac{1}{|c|} \sum_{t \in c} \varphi(t), \tag{2}$$

where c denotes the text chunk or sequence, $|c|$ represents the total number of tokens contained within c, $\varphi(t)$ is the d-dimensional embedding vector of an individual token t, and $e(c)$ is the resulting aggregated chunk embedding. So, as $|c|$ grows $e(c)$ regresses toward the corpus mean μ and discriminability falls ($\|e(c) - \mu\| \to 0$). Fragmentation favours larger chunks, dilution smaller ones, yielding the inverted-U of quality against chunk size (Fig. 2).

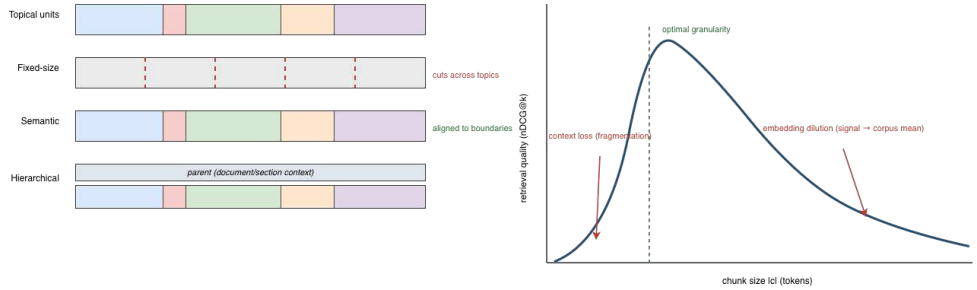


Fig. 2 Chunking strategies (a) and the fragmentation–dilution trade-off (b). Small chunks lose context; large chunks dilute the embedding toward the corpus mean; an intermediate granularity is optimal.

Tanyildiz et al. pair dynamic chunk sizing with `all-mpnet-base-v2` embeddings and cross-encoder re-ranking, improving relevance over static baselines. The canonical adaptive method is semantic chunking: given consecutive sentence embeddings, a boundary is placed where consecutive cosine distance exceeds a percentile threshold τ_p,

$$d_j = 1 - \cos(e_j, e_{j+1}); \qquad \text{boundary after } j \iff d_j \geq \tau_p, \tag{3}$$

where j is an index representing the current position in a sequence, e_j and e_{j+1} are the embedding vectors of the items at position j and $j+1$, $\cos(e_j, e_{j+1})$ is the cosine similarity between the two vectors, d_j is the cosine distance between item j and item $j+1$, and τ_p is a predefined threshold parameter for triggering a boundary. So, topic transitions become chunk edges. Danter et al. [26] evaluate this family in e-learning and find chunking and search strategy interact rather than contribute independently, an early signal of the stage-coupling discussed in Section 3.11. The limitation is cost and brittleness: every sentence must be embedded before indexing, and τ_p is a corpus-dependent hyper-parameter that studies rarely report.

A second group preserves the structure that naive chunking destroys. Lin introduces enhanced PDF structure recognition so ingestion retains layout, tables, and section hierarchy in professional documents. Che et al. generalise to scientific papers via hierarchical context augmentation, annotating each fine unit with section- and document-level context so locally retrieved passages remain globally interpretable. Hierarchical indexing decouples retrieval granularity from context granularity, but presupposes a reliable structure parser.

A third group repairs chunks after retrieval: Oreo is a pluggable context reconstructor that refines retrieved chunks before generation, shifting part of the chunking burden to query time (and overlapping with Section 3.5). A fourth group enriches the index itself: Bashir et al. [12] show that the choice of indexing technique materially affects closed-domain QA accuracy, while Mishra et al. [93] and Singh et al. [128] use LLM-generated metadata to expand each unit's retrievable surface forms, effectively query-independent index expansion that raises recall at the cost of an offline generation pass and possible metadata noise.

Finally, Kashmira et al. [54] argue the paradigm itself is limited in the sense that chunk-level text-to-text similarity cannot capture relationships spanning multiple chunks and is highly sensitive to chunking strategy. The sensitivity is formalised in equations 1 - 3 and they replace chunk retrieval with knowledge-graph retrieval, connecting directly to Section 3.4.2. Across the nine studies, researchers try a bit of everything with text chunks from tweaking their sizes, to trying to fix the context lost by chunking, to eventually giving up on the method completely. The biggest unresolved issue is that there is no one-size-fits-all solution. The perfect setup depends entirely on the data, the questions asked, and the AI model, yet researchers almost never explain how changing these variables affects their findings.

3.3.2 Query reformulation

Five studies improve the query side, addressing the mismatch between how users ask and how answers are written. QuIM-RAG [114, 115] inverts the matching direction: it pre-generates candidate questions from corpus chunks and matches the user query against these inverted questions, implemented over Meta-LLaMA-3-8B on a 500-page corpus to combat information dilution and hallucination. Koo et al. [62] delegate query construction to the model, optimising LLM-generated queries to maximise retrieved-document relevance and thereby reduce hallucination. Iterative variants extend this: IterKey [40] regenerates keywords across rounds until retrieval stabilises, trading retrieval calls for interpretability, and a meta-question framework [73] augments domain queries with generated meta-questions. The category is small, and query-side optimisation remains under-explored relative to corpus- and retriever-side work (Section 3.11).

3.4 Retrieval-stage enhancement

Retrieval-stage techniques target the matching mechanism and form the largest and most active region of the corpus, dividing into hybrid/dense retrieval, knowledge-graph retrieval, and multimodal retrieval.

3.4.1 Hybrid and dense retrieval architectures

Dense retrieval scores a chunk by embedding similarity, $s_{\text{dense}}(q, c) = \cos(\psi(q), e(c))$, capturing semantics but missing exact terms; sparse retrieval (BM25) captures lexical overlap but not paraphrase. Hybrid retrieval fuses both, commonly by reciprocal rank

fusion over each retriever's rank $r_i(\cdot)$,

$$s_{\text{hybrid}}(c) = \sum_{i \in \{\text{dense,sparse}\}} \frac{1}{k + r_i(c)} \tag{4}$$

with a small constant k. Blended RAG [117], one of the most cited studies in this review, instantiates this with dense vector and sparse encoder indexes under hybrid query strategies, reaching an F1 roughly 50% above prior benchmarks on SQuAD without fine-tuning. The same principle recurs: a hybrid query-response system [97], COS-Mix's fusion of cosine similarity and distance for sparse data [53], a complex-reasoning hybrid system [173], Elasticsearch-based hybrid retrieval [23], and an academic-advising chatbot [87] all report gains from combining lexical and semantic signals. Other studies refine the dense side specifically: a learning-to-rank approach for electronic medical records [168], topic-embedding-enhanced retrieval [45], position-embedding-agnostic attention re-weighting (PEAR) to counter context-length limits [134], biomedical document embeddings [61], and semantic compression with graph-augmented retrieval beyond nearest neighbours [109]. The evidence that fusion reliably helps is unusually consistent; its limits are equally clear, since Blended RAG itself reports no gain on the metadata-poor CoQA dataset and cautions that nDCG/F1 are weak proxies for human-judged quality that qualifications that foreshadow Section 3.9.

3.4.2 Knowledge-graph-augmented retrieval

With 25 studies, knowledge-graph (KG) retrieval is the largest category, motivated by a structural intuition: flat chunk retrieval returns independent units, whereas a graph supports retrieval of a connected subgraph or reasoning path, enabling multi-hop questions that no single chunk answers (Fig. 3). Formally, given a query entity set E_q, KG methods retrieve a path or subgraph maximising a relevance score subject to a hop budget L,

$$G^\star = \arg \max_{G' \subseteq G,\ \text{hops}(G') \le L} \text{rel}(q, G'). \tag{5}$$

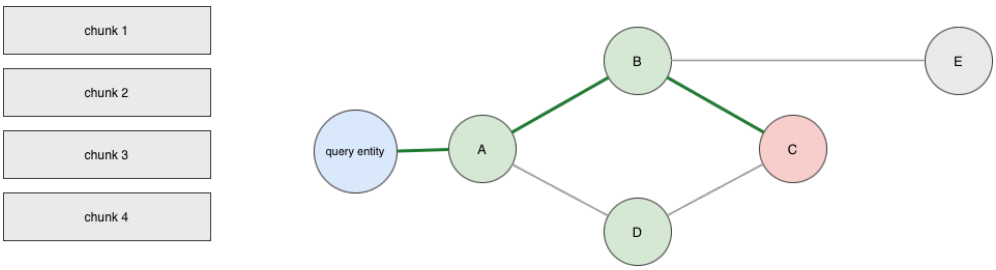


Fig. 3 Flat chunk retrieval returns disconnected units, whereas knowledge-graph retrieval traverses entities and relations to return a connected reasoning path (green), supporting multi-hop queries.

where $G^\star$ represents the optimal subgraph retrieved from the knowledge graph, which serves as the output of the retrieval process. G' is a candidate subgraph under evaluation during the optimization process. G denotes the complete knowledge graph containing all entities and their relations. The constraint $G' \subseteq G$ ensures that the candidate subgraph is a subset of the full graph. hops(G') measures the hop distance, referring to the number of edges or steps in the subgraph. L specifies the maximum hop limit, establishing an upper bound on the subgraph size and controlling the scope of retrieval. rel(q, G') is the relevance function that computes the similarity or relevance score between the query q and the candidate subgraph G'. q represents the user query or question that drives the retrieval process.

