technologies

MDPI

Review

Retrieval Augmented Generation for Large Language Models: A Methodological Survey

Vasileios P. Rekkas * and Nicholas Vretos

Visual Computing Lab, Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece; vretos@iti.gr
* Correspondence: vrekkas@iti.gr

Abstract

Over the past years, large language models (LLMs) have gained significant attention and have led to major breakthroughs in natural language processing (NLP). Despite their remarkable efficiency and contextual understanding, LLMs still remain limited by their tendency to generate hallucinated text, their reliance on static internal knowledge, and their limited interpretability. To address these limitations, retrieval-augmented generation (RAG) techniques have emerged as a hybrid architecture that couples the generative capabilities of LLMs with external knowledge retrieval processes. By drawing on external knowledge at inference time, RAG architectures are intended to enhance the accuracy and relevance of their generated output, and they have already been widely used in areas where extensive, detailed and up-to-date knowledge is needed. For this study, Scopus, IEEE Xplore, and the ACM Digital Library were explored and utilized, with Google Scholar used only for backward and forward citation checking, resulting in 191 peer-reviewed journal articles published between January 2024 and April 2026. This methodological review presents the research efforts regarding the use of RAG in LLMs for a variety of application fields. Additionally, this review summarizes the main RAG architectural approaches, deployment strategies, and evaluation metrics. Furthermore, the main open issues and future directions are examined, to motivate further research on the field.

Keywords: retrieval-augmented generation (RAG); large language models (LLMs); natural language processing (NLP); generative artificial intelligence (GenAI)

1. Introduction

Generative artificial intelligence (GenAI) has emerged as a prominent technology in natural language processing (NLP), enabling multimodal human-like outputs, including text, images, and audio. In GenAI, large language models (LLMs), a transformer-based AI architecture, have been extensively trained on extensive corpora to be able to understand, analyze and ultimately produce human-like text [1]. LLMs have been widely utilized to generate human-like, consistent, and contextually grounded text/responses across a vast spectrum of applications. Their general-purpose nature and flexibility across domain-specific settings have enabled a widespread adoption in multiple areas, including healthcare, finance, research, law, and education [2]. Compared with traditional recurrent and convolutional neural networks (RNNs and CNNs), LLMs can achieve enhanced performance and parallel processing capability, as well as more efficient training for sequence-to-sequence tasks [1].

Check for updates

Received: 4 October 2026
Revised: 20 October 2026
Accepted: 25 October 2026
Published: 2 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

However, despite their promising potential in addressing complex problems, LLMs still face critical limitations when deployed in real-world settings. Their probabilistic nature often results in incorrect or misleading answers, a phenomenon commonly known as hallucination, which raises serious concerns, particularly in settings where accuracy and trustworthiness are essential [3]. Furthermore, the static knowledge embedded in LLMs leads to a significant reduction in the accuracy of the generated text [4]. Additionally, the limited explainability of LLM-generated responses leads to reduced user trust in the AI-generated text [5].

To address these shortcomings, the LLM models can be augmented with external knowledge sources through retrieval-augmented generation (RAG) [6]. By combining retrieval-based and generative architectures, RAGs enable a generative model to query an external corpus at inference time, identify the relevant documents within a structured or unstructured knowledge base, and utilize the retrieved information to ground the generated response. Thus, in the RAG methodology, the parametric knowledge acquired during pretraining is combined with non-parametric knowledge retrieved on demand [6]. RAG systems are intended to reduce hallucination, improve transparency and increase factual accuracy, depending on the quality of what is retrieved, as irrelevant, outdated or conflicting evidence can also mislead the generator [1].

1.1. Related Work

The exponential technological breakthrough in NLP has accelerated the evolution of LLM capabilities, while the increasing need for more accurate and up-to-date AI-generated text trustworthiness in terms of explainability has stimulated significant interest in RAG architecture. As a result, RAGs have attracted significant attention from the research community, leading to several review efforts on the topic, analyzing both the architectural design and the retrieval mechanisms as well as the application domain analysis.

A comprehensive review of RAG implementation is presented in [1], where the main components of RAG architectures, the retrieval and generation methods in the legal domain are considered. In addition, the survey also analyzes privacy issues along with ethical considerations, as well as listing the most commonly utilized datasets. However, the analysis and scope of the study are restricted to the legal domain; thus they do not provide broader applications and perspectives of RAG methodologies across domains.

A survey is presented in [7] with the main focus being the use of RAG-driven methods in conversational LLMs. The review focuses on the exploration of how RAGs, vector databases, and alternative approaches (e.g., knowledge graphs) can address the challenges that arise in the conversational analysis for various sub-domains in the field of agriculture. This survey offers valuable insights concerning the memory development on conversational LLMs, but is limited without offering a wider analysis of RAG architectures and evaluation measurements in other fields.

Gao et al. [8] provide a comprehensive survey of the evolution of RAG architectures in LLMs structures and include three main paradigms: naive, advanced, and modular RAGs. Additionally, the review analyzes the retrieval, generation, and augmentation stages of the RAG architecture and presents the major methods, tasks, datasets, and challenges in a variety of applications including math, text generation, code, and code research, among others. However, given the rapid evolvement in RAG technology, a more analytic and up-to-date presentation of the RAG architectures, application, and benchmarks is required.

A scoping review of the RAG utilization for LLMs in the medical and nursing domains is presented in [9]. The survey summarizes the state-of-the-art (SOTA) RAG methodologies in the two domains, focusing on agentic, multimodal, text- and graph-based RAGs. Furthermore, the authors include a broad research on the reasoning strategies, and perform

a taxonomy of stages in the RAG architectures, namely, knowledge retrieval, integration, and generation, as well as intent identification. While this review performs a wide summarization of current SOTA and future directions, its focus on the two domains is a limitation, considering the understudied use of RAG architectures in cross-domain applications.

Zheng et al. [10] survey retrieval augmentation in computer vision, organizing existing methods into retrieval-augmented visual understanding and visual generation, and extending the discussion to embodied AI. While this survey offers a broad view of RAG for visual tasks, its focus on computer vision leaves text-centered and cross-domain applications of RAG for LLMs outside its scope.

Table 1 compares the present review with existing RAG reviews in terms of review period and corpus size, taxonomic dimension, application coverage and reproducibility of selection. Most reviews with a reported protocol searched several bibliographic databases and admitted conference papers or preprints, while the present corpus is restricted to peer-reviewed journal and magazine articles retrieved from Scopus, IEEE Xplore and the ACM Digital Library.

Table 1. Comparison of the present review with related RAG reviews; n/s, not stated.

Review	Period; Corpus	Taxonomic Dimension	Application Coverage	Reproducibility of Selection
[11]	n/s	4-stage IR pipeline (pre-/retrieval/post-/generation)	Cross-domain	No protocol stated
[12]	n/s; 200+	Graph approaches & functionalities	Cross-domain	No protocol stated
[13]	n/s	Graph indexing, retrieval, generation	Cross-domain	No protocol stated
[14]	n/s	Reasoning, acting, interacting	Cross-domain; broader than RAG	No protocol stated
[1]	2020–2024; 22	Pipeline components, precision/interpretability	Legal only	Protocol reported; 22 studies
[7]	2019–2024; 113	Memory architecture in conversational LLMs	Agriculture	PRISMA; 223 identified, 113 included
[2]	2020–Mar 2025; 51	Indexing/retrieval/generation	Education only	Single database; search date & query reported
[15]	Jan 2020–Feb 2025; 70	Methods, datasets, evaluation techniques	Healthcare	Systematic review; sources & count reported
[16]	2023–2024; 20	Clinical task, retrieval source/strategy, baseline	Biomedicine	PRISMA; 335 identified, 20 included
[9]	Nov 2022–May 2025; 67	Functional type (text, KG, agentic, multimodal, plug-and-play)	Medicine, nursing	PRISMA-ScR; 917 retrieved, 67 included
[17]	2022–2025; 30	Clinical application (diagnostic, EHR, QA)	Healthcare	PRISMA; 123 identified, 30 included; dual screening
[18]	2020–May 2025; 128	RAG stack components (retrieval, indexing, chunking, encoders, training, generation)	Cross-domain; citation-thresholded	PRISMA 2020; 4721 identified, 128 included; dual screening
[19]	2017–Jul 2025; 251	Architecture (naive/advanced/modular), data source, method, task	Medicine	PRISMA-ScR; 3980 identified, 251 included; dual screening
[8]	n/s	Paradigms (naive/advanced/modular); retrieval, generation, augmentation	Cross-domain	No protocol stated
[10]	n/s	Visual understanding, visual generation, embodied AI	Computer vision	No protocol stated
[20]	n/s	Architectures, training strategies, applications	Cross-domain	No protocol stated
[21]	n/s	Evaluation targets, datasets and metrics	Cross-domain	No protocol stated
This review	Jan 2024–27 Apr 2026; 191	Locus of architectural intervention	Cross-domain	PRISMA-ScR; strings, fields, filters, dates, counts, exclusions; subcategory mapped to rule

The reviews in Table 1 indicate that the seventeen reviews fall into two groups. Those reporting an explicit selection procedure are, with one exception, bounded by application domain: five to medicine and healthcare [9,15–17,19], one to education [2], one to the legal domain [1] and one to agriculture [7]. The narrative surveys of Gao et al. and [11–14,20,21] and the vision-specific survey of Zheng et al. report neither search strings, corpus sizes

nor eligibility rules. The one cross-domain review with a reported protocol [18] follows PRISMA 2020 with dual screening, but restricts its corpus to highly cited studies published up to May 2025 and admits conference papers and preprints. Corpus sizes among the protocol-based reviews range from 20 to 251 studies, and only 1 review [16] performs a quantitative synthesis, reporting a pooled odds ratio of 1.35 (95% CI 1.19–1.53) across 20 clinical studies.

1.2. Motivation and Contributions

The rapid evolution of the RAG ecosystem, expanding from conventional vanilla RAG pipelines to more sophisticated architectures that include graph-based knowledge utilization and agentic as well as multimodal reasoning, has led to several reviews on the topic in a wider context of describing the architecture and application of RAG models. Although these studies examine various applications of RAG methodologies, their rapid evolution and growing attention on more accurate LLMs requires a broader and more up-to-date methodological review. Additionally, the proposed taxonomy offers a principled framework for organizing the RAG literature according to the main focus of innovation within the pipeline, enabling a more systematic comparison of recent methods, and highlights the main research directions that define modern RAG systems. The current review is multidisciplinary and provides a detailed PRISMA-ScR selection process for peer-reviewed journal articles, suggesting an architectural intervention taxonomy, while classifying each article at the paper level. The review also includes a description of the level of architectural intervention and the classification of each study included. Our contributions can be summarized as follows:

- A methodological review of 191 peer-reviewed journal articles on RAG for LLMs published between January 2024 and 27 April 2026 is presented, offering a broad analysis of architectural designs, retrieval and generation strategies, and evaluation metrics across a variety of application settings and domains.
- A novel taxonomy of RAG architectures, based on the main focus of innovation within the pipeline, is introduced, generating six RAG family categories: Retrieval-Optimized RAG, Orchestrated RAG, Structured-Memory RAG, Multimodal-Grounded RAG, Governed RAG, and Task-Compositional RAG.
- A thorough review of RAG methods in LLM systems is discussed. In particular, in the review, the orchestration strategies, architectural trends, and knowledge substrates are analyzed, highlighting the RAG evolution from vanilla naive generation methodologies to more complex and advanced frameworks.
- Open challenges and future trends are discussed to highlight the increasing role of RAG in complex end-to-end decision-support systems and to motivate academia and industry to further explore the field.