The biomedical anchor KRAGEN [88] converts an Alzheimer's KG into a vector store and uses graph-of-thoughts prompting to decompose problems; multiple-choice accuracy rises from 56.6% to 70.4% (1-hop) and 53.1% to 71.8% (2-hop), and 1-hop true/false from 68.6% to 80.3%. The category spans three technical threads. The first integrates KG structure into retrieval directly: KG-enhanced Failure Mode and Effects Analysis(FMEA) for manufacturing reliability [9], community-level retrieval for clinical prediction [49], hypothesis-driven HyKGE [51], GNN-enhanced retrieval [81], query-aware GNNs [3], neural graph matching for molecules [153], triple-based TriRAG [176], path pooling for efficiency [151], and unified or hierarchical graphs (UKRAG [4], HG-RAG [123]). The second bridges graphs and language models: CoTKR [160] rewrites retrieved subgraphs into chain-of-thought text the generator can use, while document-level graphs (Document GraphRAG [60], document knowledge graphs [59]) and robustness-oriented variants [111] address representation and noise. The third applies KG retrieval to domains demanding factual precision: DO-RAG and KERAG for domain QA [98, 131], multi-hop QA [120], e-commerce support [104], telecommunications standards (KG2QA [86]), sustainable manufacturing [180], healthcare KGs [8], Kiswahili agriculture [141], and the agentic GeAR [124] and bandit-driven [137] variants. Across all three, graphs are valued for interpretability and logical structure rather than raw recall; the recurring, rarely quantified cost is constructing and maintaining the graph, and KRAGEN's flat 2-hop true/false result warns that this investment does not uniformly translate into accuracy.

3.4.3 Multimodal retrieval

Thirteen studies extend retrieval beyond text and focusing on multiple knowledge sources. For vision, enhanced multimodal RAG-LLM architectures improve visual question answering by retrieving visual evidence alongside text [163], and the convergence with the KG thread is visible in mKG-RAG [172] and the multi-hop M^3KG-RAG [102], which ground VQA in multimodal knowledge graphs. Recognising that multimodality adds failure modes, RAG-Check [94] separates retrieval-induced from cross-modal-fusion-induced error, a contribution that also belongs to Section 3.9. Applications span radiology report generation [24], image captioning with entity context [121], image-quality assessment [30], hierarchical visual retrieval [24], video generation (RagMe [106]), vision-and-text recommendation [144], and 3D Gaussian-splatting-enhanced robotic manipulation [161]. For audio, retrieval improves ASR accuracy under varied acoustics (LA-RAG [78], RAG-Boost [152]). The category is

young and its evaluation practices are still stabilising, but it is the clearest frontier for extending retrieval enhancement beyond text.

3.5 Post-retrieval enhancement: re-ranking and context refinement

Post-retrieval techniques raise the signal-to-noise ratio of the retrieved set before it reaches the generator. A retriever is tuned for recall and returns $N \approx 100$ candidates of mixed relevance; refinement reorders, selects, and compresses them into a clean top-k context (Fig. 4). Re-ranking typically replaces the bi-encoder score with a cross-encoder that jointly encodes query and passage,

$$s_{\text{ce}}(q, c) = \text{CrossEncoder}([q; c]), \qquad \text{then keep top-}k \text{ by } s_{\text{ce}}. \tag{6}$$

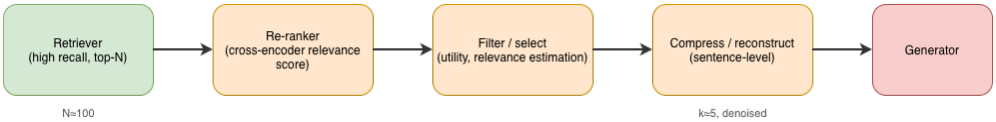


Fig. 4 Post-retrieval refinement converts a high-recall candidate set into a compact, denoised context through re-ranking, selection, and compression each stage trading latency for higher context quality.

RE-RAG [56] adds an explicit relevance estimator that scores passages, improves open-domain QA and interpretability, and degrades gracefully when retrieval is poor. DSLR [46] is unsupervised and it decomposes documents into sentences, filters irrelevant ones, and reconstructs coherent passages, improving over fixed-size passages without training, so it can be retrofitted cheaply. A utility-based passage selector is distilled into a small model for efficiency [175], and context-adaptive synthesis and compression condenses retrieved context while preserving answer-relevant content [182]. Domain pipelines stack rankers, as in the financial Multi-Reranker [68], and HyperRAG [7] reframes the reranker as a systems object, reusing reranker key-value caches to improve the quality–efficiency trade-off. The unifying insight is that recall and context quality are separable problems; the limitation, which HyperRAG targets, is that each ranking stage adds latency proportional to the candidates processed, a cost most studies leave unquantified. That only six studies address this favourable-ratio stage is itself notable (Section 3.11).

3.6 Generation-integrated enhancement

A large group improves outcomes by changing how the generator consumes retrieved evidence, across reasoning, parametric integration, and faithfulness.

3.6.1 Retrieval-augmented reasoning

Ten studies embed retrieval in a reasoning process rather than a one-shot prefix. RARE [145] extends a mutual-reasoning framework to improve reasoning accuracy and factual integrity on knowledge-intensive tasks; a retrieve–rethink–revise chain-of-verification [41] interleaves retrieval with self-verification to detect and repair inadequate evidence; and ReaRAG [69] uses knowledge-guided iterative retrieval to curb the overthinking of large reasoning models. Chain-of-thought integration recurs across settings: CoT-RAG [74], CoT+RAG for clinical phenotype reasoning [158] and corn-cultivation decision agents [21], plus methods that inject external knowledge into the reasoning process [136], emulate retrieval via prompt engineering for long contexts [103], and iterate brainstorming with hypothesis satisfaction [118]. Dynamic parametric RAG [135] injects retrieved knowledge into model parameters at test time to avoid long contexts, bridging to Section 3.6.2. The shared, rarely measured cost is computational: interleaving retrieval with verification or iteration multiplies both retrieval calls and generation passes per query.

3.6.2 Fine-tuning and parametric integration

Ten studies focus the relationship between retrieved (non-parametric) and model (parametric) knowledge. The most cited is an end-to-end case study showing a carefully engineered RAG pipeline substantially improves factual accuracy on domain-specific, time-sensitive queries over private knowledge bases [75]. Others probe the design space: a fine-tuning-enhanced RAG with a quantised influence measure as AI judge [110], comparative studies weighing domain-specific fine-tuning against retrieval [18, 66], weighted integration of retrieved evidence in the output distribution [143], parameter-efficient low-rank adaptation for RAG [25], and a mixture-of-experts–RAG combination [162]. Method-level refinements include ranking-based query rewriting folded into RAFT [165] and legal-domain fine-tuning with retrieval (ASVRI-Legal [96]); a dedicated accuracy evaluation of fine-tuning-plus-RAG against hallucination closes the group [96]. The emerging consensus is that retrieval and fine-tuning are complementary points on a spectrum rather than competitors, and the most robust systems blend them.

3.6.3 Faithfulness and factual grounding

Eleven studies make faithfulness the explicit target. Open-RAG [47] transforms a dense model into a parameter-efficient sparse mixture-of-experts trained to navigate misleading distractors, with hybrid adaptive retrieval deciding when to retrieve; on multi-hop HotpotQA it reaches EM 63.3 (vs RAG 2.0's 54 and Command R+'s 60) and improves accuracy by 9.4% and F1 by 16.14% over Adaptive-RAG, surpassing far larger proprietary systems. SCORE [170], the most cited here, maintains coherence and factual consistency across AI-generated narratives. Empirical work tempers the "more context is better" assumption: long-context RAG performance does not increase monotonically with context length [71]. Other contributions target faithfulness directly. ERAGent's multi-component accuracy/efficiency/personalisation pipeline [125], VERA's validation-and-enhancement layer [16], factuality-optimised

FoRAG for long-form web QA [20], multi-stage REVERSUM for Wikipedia tail entities [1], grounded math QA [42], and applied accuracy studies [5, 142] including a LlamaIndex practitioner framework [36]. The common cost is that the strongest methods (e.g. Open-RAG) require architectural change and training, and adaptive retrieval trades inference speed for accuracy; faithfulness, like structure, is rarely free.

3.7 Modular, multi-step, and agentic pipelines

Twelve studies restructure the pipeline itself, embedding retrieval in a control loop that can re-plan, re-retrieve, and verify before answering (Fig. 5). ERAGent [125] (also a faithfulness contribution) exemplifies the extended multi-module pipeline; a four-module synergy [126] separates retrieval management from execution to improve quality and efficiency; M-RAG [154] partitions external memory so the model focuses on the relevant region; and Vendi-RAG [112] adaptively trades diversity against quality to reduce redundancy. Li et al. [2025] observes that conventional iterative RAG simply piles retrieved passages into an ever-growing context, which burdens the model with redundant text and lets futile reasoning steps trigger further irrelevant retrievals. Instead of accumulating passages, it treats the LLM as an active tracker that alternates knowledge exploration (deciding which entity-relation pairs to pursue next) with knowledge completion (extracting triplets from the retrieved text), progressively expanding a question-specific knowledge graph until it suffices to answer. A post-hoc backtracing step then identifies which generations actually supported the correct answer, distilling process-supervision data for self-taught fine-tuning. On HotpotQA, 2WikiMultihopQA, and MuSiQue it outperforms iterative baselines such as IRCoT and ReAct and restructuring-based ones such as ERA-CoT, with average absolute EM gains of roughly 5.3% over IRCoT and 4.3% over ERA-CoT using LLaMA3-8B-Instruct. The agentic and memory threads include comparative clinical evidence that custom RAG and agentic augmentation outperform non-custom baselines [157], adaptive memory-based optimisation [108], multi-agent retrieval for open-domain table QA [166], mixture-of-knowledge-paths routing (MoK-RAG [37]), LLM agents for semantic code search [48], tool-enhanced urban intelligence (UrbanMind [164]), distributed agents for traffic-signal control [79], a conversational GPT-4 RAG agent [92], and social-intelligence enhancement for agents [179]. Across the group the linear pipeline gives way to dynamic, self-correcting systems, the Modular RAG trajectory anticipated in Section 3.2.

3.8 Efficiency and serving

Three studies treat retrieval enhancement as a systems problem. RAGO [50] provides a systematic performance-optimisation framework for RAG serving, motivated by the proliferation of RAG variants whose workloads differ and resist a single serving strategy. Efficient distributed retrieval lets small language models on edge devices benefit from external knowledge without prohibitive overhead [84], and a reverse-proxy-style retrieval architecture improves factual accuracy with minimal integration cost [76]. That just three of 173 studies address efficiency against well over a hundred accuracy-oriented ones is one of the review's clearest structural gaps (Section 3.11).

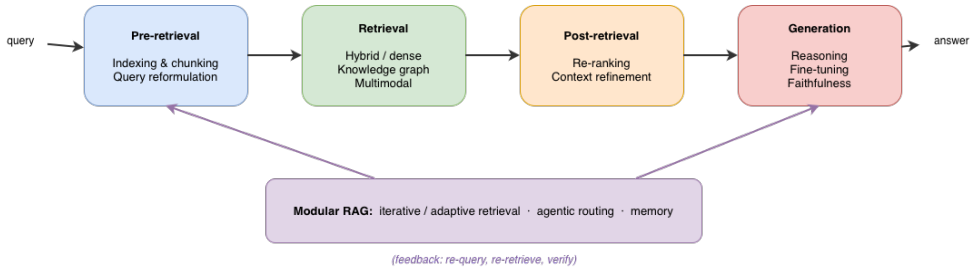


Fig. 5 Modular/agentic RAG places retrieval inside a control loop: a planner routes the query, retrieval draws on graph, hybrid, or memory sources, and a reason/verify step may trigger re-planning before the answer is emitted.

3.9 Evaluation and benchmarking practices

Twenty-three studies make evaluation their primary contribution. Retrieval quality is most often reported with rank-aware measures such as normalised discounted cumulative gain,

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k}, \qquad \mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}, \tag{7}$$

and mean reciprocal rank, while generation quality is reported with exact match, F1, and increasingly a faithfulness or groundedness score measuring the fraction of answer claims supported by retrieved evidence. The most cited evaluation, the RISE study in diabetes education [150], evaluated 43 questions across GPT-4, Claude 2, and Google Bard with clinician and patient raters, raising accurate responses by 12% on average; crucially the gain was uneven (GPT-4 +7%, Claude 2 +19%, Bard +9%), showing the benefit of retrieval is conditional on the base model's prior knowledge. Other rigorous domain evaluations include emergency triage [167], a precision-and-trust GPT-4 study [29], education [129], enterprise open-source LLMs [10], and broad accuracy assessments [15, 89]. Methodological contributions include RAG-Check for multimodal evaluation [94], span-level UncertaintyRAG for long-context modelling [82], and a cluster of robustness, uncertainty, and benchmark studies [32, 111] examining noisy or adversarial retrieval. Robustness studies have shown that retrieval-augmented generation is sensitive not only to retrieved evidence but also to prompt formulation. Hu et al. [2024] demonstrate that even small prompt perturbations can substantially alter model outputs, highlighting prompt sensitivity as an independent source of unreliability. On the measurement side, Oosterhuis et al. [2024] derive reliable confidence intervals for IR evaluation using generative judges, addressing the statistical validity of the LLM-as-judge protocols that faithfulness scoring increasingly depends on. Two survey papers are included as comparative context rather than primary contributions, following the criteria outlined in Section 2.3: a comprehensive survey of RAG architectures and robustness [119] and a survey of precision and interpretability in

legal-domain RAG systems [43]. Two patterns stand out that evaluation is heterogeneous, mixing retrieval precision/recall, end-to-end accuracy, faithfulness, and human judgment but rarely all together and seldom on shared benchmarks; and the most careful evaluations cluster in high-stakes medicine, where error cost compels rigour.

3.10 Domain applications

Thirty-five studies deploy RAG in a specific domain rather than proposing a new mechanism; their value is breadth of evidence that retrieval enhancement transfers to practice. Healthcare dominates: clinical-trial screening with GPT-4 (RECTIFIER [146]), precision oncology [67], surgical decision support (SurgeryLLM [127]), radiology contrast-media consultation [149], health information retrieval prioritising topical relevance and factual accuracy [147], evidence-based medicine implementation [99], emergency ICD-10 coding [58], and personalised physician recommendation [181], alongside condition-specific deployments, a medical-chatbot reliability strategy [35], radiology diagnostics (RadioRAG [140]), and a healthcare-wide scoping review [19], that together span the medical subset. Agriculture is a recurring non-medical theme coffee-leaf-disease remediation [64], farmer-facing AgroLLM [116], and broiler-production environmental control [70] alongside education chatbots and real-time tutoring [34, 63, 132, 139, 174], e-commerce and recommendation [13, 52, 80, 113, 155, 171], where CoRAL [2024] shows that collaborative retrieval over user–item signals improves long-tail recommendation, a setting in which parametric knowledge is weakest, legal technology [101, 122, 159], finance [31, 133], software and code [38, 177], tourism (Tibetan tourism LLMs [107]), and industrial or engineering settings [6, 90]. Table 5 summarises the cross-cutting domain distribution. It turns out healthcare is the strongest domain signal in the RAG. It has driven the field's most rigorous evaluation, but also skews the evidence base toward the requirements of clinical settings, raising the generalisation question to be discussed in the Discussion section.

Table 5 Cross-cutting distribution of studies by application domain

Application domain	n	**Representative anchor**
Healthcare & biomedicine	40	KRAGEN [88]; SurgeryLLM [127]
Industrial & engineering	11	KG-FMEA [9]
Education & academia	8	Academic chatbots [87]
E-commerce & recommendation	7	Graph-enhanced support [104]
Legal	4	Legal RAG survey [43]
Agriculture	4	AgroLLM [116]
Software & code	3	LLM agents for code search [48]
Finance	2	Multi-Reranker [68]
General / cross-domain	94	Blended RAG [117]

3.11 Cross-cutting observations and open challenges

Four patterns recur across categories. First, effort is unevenly distributed: retrieval-stage work, especially knowledge-graph and hybrid retrieval, dominates, while query reformulation (5), post-retrieval refinement (6), and efficiency (3) remain thin despite favourable cost-to-benefit ratios, and the stages are optimised in isolation though they demonstrably interact (Sections 3.3.1,3.4.1). Second, evaluation is fragmented: heterogeneous metrics and the scarcity of shared, stage-decoupled benchmarks (Section 3.9) make cumulative comparison difficult, and the RISE result shows headline gains can mask large base-model-dependent variation. Third, the most effective techniques carry rarely quantified costs-graph construction (Section 3.4.2), multiplied retrieval and generation calls in reasoning and modular loops (Sects. 3.6,3.7), and re-ranking latency (Section 3.5) which the three-study efficiency literature only begins to address. Fourth, the evidence base is healthcare-skewed, advancing rigour there while leaving open how techniques generalise to lower-stakes or lower-resource domains. These observations frame in the Section 4.

4 Discussion

This section interprets the results reported in Section 3. We first state the principal findings against each research question (Section 4.1). We then develop the four cross-cutting challenges identified in Section 3.11 into substantive discussions, each paired with concrete research directions: the uneven optimisation of the pipeline (Section 4.2), the fragmentation of evaluation practice (Section 4.3), the unmeasured cost of accuracy-oriented techniques (Section 4.4), and the domain skew of the evidence base (Section 4.5). Section 4.6 mentions field's convergence toward modular and agentic architectures, Section 4.7 distils implications for practitioners, and Section 4.8 consolidates a research agenda. Threats to the validity of the review are considered in Section 4.9.

4.1 Principal findings

Most RAG research tweaks just one step in the standard retrieve-then-answer pipeline, with only a small minority exploring genuinely different approaches using memory, routing, or decision-making though these sophisticated alternatives are growing. The field's biggest weakness is measurement fragmentation, i.e, studies test success wildly differently, some checking retrieval accuracy, others final answer quality, others trustworthiness. No shared benchmarks exist, making it impossible to fairly compare results across studies. This leaves the field facing four recurring problems: different pipeline stages get unequal attention, there's no standard testing methodology, computational costs are rarely reported, and the evidence base is heavily skewed toward healthcare examples rather than diverse applications.

4.2 An unevenly optimised pipeline

The most conspicuous structural feature of the literature is that attention is unequal across pipeline stages. Hybrid, knowledge-graph, and multimodal retrieval together

account for 49 studies, whereas the stages immediately adjacent to retrieval are sparsely populated. Query reformulation comprises five studies and post-retrieval refinement only six. Retriever-stage innovation produces visible, headline accuracy gains and aligns with the field's dominant framing, of which Blended RAG [117] is the exemplar. Yet the evidence suggests the neglected stages offer favourable cost-to-benefit ratios. Query-side interventions such as inverted-question matching [114] and LLM-optimised query generation [62] report substantial precision improvements from changes that require no modification to the retriever, and sentence-level refinement [46] and explicit relevance estimation [56] recover accuracy from existing retrieval results at modest cost. The field has, in effect, invested heavily in retrieving better while under-investing in asking better and filtering better.