The remainder of this article is organized as follows. Section 2 introduces the proposed six-family RAG taxonomy. Section 3 categorizes the reviewed corpus based on application domains, followed by a statistical analysis of the taxonomy in Section 4. Section 5 outlines the foundational concepts and architectural evolution of RAG systems. Section 6 systematically presents the evaluation metrics utilized for assessing RAG pipelines. Section 7 discusses the literature grouped by the proposed methodological families. Finally, Section 8 highlights open challenges and future research directions in the field, and Section 9 presents the conclusions derived from the review.

2. Inclusion and Proposed Taxonomy

2.1. Research Questions

The scope of this study is defined by a set of research questions formulated before the search was executed. These research questions determine the eligibility criteria, the data extracted from each included study, and the structure of the synthesis:

- RQ1. Along which architectural dimension do the peer-reviewed RAG systems for LLMs principally innovate, and can those dimensions be organized into a set of mutually distinguishable methodological families?
- RQ2. How are the reviewed RAG systems distributed across the proposed families, subcategories and application domains?
- RQ3. Which mechanisms characterize each family, and which benefits and trade-offs do the reviewed studies report?
- RQ4. Which evaluation metrics do the reviewed studies report in their results, and which aspects of retrieval, grounding and generation quality remain unmeasured?
- RQ5. Which open problem does each family leave unresolved, and which research directions follow from the cost and evaluation gaps identified under RQ3 and RQ4?

Table 2 shows the research questions, the data extracted from the included studies, and the section where they are addressed. The answers to the five questions are summarized in Section 7.7.

Table 2. Mapping of research questions to extracted data and answering sections.

RQ	Data Extracted per Study	Answered In
RQ1	Architectural component being the principal contribution → assigned family → governing rule	Section 2
RQ2	Primary family → subcategory → application domain	Sections 3 and 4
RQ3	Mechanism → reported benefit; reported cost or limitation	Section 7
RQ4	Metric groups with reported values, coded for every included study	Section 6
RQ5	Limitations and future work stated by the study; open question per family	Section 8

2.2. Search Strategy and Study Selection

This study's main focus is the way the RAG architecture has developed since search augmentation became a standard component in deployed LLM systems. This reflects a point in time when graph-structured memories, agent-based control, multimodal routing and retrieval, and governance-aware retrieval are accepted as methodological advances. It is this diversification which our taxonomy seeks to encompass and to use to outline the recent evolution of RAG systems. This study adopts a methodological review approach, reported in accordance with the PRISMA extension for scoping reviews (PRISMA-ScR) [22], an extension of the PRISMA 2020 statement [23]. The scoping extension was selected in preference to the core PRISMA 2020 checklist because the objective of this study is to map the methodological structure of a rapidly diversifying field, namely, RAG. Consequently, the PRISMA 2020 items concerning effect measures, risk-of-bias assessment, meta-analytic synthesis, and certainty of evidence are not applicable to this synthesis. The protocol of this review was defined internally before the final search was executed and is reported in full below. PRISMA-ScR was preferred to the Kitchenham guidelines for systematic reviews in software engineering [24] because the corpus spans clinical, engineering and social-science venues, for which PRISMA is the common reporting standard. The review protocol was not registered.

A search of the literature was conducted iteratively between January and April 2026, with the initial search on 13 January 2026 and repeated at intervals to capture newly indexed records, while the final consolidating search was performed on 27 April 2026. All counts

reported in this section and in Figure 1 derive from that final search, which supersedes the intermediate iterations, reflecting the rapid publication rate in this field.

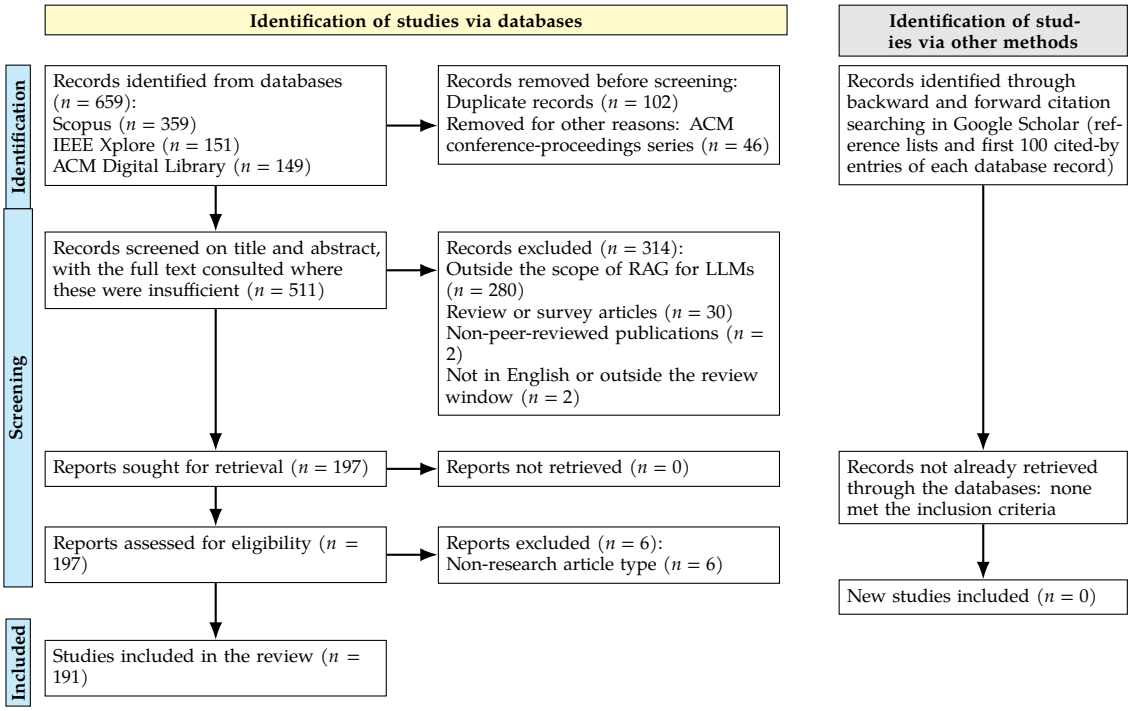


Figure 1. PRISMA 2020 flow diagram of study selection [23]. No study identified through citation searching was added to the corpus.

To be included in this review, studies had to meet the following criteria:

- Be written in English;
- Be published in peer-reviewed journals or magazines;
- Be published within the selected review period (January 2024 to 27 April 2026);
- Focus on RAG utilization for LLMs, including architectural design, retrieval and generation strategies, evaluation, or domain-specific deployment;
- Report original research, namely, an RAG method or system together with its implementation or evaluation, or an empirical evaluation of RAG.

The following exclusion criteria were applied:

- Book chapters, editorials, theses, white papers, and other non-peer-reviewed publications;
- Conference proceedings;
- Duplicate records;
- Review or survey papers;
- Article types that do not report original research, such as catchwords, perspectives, commentaries and conceptual or vision articles without an implementation or evaluation;
- Studies whose primary focus falls outside the scope of RAG, such as general LLM pretraining, standalone fine-tuning, hardware optimization, or unrelated AI applications.

Publication-type restriction. The classified corpus was restricted to peer-reviewed journal and magazine articles. This was a scoping decision rather than a judgment of quality: conferences such as ACL, EMNLP, NeurIPS and SIGIR apply rigorous peer review, and much influential RAG research appears there. The restriction was adopted for two reasons. First, journal and magazine articles are archival versions of record, while conference contributions are frequently revised or extended into later journal versions, which would have introduced duplicate or partially overlapping records. Second, journal and magazine articles in the disciplines covered by the corpus follow a broadly common reporting standard,

typically including the full method, the deployment setting and the evaluation, which made classification by architectural contribution feasible across fields as different as medicine, construction and cybersecurity. Magazine articles were retained because the magazines indexed in these databases are peer-reviewed archival venues that publish complete system descriptions. As a consequence, the corpus represents journal-published RAG research, and conference-first innovations enter it only once they reach journal publication.

A single core search string was adapted to the syntax of each database. The string combined terms for RAGs and LLMs:

```
( "retrieval-augmented generation" OR "retrieval augmented generation"
  OR "RAG" OR "retrieve-and-generate" )
AND
( "large language model" OR "LLM" OR "generative language model"
  OR "foundation model" OR "GPT" )
```

The core Boolean string above was adapted to the syntax and field structure of each source, and in Scopus was applied to the TITLE-ABS-KEY field, restricted to document type "Article" and to the publication years 2024–2026. In IEEE Xplore it was restricted to content type "Journals" over the same period. In the ACM Digital Library it was applied to the Title, Abstract and Keywords fields, restricted to journal articles over the same period. No language restriction was imposed at the query level, but only English-language records were retained during screening.

Google Scholar procedure. Google Scholar was not used for the primary searches, as it has no reliable API or bulk export, its Boolean search results are not stably reproducible, and it contains a considerable amount of non-peer-reviewed material. Its use was limited to backward and forward citation checking. For each record retrieved from the primary databases, we also screened the reference list and the first 100 relevance-sorted entries of the "cited by" list, and we only included a paper if it also met all the other requirements of the primary databases. Preprints and other non-peer-reviewed documents found in this way were excluded at screening; therefore, the primary-database search is reproducible even with the exclusion of Google Scholar.

Deduplication. All records were exported in BibTeX format and merged into a single reference-management library. Scopus gave 359 results from the TITLE-ABS-KEY field, IEEE Xplore gave 151, with the search restricted to document titles, and ACM Digital Library gave 149, of which 46 appeared in the Proceedings of the ACM and Proceedings of the VLDB Endowment series and were excluded by the publication-type criterion. The remaining 103 ACM journal records were merged with the Scopus and IEEE Xplore records, giving 613 records in total, or 659 before the removal of the ACM conference-proceedings series. Then duplicates were identified in two steps: (i) by comparing the Digital Object Identifier and, in the absence of a DOI, by matching a normalized title–first author–year triple, with titles lowercased and stripped of punctuation and diacritics, and (ii) a manual review of every flagged pair to distinguish true duplicates from distinct works with similar titles. This removed 102 duplicates, leaving 511 unique records.

Eligibility assessment. The 511 unique records were assessed against the inclusion and exclusion criteria stated above. Title and abstract were reviewed first, and the full text was checked whenever these were not enough to reach a decision. A total of 314 records were excluded, 280 of them for being outside the scope of RAG for LLMs, including studies of general LLM pretraining, standalone fine-tuning, hardware optimization, and records retrieved through acronym collisions on the term "RAG". A further 30 were review, survey, tutorial or scoping-review articles, identified by title keywords in the merged export; 2 were magazine items whose peer-review status could not be established; and 2 were not in English or fell outside the review window. As a final eligibility check, every included study was

re-examined against the article-type criterion: its title, venue and content were checked, and the article type assigned by the publisher was verified for each study whose title, venue or content suggested a non-research article. Six studies were excluded at this stage, namely, a catchword, a current perspective, a conceptual overview and three conceptual or vision articles without an implementation or evaluation. Research letters and tutorials were retained when they report an original system or evaluation. This process yielded a final corpus of 191 peer-reviewed journal articles.

2.3. Proposed Taxonomy

The proposed taxonomy introduces six distinct methodological families within the RAG ecosystem. Instead of organizing them by domain or surface-level architecture, it arranges the studies by their core construction, and by the architectural component that is most substantially redesigned. This approach separates RAG methods based on how they search and rank data, and how they manage the retrieval process step by step. The families are therefore defined by the locus of the principal intervention relative to a baseline retrieve-then-generate pipeline, whether this locus is the retrieval mechanism, the control policy, the knowledge substrate, the evidence modality, a governance constraint or the role of retrieval within a wider workflow. As a result, this survey organizes the literature into six primary families: Retrieval-Optimized RAG, Orchestrated RAG, Structured-Memory RAG, Multimodal-Grounded RAG, Governed RAG, and Task-Compositional RAG. Figure 2 presents the proposed taxonomy and the corresponding methodological families.

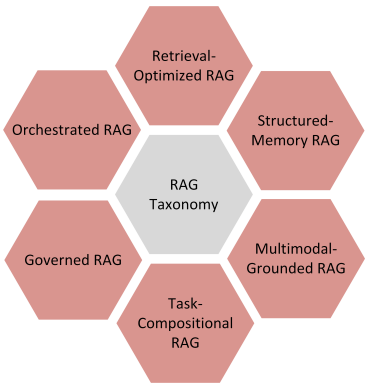


Figure 2. Proposed RAG taxonomy.

In this proposed taxonomy, each family addresses a different aspect of the architecture and tries to resolve a specific kind of problem, such as weak retrieval, fragmented knowledge, cross-modal grounding, or workflow integration. Thus, the taxonomy becomes more robust and analytical as it does not only consider domain application or model architecture.

The families are defined by the locus of the principal intervention relative to a baseline retrieve-then-generate pipeline: whether this locus is the retrieval mechanism, the control policy, the knowledge substrate, the evidence modality, a governance constraint, or the role of retrieval within a wider workflow. This principle yields a single primary assignment per study, as secondary properties are recorded separately and do not affect the partition. Adopting this primary classification scheme yields two direct consequences regarding how the corpus is partitioned and interpreted. Firstly, although a system is assigned to a single primary category to create a non-redundant partition for counting purposes, this purely practical convention does not imply that other branches of the taxonomy are not represented in a given implementation. Secondly, as the distribution reported in the taxonomy table reflects where authors situated their principal methodological contribution, these figures naturally understate the actual prevalence of individual architectural techniques, as prop-

erties such as structured substrates or adaptive control, as secondary components within systems whose core novelty resides elsewhere and are categorized under those primary headings instead.

2.3.1. Retrieval-Optimized RAG

This family of RAGs focuses on the enhancement of retrieval quality, while maintaining the core RAG architecture. The knowledge base remains primarily ordinary text segmented into documents or chunks, and the output text generator is kept largely unchanged. The improvements of this class are implemented on the retrieval side, utilizing better chunking mechanisms, enhanced embedding adaptation, hybrid search, multistage reranking, query refinements, and verification that the retrieved output is adequate and accurate [25–29].

The central goal of this family is to improve the relevance, completeness, and contextual precision of the evidence passed to the generator. As a result, these systems address common failure modes such as semantic mismatch, low recall, chunk granularity that misses the mark, and ranking that fails to prioritize appropriately. These systems rest on the assumption that a substantial share of hallucination and factual drift in RAG does not occur from the generator but from substandard retrieval evidence. This family also includes approaches that are structure parsing centered or focus on the reformulation of the architecture, the embedding-space alignment, the similarity assessments, the ones centered around similarity methods, and the retrieval alignment and benchmarking [30–32].

2.3.2. Orchestrated RAG

Orchestrated RAGs do not focus on enhanced retrieval scoring alone, but make their principal contribution by optimizing the control policy of retrieval. These systems treat the retrieval process as a dynamic or multistage procedure that may involve query rewriting, decomposition, routing to different external sources, iterative refinement, validation of the retrieved evidence, or agent-based planning over the full retrieval process [33–37].

The systems included in this family utilize an orchestration layer that determines the method, timing, and source of retrieved knowledge. Retrieval is transformed from a static lookup into a sequence of controlled decisions conditioned on query characteristics, intermediate outputs, or domain-specific heuristics. This family is especially relevant in hard settings characterized by underspecified queries, multi-intent questions, knowledge-dense inputs, or information distributed across heterogeneous sources [38–40].

2.3.3. Structured-Memory RAG

Structured-Memory RAG includes methods that modify the representational form of the external memory used by the RAG system. Rather than relying solely on flat text corpora and similarity-based retrieval over chunks, these approaches retrieve from explicitly structured substrates such as knowledge graphs, relational graphs, dual-index architectures, path-based stores, or other graph-like memory organizations [41,42].

The external knowledge base is organized so that relations between entities remain explicit instead of being dispersed across text chunks. Retrieval can then exploit the graph topology, the relations between nodes and query-relevant subgraphs, and can support multihop reasoning over the structured memory [43–45].

2.3.4. Multimodal-Grounded RAG

Multimodal-Grounded RAG comprises approaches in which retrieval and generation are grounded not just in text but in multimodal data, such as images, videos, remote sensing imagery, tables, visual documents, or mixed structured and visual evidence, while ensuring

evidence alignment. These systems integrate cross-modal evidence alignment, though sometimes the alignment is partial or task-dependent [46–48].

The context being retrieved in the RAGs of this family is not represented adequately as plain text only. Instead, the generation process relies on how the different evidence pieces, such as text, image, video or other modalities, semantically align with each other. Consequently, these systems must address relevance, modality fusion, and evidence integration alignment. This family is particularly important in domains where critical knowledge is distributed across multimodal data [46,49,50].

2.3.5. Governed RAG

Governed RAG denotes a class of methods that treat retrieval as a regulated interface that obeys rules beyond semantic relevance alone. In this family, the emphasis is on privacy, trustworthiness, authorization, policy compliance, and controlled knowledge exposure, such that the architecture explicitly governs the retrieval, exposure, retention, or suppression of knowledge [51].

In this family, retrieval should follow governance rules and operational limitations, which may conflict with raw performance. These systems therefore incorporate mechanisms such as encrypted search, access-control policies, privacy-preserving perturbation, secure retrieval, provenance-aware filtering, and safeguards that determine which evidence may be exposed to the generator or the user [52,53]. For this defined family, the core mechanism of the architecture is formulated not only on privacy alone, but includes tweaking the core design based on access and control parameters.

2.3.6. Task-Compositional RAG

Task-Compositional RAG refers to methods in which RAG functions as one component within a broader task-oriented workflow. In such systems, the final output is not limited to free-form question answering; instead, the retrieval and generation parts are tied into a structured process that produces a task-specific output, such as a diagnosis, classification, compliance checklist, executable query, recommendation, report, control plan, task-bounded guidance, or other downstream operational output [54–56].

For this family, the retrieval serves as an intermediate grounding step in a larger chain of reasoning or action. Consequently, system performance is evaluated not only in terms of answer faithfulness, but also by how useful the output is for the next step, the next decision, or the operational requirement of the surrounding pipeline. This family is especially common in enterprise, healthcare, industrial, legal, and operational decision-support settings, where RAG is one stage in an end-to-end intelligent workflow rather than a standalone question-answering module [57–60]. The family is therefore defined by the position of retrieval in the system rather than by the form of its output. Question-answering and guidance assistants belong to it only when they are built for a defined operational task, such as postoperative care, customs clearance or security-event response, and are evaluated at the level of that task. General question answering that is not tied to such a task, and comparisons of RAG with non-retrieval baselines, are assigned by rules 4–8.

2.4. Taxonomy Assignment Rules

To improve reproducibility and reduce bias, each paper in the corpus was classified using the rules presented below. As many RAG systems are complex systems combining various mechanisms, the rules were devised to isolate the central architecture of each study from its other properties. The rules are presented here as they were used in classifying the articles:

1. Main focus of innovation: Each paper was assigned to the family corresponding to the architectural component that constitutes its main methodological contribution relative to a baseline RAG pipeline.
2. Unique Primary Family: Every paper received exactly one primary family, even where its contributions spanned several categories, keeping the taxonomy partitioned and preventing the same paper from being counted twice across families.
3. Secondary Characteristics: The characteristics of the other families, if a paper contains these, were considered in the decision to break a tie under rule 10 and were entered in the justification recorded with the decision.
4. Retrieval Improvement Without Control Policy: If the main contribution concerns chunking, embeddings, reranking, hybrid retrieval or retrieval benchmarking within an otherwise one-shot pipeline, the paper was assigned to Retrieval-Optimized RAG.
5. Explicit Retrieval Policy: If the main contribution introduces query rewriting, decomposition, self-correction, or planning over retrieval actions, the paper was assigned to Orchestrated RAG.
6. Structured External Memory: If the main contribution redesigns the external knowledge substrate into a graph, knowledge graph, or another explicitly structured memory form, the paper was assigned to Structured-Memory RAG.
7. Cross-Modal Evidence Alignment: If the principal mechanism requires joint grounding across text and non-text modalities, the paper was assigned to Multimodal-Grounded RAG. In cases where multimodality did not shape the redesign of a paper's baseline RAG architecture, it was classified according to its dominant mechanism, and its multimodal properties were considered secondary.
8. Governance-Centric Retrieval: If the main contribution concerns privacy, authorization, security constraints, or forgetting/unlearning at the retrieval layer, the paper was assigned to Governed RAG.
9. Workflow-Level Integration: If RAG is primarily embedded within a broader task pipeline whose output is a structured decision artifact, executable result, recommendation, control sequence, or task-bounded guidance consumed within a defined operational process, the paper was assigned to Task-Compositional RAG.
10. Tie-Breaking Criterion: Where a paper appeared to fit several families, it was assigned to the family whose removal would eliminate the paper's core methodological identity. This rule ensures that classification is based on principal contribution.

Coding procedure. The classification rules were drawn up jointly by both authors, and they were revised whenever a study revealed an ambiguity in their formulation. The first author then applied the final rules to all the included studies, recording the primary family, subcategory, governing rule, application domain, and a short explanation for each assignment. Where several families were relevant, rule 10 was applied.

Traceability of assignments. Each individual study is listed under its subcategory in Section 7 and in Supplementary Table S1.

Verification of study descriptions. Before the synthesis, the description of each included study in the text was compared with the study's title, abstract and full text. Inaccurate descriptions were corrected, and studies whose corrected description changed the classification were reclassified. In total, eight studies were reclassified, six to a different primary family and two to a different subcategory within the same family, including the two studies reassigned under the criterion for Task-Compositional RAG stated in Section 2.3, while all changes are listed in Supplementary Table S1.

Coding reliability. The primary classification was performed by the first author. Studies whose assignment was not clearly determined by the decision rules were discussed by both authors until consensus was reached, and these are recorded in Supplementary

Table S1. No second independent coder was available, so no inter-rater reliability statistic is reported, which is acknowledged as a limitation of the synthesis.

3. Application-Domain Categorization

The included corpus is distributed across application domains based on the main problem setting, deployment scenario, and stakeholder context. Each study was assigned to exactly one application category, and studies whose contribution was not tied to a specific domain were assigned to the General-Purpose/Cross-Domain category. This application-domain categorization describes the field in which an RAG is being applied, while the family taxonomy explains how the architecture is principally modified. Figure 3 illustrates the distribution of the included papers across the identified application categories. Studies that present a cross-domain applicability or focus on benchmarking are grouped under a separate "General-Purpose/Cross-Domain Methods and Evaluation" category.

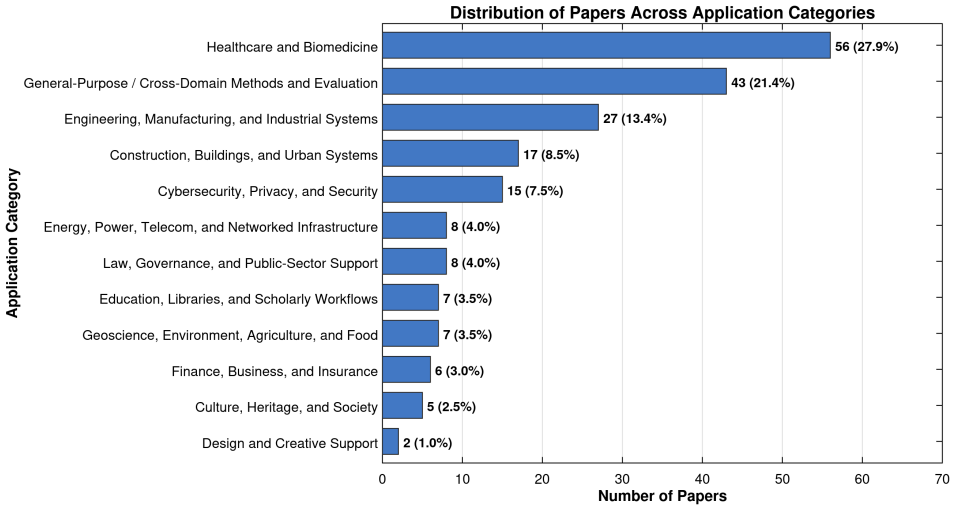


Figure 3. Distribution of the 191 reviewed studies across application categories.