A deeper consequence of stage-wise optimisation is that the stages are studied in isolation even though they interact. Chunking decisions [10, 11] determine the units a reranker can later promote; query reformulation changes the distribution of candidates a relevance estimator must score; long retrieved contexts interact with the generator's capacity to use them [71]. Almost no study in the corpus jointly optimises adjacent stages or characterises their interaction. A clear research direction therefore follows move from isolated stage optimisation toward interaction-aware, joint optimisation of the pipeline, treating chunking, query formulation, retrieval, and refinement as coupled rather than independent design choices.

4.3 The fragmentation of evaluation

Our analysis of research question 2 (Section 3.9) shows that evaluation practice, while maturing, is fragmented to a degree that impedes scientific accumulation. Because studies select different metrics, datasets, and baselines, an improvement reported by one study cannot be compared directly with that of another, and the field lacks a reliable means of establishing which techniques represent genuine progress. The two embedded surveys reach the same conclusion from a broader vantage point, noting unresolved robustness and interpretability gaps as open frontiers [43, 119]. The problem is compounded by a conflation of retrieval quality with end-to-end answer quality. A system may score well on answer accuracy for reasons unrelated to retrieval, masking retrieval failures that a stage-decoupled evaluation would expose. RAG-Check [94] is notable precisely because it separates retrieval-induced from fusion-induced error, and points toward the kind of diagnostic evaluation the field needs.

Two main research directions are suggested. First, the field would benefit from standardized benchmarks that evaluate different stages separately such as retrieval accuracy, factual correctness (faithfulness or groundedness), and final task performance, so it becomes clear which part of the system is improving. Second, stronger reporting standards should be adopted, including clearly defined baselines, fully documented datasets, and explicit metric definitions, to improve consistency and comparability across studies. The most rigorous evaluations in the corpus cluster in high-stakes medical settings [150], where the cost of error compels discipline; generalising that discipline to lower-stakes settings is a precondition for trustworthy cumulative progress.

4.4 The unmeasured cost of accuracy

Knowledge-graph augmentation, the largest category, depends on building and maintaining domain graphs whose cost is acknowledged but seldom measured [9, 60, 88]. Other techniques carry similar hidden costs. Iterative and reasoning-based pipelines multiply retrieval and inference calls [41, 69], and layered re-ranking adds latency in proportion to the number of stages [46, 68]. Expanding the retrieved context to improve grounding is doubly expensive: it raises inference cost and, beyond a point, even degrades performance [71]. Test-time parametric injection tries to sidestep this by internalising retrieved knowledge rather than lengthening the prompt [135]. Against this backdrop, only three studies treat efficiency as a primary concern [50, 76], and just one shows that quality and efficiency can be co-designed rather than traded off, through reranker cache reuse [7].

Efficiency is the field's blind spot. Barely three studies make it their primary concern, against well over 130 aimed at accuracy and that gap bites hardest exactly as RAG leaves the demonstration stage for deployment, where a technique is judged not on whether it raises accuracy, but on how much accuracy it buys per unit of latency, memory, and maintenance effort. Closing the gap means changing what we measure: reporting quality, i.e, efficiency Pareto frontiers instead of accuracy alone, treating latency, cost, and upkeep as first-class metrics, and co-designing retrieval logic with the serving infrastructure it runs on, an agenda that RAGO [50] and HyperRAG [7] only begin to illustrate.

4.5 Domain skew and the generalisation question

The evidence base is markedly skewed toward healthcare: 40 studies carry a medical tag, and the most rigorous applied evaluations, clinical trial screening [146], precision oncology [63], surgical decision support [127], and radiology consultation [149], are concentrated. This concentration has a double edge. On one hand, it has driven the field's most careful work, because clinical stakes demand explicit faithfulness evaluation, guideline grounding, and privacy-preserving local deployment. On the other hand, it raises an unresolved generalisation question. Techniques tuned to the retrieval of structured, authoritative clinical guidelines may not transfer to domains with noisier, less curated, or lower-resource knowledge sources, which remain thinly represented, agriculture, finance, and software engineering each account for a handful of studies. The risk is that field-level conclusions implicitly inherit the assumptions of the medical setting in which they were validated.

The research direction follows directly: run explicit cross-domain transfer studies. Take a single retrieval-enhancement technique and test it across domains that differ in knowledge-source quality, query complexity, and tolerance for error, then report where it holds up and where it breaks. Until such studies exist, any claim that a technique improves RAG in general should be treated as provisional.

4.6 Convergence toward modular and agentic RAG

This review tells a coherent story that the four challenges above all reflect the stage-wise optimisations of Advanced RAG are beginning to be composed into the dynamic,

self-correcting systems of Modular RAG. Extended multi-component pipelines [125], explicit module-separation designs [126], partitioned memory [154], and adaptive diversity–quality trade-offs [112] each restructure the pipeline rather than tuning a single stage, while retrieval-augmented reasoning [145, 41, 69] dissolves the boundary between retrieval and generation by embedding retrieval inside a reasoning loop. Comparative clinical evidence that custom and agentic-augmented systems outperform non-custom baselines [157] indicates that this restructuring delivers measurable benefit, not merely architectural novelty.

The developmental framework proposed by Gao et al. [33] and the evidence synthesized in this review suggest that orchestration represents an important direction for future RAG research. Such systems would dynamically select retrieval and reasoning strategies according to query characteristics, routing multi-hop relational queries to graph-based retrieval, fact-centric queries to hybrid retrieval approaches, and invoking refinement or additional retrieval steps when confidence is insufficient. They would also incorporate mechanisms for managing conversational memory across interactions. Although early examples of such architectures appear within the reviewed literature, the evidence remains limited, with only 12 studies adopting modular designs and little consistency in how components are integrated or evaluated. Consequently, the development of robust orchestration frameworks, together with evaluation methodologies that separately assess system stages while accounting for efficiency considerations, remains an important challenge for advancing RAG systems.

4.7 Implications for practice

The evidence synthesized in this review suggests several considerations for the development of retrieval-augmented systems. Across the reviewed studies, hybrid retrieval methods that combine sparse and dense representations are consistently associated with improved retrieval and downstream task performance, making them a commonly adopted approach [117]. The findings also indicate that retrieval effectiveness depends substantially on the quality of document indexing and preprocessing, highlighting the importance of structure-preserving ingestion and appropriate chunking strategies [10, 11]. In many cases, post-retrieval refinement and relevance estimation techniques have been shown to improve performance without requiring major architectural modifications [46, 56]. The literature further suggests that retrieval augmentation and domain-specific fine-tuning are not mutually exclusive approaches but can be combined to support domain adaptation effectively [18]. Finally, studies conducted in high-stakes domains, particularly healthcare, emphasize the importance of evaluating faithfulness explicitly rather than relying solely on answer accuracy, while maintaining appropriate levels of human oversight [146, 127].

4.8 A research agenda

Building on the findings synthesized throughout this review, several priorities for future research emerge:

1. **Interaction-aware optimisation.** Existing studies largely evaluate chunking, query formulation, retrieval, and refinement independently. Future work should

investigate these stages as interdependent components whose effectiveness may depend on interactions across the retrieval pipeline (Section 4.2).

2. **Stage-decoupled benchmarks and reporting standards.** The development of shared evaluation frameworks that separately measure retrieval performance, faithfulness, and end-task accuracy would improve comparability and enable more cumulative assessment of methodological advances (Section 4.3).
3. **Quality–efficiency co-evaluation.** Given the limited attention paid to computational cost in the current literature, future studies should report latency, memory requirements, and maintenance costs alongside effectiveness metrics to facilitate more realistic assessments of retrieval-enhancement techniques (Section 4.4).
4. **Cross-domain validation.** Additional evidence is needed regarding the transferability of retrieval-enhancement techniques across domains with differing knowledge characteristics, data quality, and tolerance for error, particularly beyond the heavily represented healthcare setting (Section 4.5).
5. **Modular orchestration frameworks.** Although modular and adaptive retrieval architectures have begun to emerge, there remains limited evidence regarding effective approaches to component integration, orchestration, and evaluation. Further research is needed to establish common design and assessment practices (Section 4.6).
6. **Under-investigated retrieval stages.** Query reformulation and post-retrieval refinement remain comparatively underrepresented in the literature despite evidence suggesting their potential contribution to system performance. More systematic investigation of these stages may therefore be warranted (Section 4.2).

4.9 Limitations of This Review

Several limitations should be considered when interpreting the findings of this review. First, the inclusion criteria prioritised studies that explicitly framed their contributions as retrieval enhancements. Consequently, the identified trends reflect the self-described retrieval-enhancement literature and may not capture all relevant advances reported under alternative terminology. Second, although the qualitative synthesis drew on representative studies selected using transparent criteria, the emphasis on influential publications may have reduced the visibility of some recent contributions. Third, while thematic categories were developed through explicit coding rules and manual reconciliation, classification of certain studies involved subjective judgement, particularly where boundaries between retrieval, reasoning, and modular orchestration approaches were not clearly defined. Finally, the review synthesised reported findings qualitatively rather than conducting a quantitative meta-analysis. As a result, the conclusions identify broad patterns, areas of convergence, and research gaps, but do not provide comparative estimates of effect size across techniques.

Despite these limitations, the review provides a structured synthesis of recent research on RAG retrieval enhancement and highlights both established directions and areas requiring further investigation.