Table 3 cross-tabulates the application domains with the methodological families. Healthcare and biomedicine is the most frequent domain for Retrieval-Optimized (16 of 41), Multimodal-Grounded (7 of 18) and Task-Compositional RAG (20 of 59). Orchestrated RAG is most often applied to general-purpose methods (11 of 27), more than half of the Governed RAG studies address cybersecurity and privacy (7 of 12), and Structured-Memory RAG is divided equally between engineering and industrial settings and healthcare (8 studies each).

Table 3. Included studies by application domain and methodological family.

Domain	RO	OR	SM	MG	GV	TC	Total
Healthcare and biomedicine	16	4	8	7	0	20	55
General-purpose/cross-domain	15	11	7	3	1	4	41
Engineering, manufacturing, industrial	3	3	8	3	2	5	24
Construction, buildings, urban	1	1	3	2	0	9	16
Cybersecurity, privacy, security	0	1	3	0	7	4	15
Education, libraries, scholarly	2	2	0	1	0	3	8
Energy, power, telecom, networks	2	2	2	0	0	2	8
Law, governance, public sector	0	1	0	0	1	6	8
Geoscience, environment, agriculture, food	1	0	1	2	1	2	7
Finance, business, insurance	0	1	0	0	0	3	4
Culture, heritage, society	1	0	1	0	0	1	3
Design and creative support	0	1	1	0	0	0	2
Total	41	27	34	18	12	59	191

4. Taxonomy Statistics

The analysis of the included papers indicates that the current RAG literature is not dominated by a single methodological direction, but several complementary structural design logics coexist in the RAG ecosystem. The distribution reported in Table 4 shows that workflow-oriented and task-integrated modifications account for the largest share of the corpus, while retrieval optimization, structured-memory design, orchestration, multimodal grounding and governed retrieval each account for substantial and distinct shares.

Table 4. Distribution of the RAG literature by methodological family.

RAG Family	Count	Percentage (%)
Task-Compositional RAG	59	30.9
Retrieval-Optimized RAG	41	21.5
Structured-Memory RAG	34	17.8
Orchestrated RAG	27	14.1
Multimodal-Grounded RAG	18	9.4
Governed RAG	12	6.3
Total	191	100.0

Within the included studies in this review, RAG methodology is treated less often as a self-contained retrieve-then-generate enhancement for question answering. Instead, the reviewed studies treat RAG as a modular systems component that can be optimized, orchestrated, structured, grounded across modalities, governed by security and trust constraints, or embedded into broader decision-support pipelines. RAG architecture should no longer be discussed as a monolithic architecture, but as a broader methodological ecosystem in which different families respond to different failure modes and deployment goals. The proposed taxonomy separates improvements in retrieval fidelity, retrieval policy, memory structure, multimodal grounding, governance constraints, and task integration into analytically distinct categories that can be compared more systematically.

The distribution of the studies as reported in Table 4 needs to be read with two interpretive cautions, with the first being that the counts show where authors located their methodological contribution in journal-published work. They may reflect the tractability of problems, the breadth of application settings or venue preferences as much as methodological merit, and they are not evidence of superior performance or of convergence in the field. Secondly, the category Task-Compositional RAG is, by design, the most comprehensive category in the proposed taxonomy, allowing systems that are evaluated by the quality of their downstream product, regardless of their retrieval mechanism, and, thus, includes seven composition patterns, which share a common role rather than a similar architecture. Its leading share should therefore be interpreted as showing that, within the corpus, RAG is frequently embedded as a component of applied workflows, rather than as evidence that workflow integration is the dominant methodological innovation in the field. Under a narrower definition restricted to systems whose principal novelty lies in the composition logic itself, this family would be substantially smaller. The 20 grounded-guidance studies (Section 7.6.6) are the part of the family closest to question answering; excluding them would reduce it to 39 studies (20.4%), slightly below Retrieval-Optimized RAG (41 studies, 21.5%). No conclusion of this review depends on Task-Compositional RAG being the largest family.

Table 5 reports the included studies by publication year. All families except Multimodal-Grounded and Governed RAG, which have no included study from 2024, are represented in every year. Retrieval-Optimized and Structured-Memory RAG are concentrated in 2025, while Orchestrated and Governed RAG account for a larger share of the 2026 studies than

of the 2025 studies. Because 2026 is covered only up to 27 April, its counts describe a partial year and are not interpreted as a trend.

Table 5. Included studies by publication year and methodological family.

Year	RO	OR	SM	MG	GV	TC	Total
2024	7	1	2	0	0	2	12
2025	23	13	24	11	5	38	114
2026 (to 27 April)	11	13	8	7	7	19	65
Total	41	27	34	18	12	59	191

5. Retrieval-Augmented Generation (RAG)

RAG is an advanced AI methodology that combines information retrieval (IR) with generative models to improve factual accuracy and contextual relevance. Unlike traditional IR systems, which cannot generate new text, or LLMs, which rely solely on static training data, RAG first retrieves relevant documents from an external index and then conditions the LLM's generation on them. The workflow comprises three stages: indexing (document storage), retrieval (fetching relevant passages per query), and generation (synthesizing an answer from query and retrieved text. Recent RAG systems preserve this core retrieve–ground–generate logic, but many of them redesign the retrieval layer through query control, structured memory, multimodal grounding, security constraints, and workflow-level integration [35,43,51,61,62].

The retrieval phase is foundational, as poor retrieval leads to poor generation. Modern IR has evolved from sparse methods (e.g., BM25, TF-IDF) to dense transformer-based embeddings (e.g., DPR [63]), with hybrid and knowledge-graph-enhanced approaches further improving performance, especially for complex tasks. Retrieval quality depends on chunking strategy, embedding models, and query optimization; advanced RAG pipelines also incorporate pre- and post-retrieval steps such as relevance filtering, reranking, and information compression. Integration of retrieved content into generation can be achieved via simple concatenation (which may exceed context limits), knowledge aggregation (e.g., filtering or compressing), Fusion-in-Decoder (FiD) [64], which processes documents in parallel, or adaptive fusion using attention mechanisms to dynamically weight retrieved evidence and reduce hallucination. Recent architectures combine vector search with graph traversal and similarity validation, so that retrieval quality depends not only on ranking accuracy but also on segmentation, alignment and filtering before the evidence reaches the generator [29,32,65]. The foundational methods referred to in this section, such as DPR, FiD, ColBERT, HyDE, FLARE, Self-RAG, Corrective RAG, RAPTOR and GraphRAG, are cited as background and are not part of the classified corpus.

5.1. Naive (Standard) RAG

Naive RAG refers to the simplest baseline instantiation of retrieval-augmented generation. It follows a basic three-stage pipeline consisting of indexing, retrieval, and generation, without introducing more sophisticated optimization mechanisms for complex queries. During indexing, documents are segmented into chunks and encoded into embeddings for storage in a searchable database. During retrieval, the user query is embedded in the same space and matched against the indexed chunks so that the most relevant evidence can be selected. These retrieved chunks are then appended to the query and passed to the language model for response generation [6]. If no sufficiently relevant chunk is found, the prompt may be passed to the LLM without augmentation. This vanilla retrieve-then-generate instan-

tiation, also referred to as standard RAG, couples the retriever and the generator without additional orchestration, structured-memory redesign or multimodal alignment.

5.2. Advanced RAG

Advanced RAG keeps the single retrieve-then-generate pass of Naive RAG but adds processing steps before and after retrieval. Before retrieval, the corpus is prepared with more care, for example, by segmenting documents along their logical structure and attaching metadata to each chunk, and the query can be rewritten or expanded so that it better matches the wording of the indexed text. A query-side variant generates a hypothetical answer document and uses it to guide dense retrieval, as in HyDE [66]. After retrieval, candidate chunks are filtered by a similarity threshold, reordered by a reranker, and compressed or truncated so that the most useful evidence fits the context window of the generator. These steps target the typical failures of Naive RAG, namely, missing evidence, irrelevant evidence and poorly integrated context [29,65], at the price of additional computation on every query. In the proposed taxonomy, studies whose principal contribution is such a refinement are assigned to Retrieval-Optimized RAG.

5.3. Modular RAG

The Modular RAG paradigm emerged as a later evolution in RAG methodologies, emphasizing flexibility, scalability, and domain adaptability. Modular RAG decomposes the traditional retrieve-then-generate process into discrete, reconfigurable modules. These modules independently handle different stages, including query reformulation, retriever selection, context aggregation, reranking, or routing. In this way, retrieval, reranking, memory access, routing, and answer synthesis can be specialized independently [11,19]. Adaptive and self-reflective variants decide when to retrieve and assess the retrieved evidence before or during generation, as in FLARE [67], Self-RAG [68] and Corrective RAG [69].

5.4. Sparse Retrieval RAG

Sparse retrieval represents queries and text units as high-dimensional vectors whose non-zero entries correspond to the terms they contain, and it matches them through exact term overlap. An inverted index stores, for every term, the text units in which it occurs and its frequency in each of them, together with the number of units in the collection that contain it, so that candidate units can be scored without scanning the whole corpus [70].

BM25, the scoring function of the probabilistic relevance framework, weights each query term by its inverse document frequency, lets the contribution of repeated occurrences saturate, and normalizes for the length of the text unit. The relevance score is [70] as follows:

$$\mathrm{Score}(q,d) = \sum_{t \in q} \frac{\mathrm{IDF}(t) \cdot \mathrm{TF}(t,d) \cdot (k_1 + 1)}{\mathrm{TF}(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)} \quad (1)$$

where q represents the query and d denotes the text unit; $\mathrm{TF}(t,d)$ is the frequency of keyword t within d; $\mathrm{IDF}(t)$ is the inverse document frequency of term t; $|d|$ and avgdl represent the document length and the average document length across the collection, respectively; and the parameters k_1 and b calibrate term-frequency saturation and document-length normalization.

Because it matches surface terms, sparse retrieval is efficient, interpretable and reliable for identifiers, codes, legal phrases and technical terminology, but it misses relevant text that is phrased differently from the query. For this reason, several of the reviewed systems combine it with dense retrieval in the hybrid designs described below [46,59,71].

5.5. Dense Retrieval RAG

Dense retrieval utilizes high-dimensional semantic embeddings maintained within a dense storage module to evaluate conceptual alignment between user queries and text units. Unlike traditional methods that depend on lexical overlap, this approach measures similarity in a continuous vector space.

The process involves transforming the user query into a semantic embedding, denoted as v_q, using a pretrained encoder. Concurrently, each text unit within the storage module is represented as a dense vector v_i. The degree of semantic correspondence between the query and a document is determined via cosine similarity, calculated as follows:

$$\text{Similarity}(v_q, v_i) = \frac{v_q \cdot v_i}{\|v_q\|\|v_i\|} \tag{2}$$

Based on these similarity scores, the system ranks the text units and retrieves the top-k candidates. The primary advantage of dense retrieval over sparse methods lies in its ability to capture latent semantic relationships, enabling retrieval of relevant documents even under vocabulary mismatch. Representative dense retrievers include bi-encoders such as DPR [63] and late-interaction models such as ColBERT [72].

Dense retrieval is particularly advantageous when the query and the relevant evidence are semantically related but lexically different.

5.6. Hybrid Retrieval RAG

To enhance retrieval robustness, a hybrid framework integrates the outputs of both sparse and dense retrievers into a unified candidate set. This architecture strategically combines lexical precision with semantic depth. By merging these streams, the system ensures that it captures both exact keyword matches and contextually relevant information that might lack direct vocabulary overlap. The combined context pool is processed through a scoring mechanism, such as a reranker or weighted relevance average, to select the top-k most pertinent segments for the generation phase:

$$C_q = \text{TopK}(C_{\text{sparse}} \cup C_{\text{dense}}, k) \tag{3}$$

where C_{sparse} and C_{dense} represent the document sets retrieved by the sparse and dense modules, respectively, and $\text{TopK}(\cdot, k)$ denotes the ranking function that returns the k highest-scoring segments for downstream answer generation. Hybrid retrieval combines the lexical precision of sparse search with the semantic generalization of dense retrieval. This is especially effective when exact terminology matters but does not by itself capture the complete information need [27,73,74].

5.7. Graph-Based RAG

Graph-based RAG has emerged as a major paradigm in recent years. Representative systems perform community detection, subgraph retrieval, or graph traversal over an external graph structure to identify query-relevant local regions and then generate answers by integrating both local and global information. Such systems are better suited than traditional flat-text RAG to exploit entity connectivity, relation structure, and multihop dependencies.

Graph-based RAG differs from flat-text retrieval by making topology, entity connectivity, and multihop dependency paths explicit during evidence selection. Rather than retrieving only semantically similar passages, graph-based systems can localize relevant communities, traverse relation paths, and aggregate evidence across linked graph regions [43,45,61]. Representative examples are GraphRAG, which builds an entity graph with community summaries for query-focused summarization [75], and RAPTOR, which organizes a corpus into a recursive tree of abstractive summaries [76].

Knowledge graphs (KGs) are structured databases that model real-world entities and relationships as graphs, making them highly amenable to machine processing. They enable efficient querying to retrieve all entities related to a given entity, a task that would be significantly more challenging with unstructured text databases. KG-RAG can be defined as an RAG variant in which the retrieval layer operates over a knowledge graph or a graph-enriched retrieval substrate instead of relying only on flat chunk similarity. KG-RAG can be understood as a more explicitly semantic form of graph-based retrieval because its memory is organized around typed entities, typed relations, and often domain constraints or ontological structure. This allows the retrieval stage to preserve provenance, relation semantics, and multihop reasoning paths that would otherwise remain hidden in flat text corpora [43,45,61].

5.8. Multimodal RAG

Multimodal RAG integrates more than one evidence modality, such as text, image, video or domain-specific question answering and reasoning. The model accesses an external non-parametric multimodal memory, allowing it to retrieve relevant evidence across modalities. Pretraining or downstream adaptation is often performed using image-text and text-only corpora with joint contrastive and generative objectives. Multimodal RAG extends retrieval augmentation beyond text by enabling evidence cross-modal across text, images, video, and other non-text [47,48,50,62].

6. Evaluation Metrics for Retrieval-Augmented Generation Systems

A comprehensive RAG evaluation should jointly assess the retrieval module, the generation module, and the end-to-end pipeline. Evaluation should therefore not focus only on whether the retriever returns relevant evidence, but also on whether the generated answer is grounded, semantically aligned with references, operationally efficient, and useful for downstream decision-support settings [30,54,59,61,65,77–79]. Several of the RAG-specific measures defined below, namely, context precision, context recall, faithfulness and answer relevance, follow the reference-free evaluation framework introduced by RAGAS [80], which is cited here as background.

The RAG-specific evaluation dimensions used in the reviewed studies include Context Precision, Context Recall, Context Relevance, Faithfulness, Answer Relevance, Answer Correctness, Answer Similarity, Hallucination Rate, Token Consumption, Generation Time, Latency, and Cost [30,35,59,61,79,81–83]. In addition, studies also report lexical and semantic reference-based metrics such as BLEU, ROUGE, BERTScore, SBERT cosine similarity, and MPNet similarity, as well as downstream task metrics such as accuracy, precision, recall, sensitivity, specificity, F1-score, Matthews correlation coefficient (MCC), AUROC/AUC, MAE, RMSE, R^2, exact match (EM), and Top-k accuracy [62,77–79,83–87]. How often each metric group is actually reported is quantified in Section 6.8.

The underlying evaluation methods of this review can be organized into six complementary groups and can be grouped into retriever-level and generator-level metrics, reference-based lexical and semantic metrics, downstream task performance, hallucination, robustness and efficiency parameters, and human-centered qualitative metrics.

6.1. Retriever-Level Metrics

The first analytical layer in evaluating retrieval-augmented generation (RAG) systems concerns the retrieval module itself, meaning its capacity to locate evidence that is both relevant to the query and suitably ranked for the generator's conditioning. In the reviewed literature, retriever performance is typically assessed through context-oriented measures such as Context Precision and Context Recall, along with rank-sensitive information re-

trieval metrics including Precision@k, Recall@k, F1@k, Hit@k, mean reciprocal rank (MRR), mean average precision (MAP), and normalized discounted cumulative gain (nDCG) [30,59,61,65,81,82]. Together, these measures describe complementary aspects of retrieval quality, covering relevance concentration, evidence coverage, and ranking effectiveness.

Let $\mathcal{Q} = \{q_1, \ldots, q_{|\mathcal{Q}|}\}$ represent the evaluation query set. For a given query q_j, let $\mathcal{R}_j$ denote the set of retrieved evidence units, and $\mathcal{G}_j$ the corresponding set of ground-truth relevant evidence units.

In its set-based form, Context Precision measures the fraction of retrieved evidence that is relevant; RAGAS defines a rank-weighted variant that averages precision at the positions of the relevant items [80]:

$$\mathrm{CP} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\mathcal{R}_j \cap \mathcal{G}_j|}{|\mathcal{R}_j|}. \tag{4}$$

Context Recall, by contrast, evaluates the proportion of relevant evidence successfully recovered:

$$\mathrm{CR} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\mathcal{R}_j \cap \mathcal{G}_j|}{|\mathcal{G}_j|}. \tag{5}$$

These two measures are especially significant in RAG settings because they jointly indicate whether the retriever supplies enough evidentiary support while limiting irrelevant or distracting context.

Context Relevance quantifies the semantic alignment between each query and its retrieved evidence:

$$\mathrm{CtxRel} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{1}{|\mathcal{R}_j|} \sum_{r \in \mathcal{R}_j} s(q_j, r), \tag{6}$$

where $s(\cdot, \cdot)$ denotes a semantic similarity function, often implemented as cosine similarity in an embedding space.

Averaged embedding similarity is the most frequently encountered in the reviewed corpus. Relevance can also be measured either as LLM-as-judge relevance grading of each retrieved unit, binary or graded human annotation of relevance, and claim-level entailment between the query and the retrieved evidence.

To assess the sensitivity of retrieval performance to rank position, the literature often adopts Precision@k, Recall@k, and F1@k. Let $\mathrm{Top}_k(\mathcal{R}_j)$ denote the set of top-k retrieved items for query q_j. Then,

$$\mathrm{Prec@}k = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\,\mathrm{Top}_k(\mathcal{R}_j) \cap \mathcal{G}_j|}{k}, \tag{7}$$

$$\mathrm{Rec@}k = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\,\mathrm{Top}_k(\mathcal{R}_j) \cap \mathcal{G}_j|}{|\mathcal{G}_j|}, \tag{8}$$

$$\mathrm{F1@}k = \frac{2 \cdot \mathrm{Prec@}k \cdot \mathrm{Rec@}k}{\mathrm{Prec@}k + \mathrm{Rec@}k}. \tag{9}$$

Hit@k evaluates whether at least one relevant item appears within the top-k retrieved results [81,82]:

$$\mathrm{Hit@}k = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \mathbb{I}(h_j \le k), \tag{10}$$

where h_j denotes the rank position of the first relevant item for query q_j, and $\mathbb{I}(\cdot)$ is the indicator function. If no relevant item is retrieved for q_j, we set $h_j = \infty$, so that the query contributes zero both to Hit@k and to the MRR defined below.

Mean reciprocal rank (MRR) highlights the prompt retrieval of the first relevant result and penalizes systems where relevant evidence appears only at deeper ranks [79]:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{h_j}. \tag{11}$$

When relevance is graded rather than binary, normalized discounted cumulative gain (nDCG) becomes particularly useful. For query q_j, the discounted cumulative gain at cutoff k is defined as

$$\mathrm{DCG}_k^{(j)} = \sum_{i=1}^{k}\frac{2^{g_{j,i}}-1}{\log_2(i+1)}, \tag{12}$$

where $g_{j,i}$ is the graded relevance assigned to the item at rank i. The normalized form is given by

$$\mathrm{nDCG}_k^{(j)} = \frac{\mathrm{DCG}_k^{(j)}}{\mathrm{IDCG}_k^{(j)}}, \tag{13}$$

where $\mathrm{IDCG}_k^{(j)}$ is the ideal discounted cumulative gain for query q_j. The average over the entire query set is

$$\mathrm{mNDCG}_k = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\mathrm{nDCG}_k^{(j)}. \tag{14}$$

Mean average precision (MAP) offers a rank-aware summary of retrieval quality by averaging precision values at all ranks where relevant items occur [30,81,88]. For query q_j, average precision is defined as

$$\mathrm{AP}(q_j) = \frac{1}{|\mathcal{G}_j|}\sum_{i=1}^{n_j}\mathrm{Prec}^{(j)}@i\cdot\delta_{j,i}, \tag{15}$$

where n_j is the number of retrieved items, $\mathrm{Prec}^{(j)}@i$ is the precision at rank i for query q_j, and $\delta_{j,i} \in \{0,1\}$ indicates whether the item at rank i is relevant. The corpus-level MAP is then

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\mathrm{AP}(q_j). \tag{16}$$

For applications in which the highest-ranked retrieved item is decisive, depth-specific variants are also reported [30,81,88]. Two distinct quantities are frequently conflated in the literature, as Hit@1 records whether the top-ranked item is relevant, while Recall@1, that is Rec@k with $k = 1$, records the proportion of the relevant evidence set recovered at rank one.

The two coincide only when there is exactly one relevant item per query, that is, when $|\mathcal{G}_j| = 1$. When $|\mathcal{G}_j| > 1$, Recall@1 is bounded above by $1/|\mathcal{G}_j|$ and cannot reach unity, so Hit@1 is the more informative measure at this depth. Values reported in the corpus under the name "Recall@1" should therefore be interpreted with reference to the cardinality of the underlying relevance set.

The nDCG@1 values frequently reported in the corpus correspond to the special case $k = 1$ of the mean nDCG defined above.

Finally, Coverage, the proportion of queries for which at least one relevant evidence unit is returned, equals Hit@k evaluated at the full depth of the retrieved list, $k = |\mathcal{R}_j|$.

6.2. Generator-Level Metrics

The second evaluative layer concerns the generation module, that is, whether the synthesized response is grounded in the retrieved evidence and responsive to the user query. The generator-level metrics used in the reviewed studies include Faithfulness, Answer Relevance, Answer Correctness, Answer Similarity, and, in some studies, Answer Accuracy [59,61,65,78].