5 Conclusion

This review has examined a rapidly expanding and fragmented body of work on techniques proposed between 2024 and 2026 for improving RAG systems. Using a PRISMA-based search strategy, we identified 173 studies, categorised them according to their underlying technical contributions and application domains, and organised them along the RAG pipeline stages they target: pre-retrieval, retrieval, post-retrieval, generation, and end-to-end pipeline designs. This organisation provides a unified perspective on a literature that is otherwise highly heterogeneous.

The resulting synthesis reveals a pronounced concentration of effort on retrieval-stage methods, particularly those involving knowledge graph–based approaches, while comparatively fewer studies investigate pre-retrieval query transformation or post-retrieval refinement. Beyond this imbalance, three recurring limitations are evident across the literature. First, comparability remains limited due to inconsistent use of evaluation metrics, datasets, and baselines. Second, computational cost is rarely reported, despite the non-trivial overhead introduced by graph construction, iterative retrieval, and long-context generation. Third, the evidence base is skewed toward healthcare applications, which, while methodologically rigorous, raises questions regarding generalisability to other domains.

Despite these limitations, the literature shows a clear progression toward more integrated system designs. Recent approaches increasingly combine retrieval, reasoning, self-verification, and iterative querying, indicating a shift from static pipelines toward more adaptive and modular architectures. Based on this synthesis, future work would benefit from jointly optimising pipeline stages, establishing shared benchmarks that decouple retrieval and generation performance, explicitly reporting efficiency alongside accuracy, evaluating methods across diverse domains, and further developing orchestration mechanisms that coordinate multiple retrieval and reasoning components.

Overall, the primary contribution of this review is a structured mapping of where research effort has been concentrated, where systematic gaps remain, and how these gaps may inform future work in retrieval-augmented generation systems.

Declarations.

- **Funding:** No funding was received to assist with the preparation of this manuscript.
- **Competing interests:** The author declares no competing interests relevant to the content of this article.
- **Data availability:** The search-and-screening pipeline, inclusion/exclusion criteria, and extracted study metadata supporting this review are implemented in the open-source `litreview` library, available at https://github.com/abdullahkarasan/litreview.
- **Code availability:** The `litreview` library used to query bibliographic sources, deduplicate records, and export structured metadata is openly available at https://github.com/abdullahkarasan/litreview.

- **Author contributions:** A. Karasan conceived the review, designed and executed the search strategy, performed study screening, data extraction, and synthesis, and wrote and revised the manuscript.

References

[1] Adak S et al. (2025) REVERSUM: A Multi-staged Retrieval-Augmented Generation Method to Enhance Wikipedia Tail Biographies through Personal Narratives. International Conference on Computational Linguistics. https://doi.org/10.48550/arXiv.2502.12137

[2] Agnihotram G, Joydeep Sarkar (2025) Evaluating Precision and Recall at Retrieval Time in Retrieval-Augmented Generation (RAG) Systems. American Journal of Computer Science and Technology. https://doi.org/10.11648/j.ajcst.20250804.11

[3] Agrawal V, Fay Wang, R. Puri (2025) Query-Aware Graph Neural Networks for Enhanced Retrieval-Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2508.05647

[4] Alkouz A (2025) UKRAG: A Unified Knowledge Graph to Enhance Retrieval Augmented Generation Performance. Communications in Computer and Information Science. https://doi.org/10.1007/978-3-031-82931-4_1

[5] Alnefaie S (2024) Using the Retrieval-Augmented Generation Technique to Improve the Performance of GPT-4 in Answering Quran Questions. 2024 6th International Conference on Natural Language Processing Icnlp 2024. https://doi.org/10.1109/ICNLP60986.2024.10692797

[6] Amendola M et al. (2025) Spatially-Enhanced Retrieval-Augmented Generation for Walkability and Urban Discovery. arXiv [cs.IR].

[7] An Y et al. (2025) HyperRAG: Enhancing Quality-Efficiency Tradeoffs in Retrieval-Augmented Generation with Reranker KV-Cache Reuse. arXiv.org. https://doi.org/10.48550/arXiv.2504.02921

[8] Anuyah S et al. (2025) Domain-Specific Knowledge Graphs in RAG-Enhanced Healthcare LLMS. BigData Congress [Services Society]. https://doi.org/10.1109/BigData66926.2025.11400857

[9] Bahr L et al. (2024) Knowledge Graph Enhanced Retrieval-Augmented Generation for Failure Mode and Effects Analysis. Journal of Industrial Information Integration. https://doi.org/10.1016/j.jii.2025.100807

[10] Balakrishnan G, Anupam Purwar (2024) Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability. IEEE India Conference. https://doi.org/10.1109/INDICON63790.2024.10958508

[11] Baqar M, Rajat Khanda (2025) Hallucinations and Truth: A Comprehensive Accuracy Evaluation of RAG, LoRA and DoRA. arXiv.org. https://doi.org/10.48550/arXiv.2502.10497

[12] Bashir B et al. (2025) Analyzing the Impact of Various Indexing Techniques on Retrieval-Augmented Generation (RAG) Performance in Closed-Domain Question Answering. International Journal of Science for Global Sustainability. https://doi.org/10.57233/ijsgs.v11i2.872

[13] Benita J et al. (2024) Implementation of Retrieval-Augmented Generation (RAG) in Chatbot Systems for Enhanced Real-Time Customer Support in E-Commerce. 2024

3rd International Conference on Automation, Computing and Renewable Systems (ICACRS). https://doi.org/10.1109/ICACRS62842.2024.10841586

[14] Bernardi M, Marta Cimitile (2024) Report Generation from X-Ray imaging by Retrieval-Augmented Generation and improved Image-Text Matching. IEEE International Joint Conference on Neural Network. https://doi.org/10.1109/IJCNN60899.2024.10650332

[15] Bingöl FG, D. Ağagündüz, Mustafa Can Bingöl (2025) Accuracy of Current Large Language Models and The Retrieval Augmented Generation Model in Determining Dietary Principles in Chronic Kidney Disease.. Journal of renal nutrition. https://doi.org/10.1053/j.jrn.2025.01.004

[16] Birur NA et al. (2024) VERA: Validation and Enhancement for Retrieval Augmented systems. arXiv.org. https://doi.org/10.48550/arXiv.2409.15364

[17] Brown A, Roman M, Devereux B (2025) A systematic literature review of retrieval-augmented generation: techniques, metrics, and challenges. Big Data Cogn Comput 9(12):320

[18] Budakoglu G, Hakan Emekçi (2025) Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3542334

[19] Bunnell MPDJ et al. (2025) Bridging AI and Healthcare: A Scoping Review of Retrieval-Augmented Generation - Ethics, Bias, Transparency, Improvements, and Applications. medRxiv. https://doi.org/10.1101/2025.04.01.25325033

[20] Cai T, Zhen Tan, Xiaoshuang Song, Tao Sun, Jun Jiang, Yiqun Xu, Yifan Zhang, Jia Gu (2024) FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), pp 199–210. https://doi.org/10.1145/3637528.3672065

[21] Chao A (2025) Corn Cultivation with Precision: Language Agents for Real-Time Decision Making. 2025 1st International Conference on Consumer Technology (ICCT-Pacific). https://doi.org/10.1109/ICCT-Pacific63901.2025.11012907

[22] Che T et al. (2024) A Hierarchical Context Augmentation Method to Improve Retrieval-Augmented LLMs on Scientific Papers. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). https://doi.org/10.1145/3637528.3671847

[23] Chen J et al. (2024) Optimizing Retrieval-Augmented Generation with Elasticsearch for Enhanced Question-Answering Systems. 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). https://doi.org/10.1109/ICBAIE63306.2024.11117089

[24] Chen K et al. (2025) Observe, align, and enhance: a hierarchical retrieval-augmented vision-language model for generating radiology reports. Health Information Science and Systems. https://doi.org/10.1007/s13755-025-00389-9

[25] Choi Y et al. (2025) Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation. Applied Sciences. https://doi.org/10.3390/app15084425

[26] Danter D, Heidrun Mühle, Andreas Stöckl (2024) Advanced Chunking and Search Methods for Improved Retrieval-Augmented Generation (RAG) System Performance in E-Learning. AHFE International. https://doi.org/10.54941/ahfe1005756

[27] Eskenazi J (2025) Evaluating retrieval augmented generation and ChatGPT's accuracy on orthopaedic examination assessment questions. Annals of Joint. https://doi.org/10.21037/aoj-24-49

[28] Feng N et al. (2025) Response Quality Assessment for Retrieval-Augmented Generation via Conditional Conformal Factuality. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3726302.3730244

[29] Fink A et al. (2025) Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: a proof-of-concept study. European Radiology. https://doi.org/10.1007/s00330-025-11445-z

[30] Fu K et al. (2026) Enhancing Image Quality Assessment Ability of LMMs via Retrieval-Augmented Generation. arXiv [cs.CV, cs.AI].