Faithfulness measures the degree to which the generated response is supported by the retrieved context and is central to hallucination analysis [30]. Let $\mathcal{A}_j$ denote the set of atomic claims extracted from the generated response for query q_j, and let $\mathcal{S}_j \subseteq \mathcal{A}_j$ denote the subset of claims supported by the retrieved evidence. Then,

$$\text{Faithfulness} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\mathcal{S}_j|}{|\mathcal{A}_j|}. \tag{17}$$

Answer Relevance evaluates whether the generated response remains semantically aligned with the original question. A common approach is to generate proxy questions from the answer and compare them with the input query [59]:

$$\text{AnsRel} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{1}{M_j} \sum_{m=1}^{M_j} s(q_j, \tilde{q}_{j,m}), \tag{18}$$

where $\tilde{q}_{j,m}$ denotes the m-th proxy question derived from the generated response, M_j is the number of such proxy questions, and $s(\cdot,\cdot)$ is a semantic similarity function.

In multimodal RAG settings, related measures such as Multimodal Faithfulness and Multimodal Relevance are also used to verify that the generated response is jointly supported by textual and visual evidence [62].

Answer Accuracy is often obtained through human annotation or model-based judging schemes and is typically normalized to the interval $[0, 1]$ [32,62]:

$$\text{AnsAcc} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \alpha_j, \qquad \alpha_j \in [0, 1], \tag{19}$$

where α_j is the adjudicated accuracy score assigned to the response for query q_j.

Answer Similarity measures the semantic agreement between the generated response $\hat{y}_j$ and the reference answer $y_j^\star$:

$$\text{AnsSim} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \cos\Big(\phi(\hat{y}_j), \phi(y_j^\star)\Big), \tag{20}$$

where $\phi(\cdot)$ is an embedding function and $\cos(\cdot,\cdot)$ denotes cosine similarity.

Answer Correctness is often modeled as a weighted combination of factual validity and semantic similarity. Let $\widetilde{\mathcal{T}}_j$, $\widetilde{\mathcal{F}}_j^+$, and $\widetilde{\mathcal{F}}_j^-$ denote the sets of true-positive, false-positive, and false-negative atomic statements, respectively, for query q_j. A statement-level F1 score of factual validity, distinct from the FActScore metric, may be written as

$$\text{StmtF1} = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{|\widetilde{\mathcal{T}}_j|}{|\widetilde{\mathcal{T}}_j| + \frac{1}{2}\Big(|\widetilde{\mathcal{F}}_j^+| + |\widetilde{\mathcal{F}}_j^-|\Big)}. \tag{21}$$

The composite correctness score is then

$$\text{AnsCorr} = \lambda_1 \cdot \text{StmtF1} + \lambda_2 \cdot \text{AnsSim}, \qquad \lambda_1 + \lambda_2 = 1, \tag{22}$$

where λ_1 and λ_2 control the relative contribution of factual support and semantic agreement [30,78].

6.3. Reference-Based Lexical and Semantic Metrics

In many practical RAG applications, generated responses are also evaluated using reference-dependent lexical and semantic overlap metrics [77–79,82]. These measures remain useful whenever gold-standard answers exist and when one seeks to quantify textual overlap or higher-level semantic agreement.

BLEU evaluates n-gram precision while applying a brevity penalty to discourage overly short outputs:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} \alpha_n \log p_n\right), \tag{23}$$

where BP is the brevity penalty, α_n are normalized weights satisfying $\sum_{n=1}^{N} \alpha_n = 1$, and p_n is the modified n-gram precision.

ROUGE takes a recall-oriented perspective. ROUGE-N is defined as

$$\text{ROUGE-}N = \frac{\sum_{g \in \mathcal{G}_N(y^\star)} \min\left\{\text{cnt}_{\hat{y}}(g), \text{cnt}_{y^\star}(g)\right\}}{\sum_{g \in \mathcal{G}_N(y^\star)} \text{cnt}_{y^\star}(g)}, \tag{24}$$

where $\mathcal{G}_N(y^\star)$ denotes the set of reference N-grams, and $\text{cnt}_{\hat{y}}(g)$ and $\text{cnt}_{y^\star}(g)$ denote the frequencies of g in the generated and reference texts, respectively.

ROUGE-L is based on the longest common subsequence (LCS) and captures sentence-level structural similarity:

$$\text{ROUGE-}L = \frac{(1+\beta^2) P_{\text{LCS}} R_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}, \tag{25}$$

where

$$P_{\text{LCS}} = \frac{\text{LCS}(\hat{y}, y^\star)}{|\hat{y}|}, \qquad R_{\text{LCS}} = \frac{\text{LCS}(\hat{y}, y^\star)}{|y^\star|}. \tag{26}$$

BERTScore F1 evaluates semantic overlap by aligning contextual token embeddings from the generated and reference texts [79]:

$$\text{BERTScore}_{F1} = \frac{2 P_{\text{B}} R_{\text{B}}}{P_{\text{B}} + R_{\text{B}}}, \tag{27}$$

where P_{B} and R_{B} are the embedding-based precision and recall components.

Sentence-level semantic agreement is also often assessed using SBERT or MPNet embeddings:

$$\text{Sim}_{\text{SBERT/MPNet}} = \cos(\psi(\hat{y}), \psi(y^\star)), \tag{28}$$

where $\psi(\cdot)$ maps each sentence into a dense semantic representation [79].

6.4. Downstream Task Metrics

RAG systems are not limited to open-ended question answering; many are integrated into broader classification, structured prediction, and decision-support pipelines [83,85–87]. In such contexts, intrinsic generation metrics are insufficient, and downstream task performance becomes essential for assessing practical utility.

For classification-oriented applications, the metrics used include accuracy, precision, recall, sensitivity, specificity, F1-score, and Matthews correlation coefficient (MCC) [85–87]. Let N_{TP}, N_{TN}, N_{FP}, and N_{FN} denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Then,

$$\text{Accuracy} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}}, \tag{29}$$

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}}, \tag{30}$$

$$\text{Recall} = \text{Sensitivity} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}, \tag{31}$$

$$\text{Specificity} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}}, \tag{32}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{33}$$

$$\mathrm{MCC} = \frac{N_{TP}N_{TN} - N_{FP}N_{FN}}{\sqrt{D}} \tag{34}$$

where

$$D = (N_{TP} + N_{FP})(N_{TP} + N_{FN}) \times (N_{TN} + N_{FP})(N_{TN} + N_{FN}).$$

For systems that produce structured outputs, exact match (EM) and Top-k accuracy are frequently reported. Exact match is defined as

$$\mathrm{EM} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(\hat{z}_j = z_j^\star), \tag{35}$$

where $\hat{z}_j$ and $z_j^\star$ denote the predicted and gold-standard outputs for sample j.

Top-k accuracy measures whether the correct label is contained within the top-k predicted candidates:

$$\text{Top-}k\ \text{Accuracy} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(z_j^\star \in \mathcal{Z}_j^{(k)}), \tag{36}$$

where $\mathcal{Z}_j^{(k)}$ is the set of the k highest-ranked predictions for instance j.

6.5. Hallucination, Robustness, and Efficiency Metrics

Another evaluation layer concerns operational robustness, computational efficiency, and deployment feasibility. In the reviewed literature, the metrics in this category include hallucination rate, latency, token consumption, generation time, peak memory usage, and cost [30,35,54,61,79,83]. These measures matter because high answer quality alone does not ensure that an RAG system is practical or scalable in real-world settings.

Hallucination rate measures the proportion of generated claims not supported by the retrieved evidence:

$$\text{Hallucination Rate} = \frac{\sum_{j=1}^{|Q|} |\mathcal{U}_j|}{\sum_{j=1}^{|Q|} |\mathcal{A}_j|} \times 100\%, \tag{37}$$

where $\mathcal{U}_j$ denotes the set of unsupported claims in the response to query q_j [61,79].

Latency records the end-to-end processing time of the system:

$$\text{Latency} = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\left(\tau_{\text{ret}}^{(j)} + \tau_{\text{rank}}^{(j)} + \tau_{\text{gen}}^{(j)}\right), \tag{38}$$

where $\tau_{\text{ret}}^{(j)}$, $\tau_{\text{rank}}^{(j)}$, and $\tau_{\text{gen}}^{(j)}$ denote retrieval, reranking, and generation time, respectively [35,54].

Token consumption is commonly defined as

$$\text{TokUse} = N_{\text{tok}}^{\text{in}} + N_{\text{tok}}^{\text{out}}, \tag{39}$$

where $N_{\text{tok}}^{\text{in}}$ and $N_{\text{tok}}^{\text{out}}$ are the input and output token counts [61,83].

Generation time isolates the decoding cost of the generator:

$$\text{GenTime} = \frac{1}{|Q|}\sum_{j=1}^{|Q|} \tau_{\text{gen}}^{(j)}. \tag{40}$$

Peak memory usage measures the maximum memory footprint during inference:

$$\text{Mem}_{\text{peak}} = \max_t U(t), \tag{41}$$

where $U(t)$ denotes memory consumption at time t [83].

When monetary considerations are explicitly modeled, computational cost can be approximated as

$$\text{Cost} = c_{\text{in}} N_{\text{tok}}^{\text{in}} + c_{\text{out}} N_{\text{tok}}^{\text{out}} + c_{\text{ret}} N_{\text{ret}}, \tag{42}$$

where c_{in} and c_{out} are token pricing coefficients for input and output, c_{ret} is the cost of a retrieval operation, and N_{ret} is the number of retrieval calls.

6.6. Human Evaluation Metrics

Human-centered evaluation remains essential in high-stakes applications, particularly in medical, legal, and multimodal RAG systems [54,62,84]. In such cases, qualitative judgments are usually summarized through mean Likert score, percent agreement, and intraclass correlation coefficient (ICC).

The mean Likert score is the average of ordinal ratings assigned by expert annotators:

$$\overline{L} = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \ell_i, \tag{43}$$

where ℓ_i is the score assigned to the i-th evaluated item.

Percent agreement measures the proportion of observations where independent evaluators reach the same judgment:

$$\text{PA} = \frac{N_{\text{agr}}}{N_{\text{obs}}} \times 100\%, \tag{44}$$

where N_{agr} is the number of agreement cases and N_{obs} is the total number of observations.

The intraclass correlation coefficient (ICC) provides a variance-based measure of inter-rater reliability:

$$\text{ICC} = \frac{\sigma_{\text{sub}}^2}{\sigma_{\text{sub}}^2 + \sigma_{\text{rat}}^2 + \sigma_{\text{err}}^2}, \tag{45}$$

where σ_{sub}^2, σ_{rat}^2, and σ_{err}^2 denote the subject, rater, and residual variance components, respectively.

The formulation of ICC corresponds to a two-way random-effects, single-measurement, absolute-agreement intraclass correlation coefficient. This formulation is appropriate when a single, constant panel of judges scores every response, and when the absolute scores are compared, which is the mode most frequently used in the literature. A one-way model is appropriate when judges differ across items, and a consistency model is suitable when only the ranking order is important [89].

6.7. An Analytic View of RAG Evaluation

The preceding analysis shows that no single metric suffices for a rigorous evaluation of an RAG system. Retriever-level metrics alone cannot reveal whether the generation module used the retrieved evidence correctly, while generator-level metrics alone cannot determine whether failure stems from inadequate retrieval, weak grounding, or poor response synthesis. Similarly, lexical overlap measures do not ensure factual correctness, and downstream task indicators often conflate the contributions of retrieval and generation subsystems.

Therefore, a scientifically robust RAG evaluation framework should combine six complementary groups of metrics:

1. Retriever-level metrics, assessing evidence identification and ranking quality;
2. Generator-level metrics, evaluating grounding, answer relevance, and hallucination behavior;
3. Reference-based lexical and semantic metrics, quantifying textual overlap and semantic agreement;

4. Downstream task metrics, measuring functional effectiveness in application-specific contexts;
5. Hallucination, robustness, and efficiency metrics, capturing grounding-failure rates together with latency, token consumption, memory footprint and cost, that is, the operational feasibility of deployment;
6. Human-centered metrics, assessing expert-judged answer quality and the reliability of that judgment through inter-rater agreement.

When considered together, these six evaluation dimensions provide a more comprehensive and analytically meaningful characterization of RAG system performance, spanning retrieval fidelity, generation faithfulness, semantic correctness, downstream utility, operational readiness, and expert trust.

6.8. Metric Usage in the Reviewed Corpus

To characterize evaluation practice, every included study was coded for the metric groups of Section 6.7 and for the use of statistical significance testing. The coding drew on the evaluation and results sections and the tables of each study, and a metric group was recorded as used only when the study reports a numerical value for at least one metric of that group; metrics that are only mentioned, for example, in the related work or in the description of other systems, were not counted. Candidate passages were located with a scripted keyword search of each full text, and each code was assigned after the passage had been read in context. The coding was carried out with the assistance of an AI tool (Claude, Anthropic) and was checked by the authors. For 20 studies whose full text could not be obtained, the codes rest on the published abstract; and 6 studies were coded from their preprint versions—these studies are marked in Supplementary Table S1, which reports the codes of every study. Table 6 reports the number and share of studies that use each group, overall and by family. Restricting the analysis to the 165 studies coded from the published full text changes no overall share by more than 2.2 percentage points.

Table 6. Metric groups with reported values in the 191 included studies, overall (number and %) and by methodological family (% of the studies in each family).

Metric group	All, n (%) (191)	RO (41)	OR (27)	SM (34)	MG (18)	GV (12)	TC (59)
Retriever-level metrics	52 (27.2)	31.7	25.9	35.3	27.8	16.7	22.0
Generator-level metrics	41 (21.5)	26.8	22.2	29.4	16.7	8.3	16.9
Reference-based lexical and semantic metrics	38 (19.9)	19.5	25.9	8.8	27.8	33.3	18.6
Downstream task metrics	134 (70.2)	58.5	77.8	82.4	66.7	66.7	69.5
Hallucination, robustness and efficiency metrics	75 (39.3)	29.3	63.0	26.5	50.0	58.3	35.6
Human evaluation	49 (25.7)	24.4	14.8	26.5	27.8	8.3	33.9
Statistical significance testing	51 (26.7)	36.6	22.2	29.4	11.1	8.3	28.8

Evaluation gaps. Downstream task metrics are the most widely reported group, used in 134 of the 191 studies (70.2%), followed by hallucination, robustness and efficiency metrics (75, 39.3%). Retrieval quality and grounding are measured far less often: retriever-level metrics are reported in 52 studies (27.2%), generator-level metrics such as faithfulness and answer relevance in 41 (21.5%), and both together in only 26 (13.6%). In 69 studies (36.1%) a downstream task metric is the only quality measure, with no retriever-level, generator-level, reference-based or human evaluation, so that the contribution of retrieval to the reported result is not isolated. Human evaluation (49, 25.7%) and statistical significance testing (51, 26.7%) are each reported in about one study in four. The families differ in emphasis: retriever-level metrics are most frequent in Structured-Memory RAG (35.3%),

hallucination, robustness and efficiency metrics in Orchestrated RAG (63.0%), and human evaluation in Task-Compositional RAG (33.9%), while only 1 of the 12 Governed RAG studies reports a human evaluation, and only 1 a significance test.

7. Discussion

This section reviews the literature family by family, focusing on the main methodological contribution, the shared design logic, and the key subcategories found in the corpus. The reviewed studies are not evenly distributed across the proposed taxonomy; rather, they show distinct patterns of architectural emphasis, ranging from retrieval-quality improvement and orchestration policies to structured memory, multimodal grounding, governance constraints, and workflow-level integration. Tables 7 and 8 summarize these patterns across the six families, Table 9 lists the studies in each family and subcategory, and the following subsections consider each family in turn. Read comparatively, the families of the taxonomy trade quality for cost at different phases of the architecture and the process. Retrieval-Optimized RAG adds a one-off cost when documents are segmented and indexed or embeddings are adapted and, when it introduces reranking, hybrid fusion or relevance filtering, a further cost on every query. Structured-Memory RAG pays once to build and maintain its store and again on every query for graph traversal or subgraph retrieval, while Orchestrated RAG pays for extra retrieval rounds and tokens on each request. The Governed RAG family is the only one where the intervention deliberately reduces the evidence available to the generator; thus, its evaluation must always include both utility and protection. In Task-Compositional RAG, retrieval is coupled with the other components of a wider workflow, so that its contribution to the reported outcomes is harder to isolate.

Table 7. Comparative synthesis of the six methodological families.

Family	Problem Addressed	Architectural Intervention	Intended Benefit (Reported In)
Retrieval-Optimized	Evidence missed, mis-segmented, or ranked below the cut-off	Structure-aware chunking; domain-adapted embeddings; sparse–dense fusion; reranking; relevance filtering	Better evidence for a fixed pipeline [30,90]
Orchestrated	One retrieval pass fails multi-intent, underspecified, or multihop queries	Query rewriting, decomposition, routing; iterative retrieval; self-correction; agentic planning	Answers no single pass can satisfy [35,38]
Structured-Memory	Relational and multihop structure lost in flat chunks	Knowledge graphs and traversal; ontology-grounded stores; vector–graph and dual-index designs	Relation-preserving multihop grounding [42,43]
Multimodal-Grounded	Decisive evidence is non-textual	Vision–language retrieval; multimodal knowledge graphs; cross-modal alignment; dual-channel retrieval	Grounding in visual and spatial evidence [91,92]
Governed	Evidence is private, access-restricted, untrusted, or poisoned	Encrypted and perturbed retrieval; federated and access-controlled designs; provenance filtering; unlearning	Privacy, access control, poisoning resistance [93,94]
Task-Compositional	Output is an operational artifact, not a free-form answer	RAG as a grounding stage in clinical, industrial, educational, or public-sector pipelines	Artifacts usable in downstream workflows [95,96]

Table 8. Comparative synthesis of the six methodological families (continued).

Family	Cost Introduced by Design	Evaluation Focus Implied	Open Question
Retrieval-Optimized	Per query: reranking, fusion and filtering add latency; once: adaptation needs domain data	Rank-sensitive retrieval quality	Whether chunking and embedding choices transfer across corpora
Orchestrated	Each round adds tokens and latency; failure attribution harder	Multihop and multi-intent performance, with cost per query	When further retrieval is needed, and when to stop
Structured-Memory	Store must be built and maintained; graph traversal adds latency per query; construction errors propagate	Multihop accuracy and provenance of retrieved relations	Keeping the store consistent as the corpus changes
Multimodal-Grounded	Non-text encoders add computation; alignment adds a failure point	Support from the non-text evidence itself	How to measure non-text grounding directly
Governed	Protection restricts or perturbs generator-visible evidence	Protection under a stated threat model, with answer utility	Utility lost relative to an unprotected baseline
Task-Compositional	Retrieval coupled with other components; contribution not isolated	Task outcome plus retrieved-evidence quality	How much performance is attributable to retrieval

7.1. Retrieval-Optimized RAG

The Retrieval-Optimized RAG family includes 41 papers and covers methods that maintain the core RAG architecture, while focusing on the enhancement of retrieval quality. In this family, methods assume that hallucination and drift are not mainly derived during the generation process, but are generated due to a significant origin upstream that occurs from segmentation, ranking noise, or poor document-query relevance [30,32,77,97,98].

7.1.1. Chunking, Parsing, and Structure-Aware Segmentation

This subgroup focuses on how documents are segmented before indexing. These studies show that retrieval quality is very sensitive to chunk size, overlap, recursive segmentation, table-aware parsing, and document layout preservation. Specifically, ref. [97] examines OCR-sensitive RAG and improves retrieval through structure-aware OCR transformation, while [32] suggests table-aware chunking for complex tabular documents. Ref. [25] compares various chunking strategies for curriculum-grounded clinical QA, and [99] looks at chunking choices in service-discovery retrieval. This subgroup demonstrates that segmentation is not just a minor preprocessing step but an important retrieval factor [25,32,97,99].

7.1.2. Embedding Adaptation and Retrieval Alignment

This subgroup aims to improve retrieval by making the embedding space better capture the relationship between user queries and stored evidence. The core idea is that retrieval performs better when queries and relevant documents are represented in a more compatible manner. The proposed methods involve multilingual and domain-specific encoders, hypothetical-query matching, retrieval strategies designed for medical contexts, and mechanisms for checking semantic similarity.

More specifically, ref. [86] adapts retrieval to classify metal-binding residues in proteins, while [87] combines search augmentation with a multiscale CNN to predict protein-DNA binding sites. Ref. [77] concentrates on multilingual medication counseling, applying retrieval alignment to boost performance across languages. In molecular design, ref. [100] aligns retrieved molecular examples with structure-based drug discovery tasks. Ref. [26]

tackles the gap between questions and answers through hypothetical query embeddings, and [29] mitigates context mismatch by verifying semantic similarity. Likewise, ref. [31] adapts RAG for technical documentation domains, ref. [65] refines query–evidence matching with inverted question matching, and [101] improves protein-site prediction using retrieval-aligned representations. Altogether, these methods aim to strengthen the system's capacity to retrieve semantically relevant evidence even when the wording of the query and the source text differ notably.

7.1.3. Sparse, Dense, and Hybrid Retrieval

Studies in this subgroup report that hybrid retrieval is well suited to specialized domains because it preserves exact terminology while still allowing semantic generalizations. In particular, ref. [90] applies improved retrieval to textbook question answering, as [71] applies enhanced retrieval augmentation to biologically inspired UAV swarm control. The authors of [27] optimize RAG for telecom queries through hybrid retrieval and neural routing, while [98] combines semantic chunking, hybrid search, reranking, and long-context support for telecom QA. Finally, Chen et al. [102] employ hybrid retrieval to recover complementary evidence in dense technical knowledge spaces. These systems can improve recall while keeping the precision benefits of exact matching [27,71,90,98,102,103]. Hybrid retrieval is further paired with feature summarization in [103] for predicting pulmonary disease from radiology reports.

7.1.4. Reranking and Retrieval Refinement

This subgroup strengthens the evidence set following the initial retrieval stage. The main mechanisms are reranking, relevance filtering, thresholding, and distraction mitigation. In particular, ref. [104] reduces distracting context in industrial QA. Kelly et al. [105] improve retrieval for cultural heritage materials, while the authors in [28] integrate a dual retrieval and ranking in a medical context. Instead of altering memory structure or orchestration logic, these RAGs refine which retrieved candidates proceed to the generator [28,104,105].

7.1.5. Retrieval Evaluation, Benchmarking, and Comparative Studies

This subgroup explores the transformation of retrieval design into a clearly measurable experimental space. The main contributions compare retrievers, rerankers, similarity metrics, chunking strategies, and indexing options, or they introduce retrieval-oriented benchmarks and evaluation frameworks. In particular, ref. [106] benchmarks retrieval-augmented pipelines for screening of the literature, while [30] directly evaluates retrievers, rerankers, and similarity choices. The authors in [107] compare retrieval-augmented and standard LLMs on radiology board-examination questions, while in [108], medical LLMs with and without retrieval augmentation are compared. Furthermore, in [109], the authors examine the effect of RAG on domain-specific educational responses, while [110] assesses retrieval-enhanced nutrition guidance. The authors in [111] compare RAG variants for accelerating research on plastic degradation by mealworm larvae, as [112] provides a systematic analysis of RAG-based medical chatbots. Additionally, ref. [113] compares retrieval augmentation and fine-tuning in knowledge-based systems, with [114] measuring the effect of retrieval on microbiology-specific QA. The clinical accuracy of RAG-based LLMs for infective endocarditis prophylaxis is evaluated in [115], and retrieval augmentation is compared with fine-tuning for construction-safety question answering in [116]. Ref. [88] introduces a QA benchmark for retrieval evaluation; ref. [117] examines RAG integration in breast-cancer nursing support; ref. [118] presents a scalable RAG evaluation framework; ref. [85] compares RAG-enhanced tabular models; and [119] introduces a multiturn conversational RAG benchmark. Two further contributions target the evaluation instruments

themselves, as a unified needle-in-a-haystack framework is proposed in [120] for assessing long-context retrieval behavior under controlled evidence placement. Also, the authors in [121] examine the fundamental design decisions underlying RAG pipelines, reporting how configuration choices interact rather than ranking systems outright. These studies matter because they make retrieval design comparable across systems instead of leaving it as an implementation detail [30,85,88,106–121].