[31] Fuangfung P (2025) Enhanced Retrieval-Augmented Generation for Intelligent Customer Service in Electricity Utility Sector : A Case Study of MEA's DTD Doc Insight System. 2025 25th Conference of the Electric Power Supply Industry (CEPSI). https://doi.org/10.1109/CEPSI66359.2025.11403553

[32] Gangavarapu R (2025) Evaluating Accuracy in Large Language Models: Benchmarking Corrective Rag Vs. Naive Retrieval Augmented Generation Approach. 2025 IEEE Conference on AI and Data Analytics Icad 2025. https://doi.org/10.1109/ICAD65464.2025.11114027

[33] Gao Y, Xiong Y, Huang X, et al (2024) Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997

[34] Gençer Bingöl F (2025) Accuracy of Current Large Language Models and the Retrieval-Augmented Generation Model in Determining Dietary Principles in Chronic Kidney Disease. Journal of Renal Nutrition. https://doi.org/10.1053/j.jrn.2025.01.004

[35] Ghanbari Haez S (2024) A Retrieval-Augmented Generation Strategy to Enhance Medical Chatbot Reliability. Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics. https://doi.org/10.1007/978-3-031-66538-7_22

[36] Gheorghiu A (2024) Building Data-Driven Applications with LlamaIndex: A practical guide to retrieval-augmented generation (RAG) to enhance LLM applications. Packt Publishing.

[37] Guo Z et al. (2025) MoK-RAG: Mixture of Knowledge Paths Enhanced Retrieval-Augmented Generation for Embodied AI Environments. arXiv [cs.LG, cs.AI].

[38] Ha S et al. (2026) A Selective RAG-Enhanced Hybrid ML-LLM Framework for Efficient and Explainable Fatigue Prediction Using Wearable Sensor Data. Bioengineering. https://doi.org/10.3390/bioengineering13010058

[39] Hao J, Lei Chen, Luyao Meng (2025) Advancing Large Language Models with Enhanced Retrieval-Augmented Generation: Evidence from Biological UAV Swarm Control. Drones. https://doi.org/10.3390/drones9050361

[40] Hayashi K et al. (2025) IterKey: Iterative Keyword Generation with LLMs for Enhanced Retrieval Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2505.08450

[41] He B et al. (2024) Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval Augmented Generation. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2410.05801

[42] Henkel O (2024) Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. Proceedings of the International Conference on Educational Data Mining. https://doi.org/10.5281/zenodo.12729824

[43] Hindi M et al. (2025) Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3550145

[44] Holzinger A, Malle B, Saranti A, Pfeifer B (2021) Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf Fusion 71:28–37

[X] Hu Z, Chenyang Wang, Yuan Shu, Hye-Young Paik, Liming Zhu (2024) Prompt Perturbation in Retrieval-Augmented Generation Based Large Language Models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), pp 1119–1130.

[45] Huseynova K, Jafar Isbarov (2024) Enhanced document retrieval with topic embeddings. Advanced Industrial Conference on Telecommunications. https://doi.org/10.1109/AICT61888.2024.10740455

[46] Hwang T et al. (2024) DSLR: Document Refinement with Sentence-Level Re-ranking and Reconstruction to Enhance Retrieval-Augmented Generation. KNOWLED-GENLP. https://doi.org/10.48550/arXiv.2407.03627

[47] Islam SB et al. (2024) Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2410.01782

[48] Jain S et al. (2024) LLM Agents Improve Semantic Code Search. arXiv.org. https://doi.org/10.48550/arXiv.2408.11058

[49] Jiang P et al. (2024) Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2410.04585

[50] Jiang W et al. (2025) RAGO: Systematic Performance Optimization for Retrieval-Augmented Generation Serving. Preprint. https://doi.org/10.1145/3695053.3731093

[51] Jiang X et al. (2025) HyKGE: A Hypothesis Knowledge Graph Enhanced RAG Framework for Accurate and Reliable Medical LLMs Responses. Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.580

[52] Jun H et al. (2025) Implementing a context-augmented large language model to guide precision cancer medicine. medRxiv. https://doi.org/10.1101/2025.05.09.25327312

[53] Juvekar K, Anupam Purwar (2024) COS-Mix: Cosine Similarity and Distance Fusion for Improved Information Retrieval. arXiv.org. https://doi.org/10.48550/arXiv.2406.00638

[54] Kashmira S et al. (2024) TOBUGraph: Knowledge Graph-Based Retrieval for Enhanced LLM Performance Beyond RAG. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2025.emnlp-industry.93

[55] Kazemzadeh H et al. (2025) DrugRAG: Enhancing Pharmacy LLM Performance Through A Novel Retrieval-Augmented Generation Pipeline. arXiv [cs.CL, cs.AI].

[56] Kim K, Jay-Yoon Lee (2024) RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation. Preprint. https://doi.org/10.18653/v1/2024.emnlp-main.1236

[57] Kitchenham BA, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01, Keele University and Durham University

[58] Klang E et al. (2024) Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders. medRxiv. https://doi.org/10.1101/2024.10.15.24315526

[59] Knollmeyer S (2024) Document Knowledge Graph to Enhance Question Answering with Retrieval Augmented Generation. IEEE International Conference on Emerging Technologies and Factory Automation ETFA. https://doi.org/10.1109/ETFA61755.2024.10711054

[60] Knollmeyer S, Oğuz Caymazer, Daniel Grossmann (2025) Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain. Electronics. https://doi.org/10.3390/electronics14112102

[61] Kong Y et al. (2024) Document Embeddings Enhance Biomedical Retrieval-Augmented Generation. IEEE International Conference on Bioinformatics and Biomedicine. https://doi.org/10.1109/BIBM62325.2024.10822781

[62] Koo H, Minseon Kim, Sung Ju Hwang (2024) Optimizing Query Generation for Enhanced Document Retrieval in RAG. arXiv.org. https://doi.org/10.48550/arXiv.2407.12325

[63] Kreimeyer K (2024) Using Retrieval-Augmented Generation to Capture Molecularly-Driven Treatment Relationships for Precision Oncology. Studies in Health Technology and Informatics. https://doi.org/10.3233/SHTI240575

[64] Kumar D et al. (2024) Overcoming LLM Challenges using RAG-Driven Precision in Coffee Leaf Disease Remediation. 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS). https://doi.org/10.1109/ICETCS61022.2024.10543859

[65] Kumar M et al. (2025) Performance Evaluation of Large Language Models in Detecting Buccal Mucosal Lesions Using Smartphone-Based Imaging. Journal of Pioneering Medical Sciences. https://doi.org/10.47310/jpms202514s0216

[66] Lakatos R et al. (2025) Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems. Machine Learning and Knowledge Extraction. https://doi.org/10.3390/make7010015

[67] Lammert J et al. (2024) Expert-Guided Large Language Models for Clinical Decision Support in Precision Oncology. JCO Precision Oncology. https://doi.org/10.1200/PO-24-00478

[68] Lee J, Minji Roh (2024) Multi-Reranker: Maximizing performance of retrieval-augmented generation in the FinanceRAG challenge. arXiv.org. https://doi.org/10.48550/arXiv.2411.16732

[69] Lee Z et al. (2025) ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2503.21729

[70] Leite MV et al. (2025) Enhancing Environmental Control in Broiler Production: Retrieval-Augmented Generation for Improved Decision-Making with Large Language Models. AgriEngineering. https://doi.org/10.3390/agriengineering7010012

[71] Leng Q et al. (2024) Long Context RAG Performance of Large Language Models. arXiv.org. https://doi.org/10.48550/arXiv.2411.03538

[72] Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol 33, pp 9459–9474

[73] Li C et al. (2024) Development of a Meta-Question Enhanced Retrieval-Augmented Generation Model and Its Application in Dermatology. International Conference on Advanced Computer Theory and Engineering. https://doi.org/10.1109/ICACTE62428.2024.10871274

[74] Li F et al. (2025) CoT-RAG: Integrating Chain of Thought and Retrieval-Augmented Generation to Enhance Reasoning in Large Language Models. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2504.13534

[75] Li J, Ye Yuan, Zehua Zhang (2024) Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. arXiv.org. https://doi.org/10.48550/arXiv.2403.10446

[76] Li P, Ya-yun Lai (2024) Augmenting Large Language Models with Reverse Proxy Style Retrieval Augmented Generation for Higher Factual Accuracy. Preprint. https://doi.org/10.31219/osf.io/ma6cq

[X] Li R, Quanyu Dai, Zeyu Zhang, Xu Chen, Zhenhua Dong, Ji-Rong Wen (2025) KnowTrace: Bootstrapping Iterative Retrieval-Augmented Generation with Structured Knowledge Tracing. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25). https://doi.org/10.1145/3711896.3737015

[77] Li S, Naren Ramakrishnan (2025) Oreo: A Plug-in Context Reconstructor to Enhance Retrieval-Augmented Generation. International Conference on the Theory of Information Retrieval. https://doi.org/10.1145/3731120.3744590

[78] Li S et al. (2024) LA-RAG:Enhancing LLM-based ASR Accuracy with Retrieval-Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2409.08597

[79] Li X et al. (2025) Retrieval Augmented Generation-Enhanced Distributed LLM Agents for Generalizable Traffic Signal Control with Emergency Vehicles. arXiv [cs.AI].

[80] Li Y et al. (2025) FinGEAR: Financial Mapping-Guided Enhanced Answer Retrieval. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2509.12042

[81] Li Z et al. (2024) Graph Neural Network Enhanced Retrieval for Question Answering of Large Language Models. North American Chapter of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.337

[82] Li Z et al. (2024) UncertaintyRAG: Span-Level Uncertainty Enhanced Long-Context Modeling for Retrieval-Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2410.02719

[83] Lin D (2024) Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition. arXiv.org. https://doi.org/10.48550/arXiv.2401.12599

[84] Liu S et al. (2025) Efficient Distributed Retrieval-Augmented Generation for Enhancing Language Model Performance. arXiv [cs.LG, cs.AI].

[85] Liu T et al. (2024) Pointwise Mutual Information as a Performance Gauge for Retrieval-Augmented Generation. North American Chapter of the Association for

Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.78

[86] Luo Z et al. (2025) KG2QA: Knowledge Graph-enhanced Retrieval-augmented Generation for Communication Standards Question Answering. arXiv [cs.CL].

[87] Maryamah M et al. (2024) Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access. International Conference on Knowledge and Smart Technology. https://doi.org/10.1109/KST61284.2024.10499652

[88] Matsumoto N et al. (2024) KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. Bioinform.. https://doi.org/10.1093/bioinformatics/btae353

[89] Meng W et al. (2025) Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application. Electronics. https://doi.org/10.3390/electronics14020386

[90] Meng Z, Zixuan Yi, Iadh Ounis (2025) KERAG_R: Knowledge-Enhanced Retrieval-Augmented Generation for Recommendation. arXiv [cs.IR].

[91] Miladi F (2024) Leveraging GPT-4 for Accuracy in Education: A Comparative Study on Retrieval-Augmented Generation in MOOCs. Communications in Computer and Information Science. https://doi.org/10.1007/978-3-031-64315-6_40

[92] Miladi F (2025) Evaluating a GPT-4 and Retrieval-Augmented Generation-Based Conversational Agent to Enhance Learning Experience in a MOOC. International Conference on Computer Supported Education Csedu Proceedings. https://doi.org/10.5220/0013366100003932

[93] Mishra P et al. (2025) A Systematic Framework for Enterprise Knowledge Retrieval: Leveraging LLM-Generated Metadata to Enhance RAG Systems. arXiv.org. https://doi.org/10.48550/arXiv.2512.05411

[94] Mortaheb M et al. (2025) RAG-Check: Evaluating Multimodal Retrieval Augmented Generation Performance. arXiv.org. https://doi.org/10.48550/arXiv.2501.03995

[95] Mourão E, Kalinowski M, Murta L, Mendes E, Wohlin C (2020) On the performance of hybrid search strategies for systematic literature reviews in software engineering. Inf Softw Technol 123:106294

[96] Octadion O et al. (2025) ASVRI-Legal: Fine-Tuning LLMs with Retrieval Augmented Generation for Enhanced Legal Regulation. arXiv [cs.CL].

[97] Omrani P et al. (2024) Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement. 2024 10th International Conference on Web Research (ICWR). https://doi.org/10.1109/ICWR61162.2024.10533345

[X] Oosterhuis H, Rolf Jagerman, Zhen Qin, Xuanhui Wang, Michael Bendersky (2024) Reliable Confidence Intervals for Information Retrieval Evaluation Using Generative AI. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), pp 2307–2317.

[98] Opoku DO, Ming Sheng, Yong Zhang (2025) DO-RAG: A Domain-Specific QA Framework Using Knowledge Graph-Enhanced Retrieval-Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2505.17058

[99] Özmen BB, Piyush Mathur (2025) Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery. Journal of Plastic Reconstructive & Aesthetic Surgery. https://doi.org/10.1016/j.bjps.2025.03.053

[100] Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

[101] Panchal D et al. (2025) LawPal : A Retrieval Augmented Generation Based System for Enhanced Legal Accessibility in India. arXiv.org. https://doi.org/10.48550/arXiv.2502.16573

[102] Park H et al. (2025) M^{3}KG-RAG: Multi-hop Multimodal Knowledge Graph-enhanced Retrieval-Augmented Generation. arXiv [cs.CL, cs.AI].

[103] Park J et al. (2025) Emulating Retrieval Augmented Generation via Prompt Engineering for Enhanced Long Context Comprehension in LLMs. arXiv [cs.CL].

[104] Patel P (2025) Graph-Enhanced Retrieval-Augmented Question Answering for E-Commerce Customer Support. International Journal of Complexity in Applied Science and Technology. https://doi.org/10.1504/ijcast.2025.10073851

[105] Perron BE et al. (2025) AI-Enhanced Social Work: Developing and Evaluating Retrieval-Augmented Generation (RAG) Support Systems. Journal of Social Work Education. https://doi.org/10.1080/10437797.2024.2411172

[106] Peruzzo E et al. (2025) RagMe: Retrieval Augmented Video Generation for Enhanced Motion Realism. International Conference on Multimedia Retrieval. https://doi.org/10.1145/3731715.3733417

[107] Qi J et al. (2024) RAG-Optimized Tibetan Tourism LLMs: Enhancing Accuracy and Personalization. arXiv.org. https://doi.org/10.1145/3703935.3704112

[108] Qin Q et al. (2025) Towards Adaptive Memory-Based Optimization for Enhanced Retrieval-Augmented Generation. Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2504.05312

[109] Raja R, A. Vats (2025) Beyond Nearest Neighbors: Semantic Compression and Graph-Augmented Retrieval for Enhanced Vector Search. Proceedings of the 3rd International Workshop on Deep Multimodal Generation and Retrieval. https://doi.org/10.48550/arXiv.2507.19715

[110] Rangan K, Yiqiao Yin (2024) A fine-tuning enhanced RAG system with quantized influence measure as AI judge. Scientific Reports. https://doi.org/10.1038/s41598-024-79110-x

[111] Rani M (2024) To Enhance Graph-Based Retrieval-Augmented Generation (RAG) with Robust Retrieval Techniques. 2024 18th International Conference on Open Source Systems and Technologies Icosst 2024 Proceedings. https://doi.org/10.1109/ICOSST64562.2024.10871140

[112] Rezaei MR, Adji Bousso Dieng (2025) Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. arXiv.org. https://doi.org/10.48550/arXiv.2502.11228

[113] Rouhollahi A et al. (2025) RAGnosis: Retrieval-Augmented Generation for Enhanced Medical Decision Making. medRxiv. https://doi.org/10.1101/2025.06.11.25329438

[114] Saha B, Utsha Saha, Muhammad Zubair Malik (2024) Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance. IEEE Access. https://doi.org/10.1109/access.2024.3513155

[115] Saha B, Utsha Saha, Muhammad Zubair Malik (2025) QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3513155

[116] Samuel DJ et al. (2025) AgroLLM: Connecting Farmers and Agricultural Practices through Large Language Models for Enhanced Knowledge Transfer and Practical Application. arXiv.org. https://doi.org/10.48550/arXiv.2503.04788

[117] Sawarkar K, Abhilasha Mangal, S. Solanki (2024) Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. Conference on Multimedia Information Processing and Retrieval. https://doi.org/10.1109/MIPR62202.2024.00031

[118] Shahmansoori A (2024) Concurrent Brainstorming & Hypothesis Satisfying: An Iterative Framework for Enhanced Retrieval-Augmented Generation (R2CBR3H-SR). arXiv [cs.IT, cs.AI].

[119] Sharma C (2025) Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. arXiv.org. https://doi.org/10.48550/arXiv.2506.00054

[120] Shavaki MA et al. (2024) Knowledge Graph Based Retrieval-Augmented Generation for Multi-Hop Question Answering Enhancement. Conference on Information and Knowledge Technology. https://doi.org/10.1109/IKT65497.2024.10892619

[121] Shen L et al. (2025) Retrieval-Augmented Image Captioning and Generation with Entity Concepts Enhancement for Baidu Multimodal Advertising. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3726302.3731957

[122] Shen Y et al. (2025) MGPRAG: Enhancing Medical Large Language Models via Precision Retrieval-Augmented Generation. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3624887

[123] Shen Z et al. (2026) HG-RAG: Hierarchical Graph-Enhanced Retrieval-Augmented Generation for Power Systems. Electronics. https://doi.org/10.3390/electronics15071445

[124] Shen Z et al. (2024) GeAR: Graph-enhanced Agent for Retrieval-augmented Generation. arXiv [cs.CL, cs.AI].

[125] Shi Y et al. (2024) ERAGent: Enhancing Retrieval-Augmented Language Models with Improved Accuracy, Efficiency, and Personalization. European Conference on Artificial Intelligence. https://doi.org/10.3233/FAIA240748

[126] Shi Y et al. (2024) Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems. arXiv.org. https://doi.org/10.48550/arXiv.2407.10670

[127] Siang OC et al. (2024) SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. npj Digital Medicine. https://doi.org/10.1038/s41746-024-01391-3

[128] Singh M et al. (2025) Leveraging Retrieval Augmented Generative LLMs For Automated Metadata Description Generation to Enhance Data Catalogs. arXiv [cs.IR, cs.CL].

[129] Staple BD et al. (2025) Comparison of Domain-Specific and Ensemble Large Language Models in Surgical Education: A Preliminary Performance Evaluation. Advances in Neurology and Neuroscience. https://doi.org/10.33140/an.08.02.03

[130] Su M et al. (2024) An Effective Retrieval Method to Improve RAG Performance. International Conference on Data Science and Information Technology. https://doi.org/10.1109/DSIT61374.2024.10881380

[131] Sun Y et al. (2025) KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering. arXiv [cs.CL, cs.AI].

[132] Superbi J et al. (2024) Enhancing Large Language Model Performance on ENEM Math Questions Using Retrieval-Augmented Generation. Anais do XVIII Brazilian e-Science Workshop (BRESCI 2024). https://doi.org/10.5753/bresci.2024.243977

[133] Swathi D, Dr. P. Jyotsna (2025) Enhanced Retrieval-Augmented Generation For Student Mental Health Support Using Generative Artificial Intelligence. International Journal of Computer Engineering & Technology. https://doi.org/10.34218/ijcet_16_01_277

[134] Tan T et al. (2024) PEAR: Position-Embedding-Agnostic Attention Re-weighting Enhances Retrieval-Augmented Generation with Zero Inference Overhead. The Web Conference. https://doi.org/10.1145/3696410.3714795

[135] Tan Y et al. (2025) Dynamic Parametric Retrieval Augmented Generation for Test-time Knowledge Enhancement. arXiv [cs.CL, cs.AI].

[136] Tang M et al. (2025) Injecting External Knowledge into the Reasoning Process Enhances Retrieval-Augmented Generation. arXiv [cs.IR, cs.CL].

[137] Tang X et al. (2024) Adapting to Non-Stationary Environments: Multi-Armed Bandit Enhanced Retrieval-Augmented Generation on Knowledge Graphs. arXiv [cs.AI, cs.CL].

[138] Tanyildiz D, S. Ayvaz, M. Amasyalı (2024) Enhancing Retrieval-Augmented Generation Accuracy with Dynamic Chunking and Optimized Vector Search. Orclever Proceedings of Research and Development. https://doi.org/10.56038/oprd.v5i1.516

[139] Taschetto L, Renato Fileto (2024) Using Retrieval-Augmented Generation to improve Performance of Large Language Models on the Brazilian University Admission Exam. Brazilian Symposium on Databases. https://doi.org/10.5753/sbbd.2024.243137

[140] Tayebi S et al. (2024) RadioRAG: Factual Large Language Models for Enhanced Diagnostics in Radiology Using Dynamic Retrieval Augmented Generation. arXiv.org. https://doi.org/10.48550/arXiv.2407.15621

[141] Telemala JP et al. (2025) Towards Enhanced Agricultural Information Access in Kiswahili: Integrating Knowledge Graphs and Retrieval-Augmented Generation. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3726302.3730271

[142] Tohir H, N. Merlina, Muhammad Haris (2024) Utilizing Retrieval-Augmented Generation In Large Language Models To Enhance Indonesian Language NLP. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). https://doi.org/10.33480/jitk.v10i2.5916

[143] Tong L, Quan Ge (2024) Achieving Higher Factual Accuracy in Llama LLM with Weighted Distribution of Retrieval-Augmented Generation. Preprint. https://doi.org/10.31219/osf.io/ctw8v

[144] Tourani A, Fatemeh Nazary, Yashar Deldjoo (2025) RAG-VisualRec: An Open Resource for Vision- and Text-Enhanced Retrieval-Augmented Generation in Recommendation. arXiv.org. https://doi.org/10.48550/arXiv.2506.20817

[145] Tran H et al. (2024) RARE: Retrieval-Augmented Reasoning Enhancement for Large Language Models. Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2412.02830

[146] Ünlü O et al. (2024) Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. medRxiv.

https://doi.org/10.1101/2024.02.08.24302376

[147] Upadhyay R, Marco Viviani (2025) Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. Discover Computing. https://doi.org/10.1007/s10791-025-09505-5

[148] Van de Schoot R, de Bruin J, Schram R, et al (2021) An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 3:125–133

[149] Wada A et al. (2025) Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. npj Digit. Medicine. https://doi.org/10.1038/s41746-025-01802-z

[150] Wang D et al. (2024) Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. Journal of Medical Internet Research. https://doi.org/10.2196/58041

[151] Wang H et al. (2025) Path Pooling: Training-Free Structure Enhancement for Efficient Knowledge Graph Retrieval-Augmented Generation. arXiv [cs.AI, cs.CL].

[152] Wang P, Sheng Li, Takahiro Shinozaki (2025) RAG-Boost: Retrieval-Augmented Generation Enhanced LLM-based Speech Recognition. arXiv [eess.AS, cs.CL].

[153] Wang R et al. (2025) Neural Graph Matching Improves Retrieval Augmented Generation in Molecular Machine Learning. International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2502.17874

[154] Wang Z et al. (2024) M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions. Preprint. https://doi.org/10.18653/v1/2024.acl-long.108

[155] Wei C et al. (2025) Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model. Journal of Artificial Intelligence Research. https://doi.org/10.1613/jair.1.17809

[156] Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, pp 1–10

[157] Woo JJ et al. (2024) Custom Large Language Models Improve Accuracy: Comparing Retrieval Augmented Generation and Artificial Intelligence Agents to Non-Custom Models for Evidence-Based Medicine.. Arthroscopy: The Journal of Arthroscopy And Related. https://doi.org/10.1016/j.arthro.2024.10.042

[158] Wu D et al. (2025) Integrating Chain-of-Thought and Retrieval Augmented Generation Enhances Rare Disease Diagnosis from Clinical Notes. arXiv.org. https://doi.org/10.48550/arXiv.2503.12286

[159] Wu JW (2025) Developing a Local Generative AI Teaching Assistant System: Utilizing Retrieval-Augmented Generation Technology to Enhance the Campus Learning Environment. Electronics Switzerland. https://doi.org/10.3390/electronics14173402

[X] Wu J, Cheng-Chu Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, Julian McAuley (2024) CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-Tail Recommendation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), pp 3391–3401.

[160] Wu Y et al. (2024) CoTKR: Chain-of-Thought Enhanced Knowledge Rewriting for Complex Knowledge Graph Question Answering. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2409.19753

[161] Xie Z et al. (2026) Zero-Shot Robotic Manipulation via 3D Gaussian Splatting-Enhanced Multimodal Retrieval-Augmented Generation. arXiv [cs.RO].

[162] Xiong X, Mingliang Zheng (2024) Merging Mixture of Experts and Retrieval Augmented Generation for Enhanced Information Retrieval and Reasoning. Research Square. https://doi.org/10.21203/rs.3.rs-3978298/v1

[163] Xue J et al. (2024) Enhanced Multimodal RAG-LLM for Accurate Visual Question Answering. arXiv.org. https://doi.org/10.48550/arXiv.2412.20927

[164] Yang K et al. (2025) UrbanMind: Towards Urban General Intelligence via Tool-Enhanced Retrieval-Augmented Generation and Multilevel Optimization. arXiv [cs.LG, cs.AI].

[165] Yang Z et al. (2025) Integrating Ranking-Based Query Rewriting into RAFT for Improved Retrieval-Augmented Generation. IEEE International Conference on Consumer Electronics. https://doi.org/10.1109/ICCE-Taiwan66881.2025.11207967

[166] Yao Y, Xiaozhong Lyu (2025) Multi-Agent Enhanced Retrieval-Augmented Generation for Open-Domain Table Question Answering. International Conference on Advanced Computer Theory and Engineering. https://doi.org/10.1109/ICACTE66284.2025.11412132

[167] Yazaki M et al. (2024) Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model. Prehospital Emergency Care. https://doi.org/10.1080/10903127.2024.2374400

[168] Ye C (2024) Exploring a learning-to-rank approach to enhance the Retrieval Augmented Generation (RAG)-based electronic medical records search engines. Informatics and Health. https://doi.org/10.1016/j.infoh.2024.07.001

[169] Ye Z et al. (2025) IHGR-RAG: An Enhanced Retrieval-Augmented Generation Framework for Accurate and Interpretable Power Equipment Condition Assessment. Electronics. https://doi.org/10.3390/electronics14163284

[170] Yi Q et al. (2025) SCORE: Story Coherence and Retrieval Enhancement for AI Narratives. arXiv.org. https://doi.org/10.48550/arXiv.2503.23512

[171] Yu X et al. (2024) DroidCoder: Enhanced Android Code Completion with Context-Enriched Retrieval-Augmented Generation. International Conference on Automated Software Engineering. https://doi.org/10.1145/3691620.3695063

[172] Yuan X et al. (2025) mKG-RAG: Multimodal Knowledge Graph-Enhanced RAG for Visual Question Answering. arXiv.org. https://doi.org/10.48550/arXiv.2508.05318

[173] Yuan Y et al. (2024) A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. arXiv.org. https://doi.org/10.48550/arXiv.2408.05141

[174] Zhang C et al. (2025) A Knowledge-Enhanced Platform (MetaSepsisKnowHub) for Retrieval Augmented Generation–Based Sepsis Heterogeneity and Personalized Management: Development Study. Journal of Medical Internet Research. https://doi.org/10.2196/67201

[175] Zhang H et al. (2025) Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. arXiv [cs.IR, cs.AI].

[176] Zhang H, M. O. Shafiq (2025) TriRAG: Enhancing Retrieval-Augmented Generation Method with Triple-Based Knowledge Graphs for Improved Question Answering. International Computer Science Conference. https://doi.org/10.1109/ICSC64641.2025.00011

[177] Zhang L et al. (2024) RAG-Enhanced Commit Message Generation. arXiv.org. https://doi.org/10.48550/arXiv.2406.05514

[178] Zhang S et al. (2025) Retrieval-Augmented Generation for Medical Question Answering on a Heart Failure Dataset: Performance Analysis.. JMIR Formative Research.

https://doi.org/10.2196/84932

[179] Zhang Z, Yang Wu, Xiaolong Zheng (2024) Social Intelligence Enhancement Methods for LLM Agents Based on Retrieval-Augmented Generation. ACM Cloud and Autonomic Computing Conference. https://doi.org/10.1109/CAC63892.2024.10865026

[180] Zhao Y et al. (2025) A Graph-Retrieval-Augmented Generation Framework Enhances Decision-Making in the Circular Economy. arXiv [cs.AI, cs.CL].

[181] Zheng Y et al. (2025) Integrating retrieval-augmented generation for enhanced personalized physician recommendations in web-based medical services: model development study. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2025.1501408

[182] Zhou P et al. (2025) Context-Adaptive Synthesis and Compression for Enhanced Retrieval-Augmented Generation in Complex Domains. arXiv [cs.CL].