7.2. Orchestrated RAG

The Orchestrated RAG family includes 27 papers and covers methods that process retrieval as an adaptive control mechanism instead of a single static lookup. Within this family, systems determine *when* to retrieve, *how* to retrieve, *where* to retrieve from, and *whether* the retrieved evidence should be revised, decomposed, or validated. The core idea is that many complex queries call for staged evidence acquisition rather than one-shot retrieval [33,35,38,40,78].

7.2.1. Query Planning, Decomposition, and Routing

This subgroup covers methods that rewrite or break down queries, direct them to various repositories, or determine which retrieval path should be triggered first. Specifically, ref. [33] improves research assistance through query rewriting, while the authors in [122] arrange retrieval in two stages for design-guideline generation. Cheng et al. in [123] merge hierarchical retrieval with prompt generation for power-system evaluation, while DAG-style task decomposition in medical RAG is applied in [124]. The authors in [125] organize knowledge translation into stages for biomimicry design, while hierarchical multimodal retrieval for robot task planning is employed in [126]. Ref. [127] treats retrieval as part of interactive AI for next-generation networking, and the authors in [39] manage routing across multiple retrieval modes in a biomedicine. Multihop questions are addressed in [34], where complex queries are broken up into separate, retrievable questions, whose answers can be recombined, enabling the construction of evidence which a single search would not have uncovered. These systems are particularly effective for underspecified, multi-intent, or heterogeneous information needs [33,34,39,122–127].

7.2.2. Iterative and Adaptive Retrieval

This subgroup implements retrieval loops instead of a single one-shot lookup, where the retrieved evidence is refined across several rounds, often relying on intermediate outputs or retrieval-quality checks to prompt additional retrieval. In particular, ref. [128] draws on conversational history to enable multiturn retrieval, and [129] applies iterative cue-memory retrieval and refinement. The authors in [35] explore the utilization of iterative retrieval with detailed knowledge refinement, while the authors in [38] propose adaptive iterative retrieval. Finally, budget-constrained online retrieval is examined in [130], and the authors [131] employ layered retrieval in RAG systems for low-resource medical QA. Such methods are useful when the first retrieval pass is incomplete or when evidence needs to be progressively narrowed [35,38,128–131].

7.2.3. Self-Reflective, Self-Corrective, and Retrieval-Necessity Control

This subgroup covers RAG systems that assess their own evidence or determine whether retrieval is needed at all. For instance, ref. [36] applies retrieval-necessity control and answer validation for dual-strategy anomaly recovery, while the authors [132] integrate self-verification with retrieval in biomedical extraction. Additionally, the authors in [133] propose evidence-aware RAG through self-corrective chain-of-thought reasoning, while in [134] the retrieval is dynamically adjusted based on the internal and external knowledge

balance. Finally, Xiong et al. [40] address RAG conflicts between internal knowledge and retrieved evidence. The key contribution lies not in retrieving more, but in retrieving only when useful and revising evidence when inconsistencies appear [36,40,132–134].

7.2.4. Agentic and Multiagent Orchestration

This subgroup extends orchestration into planner–executor or multiagent systems, where distinct agents handle retrieval, reasoning, validation, or synthesis. In particular, ref. [37] presents a multiagent knowledge-graph RAG framework for intelligent maintenance. Furthermore, the authors in [135] utilize and orchestrate dynamically expert agents for social and emotional learning. Flores et al. [78] use expert agents to assist with government transparency, while a dynamic orchestration is also utilized in [136]. Furthermore, ref. [137] use retrieval augmentation to bridge laboratory practice and the scientific literature in nanostructured material design, while in [82] the authors integrate multimodule orchestration for insurance support. Agentic coordination is applied to system assurance in [138], where cooperating agents retrieve over threat intelligence and system documentation to identify and mitigate risks in machine-learning deployments. Here, retrieval functions as one component within a broader agentic control policy [37,78,82,135–138].

7.3. Structured-Memory RAG

The Structured-Memory RAG family includes 34 papers and describes methods whose main innovation lies in replacing flat document memory with structured knowledge bases such as graphs, ontologies, dual indexes, and hybrid relational stores. Rather than treating evidence as separate chunks, these systems maintain explicit topology, schema, or relational constraints, which in turn support graph-guided retrieval, multihop evidence assembly, and structure-aware grounding [41,43,45,139,140].

7.3.1. GraphRAG and Knowledge-Graph RAG

This subgroup includes and explores RAG systems where evidence is gathered through graph traversal, entity–relation lookup, or graph-grounded answer generation. Specifically, ref. [41] grounds retrieval in structured engineering design knowledge, while the authors in [141] develop graph retrieval for process-safety QA. Cross-domain knowledge retrieval for bio-inspired solution design is developed in [142]. In a similar approach, Liang et al. [143] introduce graph-based geospatial RAG, while the authors in [144] apply knowledge-graph-enhanced RAG to failure mode and effects analysis. Additionally, retrieval over a knowledge graph of cybersecurity vulnerabilities is developed in [61], while [145] incorporates graph retrieval-augmented generation into prescriptive recommender systems. The authors in [146] employ an RAG architecture based on graph memory for diet recommendation, while dynamic knowledge graphs and explainable retrieval are used in [147] to enhance LLMs for telecommunications. Ghosh et al. [148] apply context-aware, knowledge-graph-based RAG to engineering research, while KG-RAG is applied in [149] to earthquake emergency support, from knowledge-graph construction through to retrieval-augmented generation. The authors in [150] develop a hierarchical KG-based decision support system for a generic QA scenario, while Wang et al. [151] propose a hybrid system that merges hypergraph modeling with KG-enhanced retrieval. Graph and ontology retrieval is used in [152] to improve dietary-supplement information retrieval, while [153] applies precision retrieval to enhance medical LLMs. The authors in [154] apply knowledge-graph-based retrieval to detecting emergencies in patient-portal messages, while the authors in [155] construct a cultural-heritage KG-RAG system. Similarly, the approach in [156] integrates KG-RAG and knowledge graphs into intelligent decision-support systems, and [45] employs graph inference for vulnerability-description mapping. Finally, the combination of

knowledge graphs with RAG for wireless networks is examined in [157] through a tutorial that proposes a GraphRAG framework for networking and evaluates it in an original case study on channel-gain prediction. In [158], a knowledge base built from historical vulnerability fixes supports detection, as retrieving prior fixes grounds the model's judgment about new code. In [159], a hierarchical knowledge structure for embedded-system design is described, where retrieval descends from system level to component level. Finally, Zhang et al. [160] use meta-knowledge to guide the self-correction of LLM outputs against the retrieved structure. These systems maintain explicit topology during retrieval and are particularly valuable for enterprise semantics, infrastructure knowledge, and relational reasoning tasks [41,45,61,141–160].

7.3.2. Ontology- and Rule-Grounded RAG

This subgroup draws on formal schemas, ontologies, semantic web structures, or rule-based retrieval stores. In particular, ref. [139] grounds RAG in an OWL ontology for maintenance assistance, while the authors in [161] utilize standardized industrial models to organize retrieval in their RAG systems. Xiong et al. [42] perform an extensive and direct combination of domain rules with graph retrieval. Finally, the authors in [162] use retrieval augmentation to improve automated deep phenotyping with LLMs.

7.3.3. Vector-Graph and Dual-Index Architectures

This subgroup combines vector retrieval with graph memory or keeps dual indexes for complementary access paths. In [43], vector and graph retrieval is strategically combined on Brick-schema building data. The authors of [73] develop hybrid index retrieval for manufacturing knowledge, while [163] uses knowledge graphs and hybrid indexing for mass concrete mix design. Additionally, Wang et al. [74] use property graphs to index and retrieve in building engineering, reporting decreased hallucinations and increased accuracy in their evaluation setting. Finally, in [164], the authors introduce neural-symbolic dual-indexing architectures to achieve a balanced and accurate scalable RAG architecture. The goal of these RAG systems is to maintain both semantic flexibility and explicit relational structure, enabling the system to retrieve by similarity, topology, or both [43,73,74,163,164].

7.3.4. HybridRAG, Structured Hybrid, and Multihop Retrieval

This subgroup focuses on structured hybrid retrieval, where graph-aware or multihop evidence assembly is built directly into the retrieval substrate. More precisely, a dual-source retrieval for materials design is introduced in [140], while the authors in [44] gather evidence through subgraph retrieval and triplet ordering for knowledge-graph multihop QA. The key contribution lies in explicit relational evidence composition rather than simple similarity matching [44,140].

Table 9. Distribution of the reviewed studies across methodological families and subcategories. The decision rule governing each family is given in Section 2.4 (rules 4–9).

Family	Subcategory	Count	References
Retrieval-Optimized RAG (41)	Chunking, Parsing, Structure-Aware Segmentation	4	[25,32,97,99]
	Embedding Adaptation and Retrieval Alignment	9	[26,29,31,65,77,86,87,100,101]
	Sparse, Dense, and Hybrid Retrieval	6	[27,71,90,98,102,103]
	Reranking and Retrieval Refinement	3	[28,104,105]
	Retrieval Evaluation, Benchmarking, Comparative	19	[30,85,88,106–121]

Table 9. *Cont.*

Family	Subcategory	Count	References
Orchestrated RAG (27)	Query Planning, Decomposition, Routing	9	[33,34,39,122–127]
	Iterative and Adaptive Retrieval	6	[35,38,128–131]
	Self-Reflective, Self-Corrective, Retrieval-Necessity	5	[36,40,132–134]
	Agentic and Multiagent Orchestration	7	[37,78,82,135–138]
Structured-Memory RAG (34)	GraphRAG and Knowledge-Graph RAG	23	[41,45,61,141–160]
	Ontology- and Rule-Grounded RAG	4	[42,139,161,162]
	Vector-Graph and Dual-Index Architectures	5	[43,73,74,163,164]
	HybridRAG, Structured Hybrid, Multihop	2	[44,140]
Multimodal-Grounded RAG (18)	Vision-Language, Document-Centered, Remote-Sensing	11	[47,49,50,91,165–171]
	Multimodal KG and Cross-Modal Alignment	5	[48,62,92,172,173]
	Dual-Channel and Hybrid-Modal Retrieval	2	[46,174]
Governed RAG (12)	Privacy-Preserving and Encrypted RAG	3	[51–53]
	Secure, Access-Controlled, Federated RAG	1	[93]
	Trust, Provenance, Compliance, Policy-Aware RAG	2	[79,94]
	Robustness, Unlearning, Security-Aware RAG	6	[175–180]
Task-Compositional RAG (59)	Classification and Assessment	6	[54,83,181–184]
	Information Extraction	6	[185–190]
	Code and Query Generation	6	[191–196]
	Recommendation	6	[56,197–201]
	Report and Document Generation	9	[55,95,96,202–207]
	Grounded Guidance and Question Answering	20	[57–60,81,84,208–221]
	Control, Optimization, and Other Outputs	6	[222–227]

7.4. Multimodal-Grounded RAG

The Multimodal-Grounded RAG family includes 18 papers and consists of RAG architectures where retrieval and generation rely on evidence spread across several modalities, including text, images, video, document layout, or remote-sensing imagery. In these systems, evidence cannot be simplified to plain text; rather, meaningful grounding arises from the alignment of visual, textual, spatial, and document-level signals [46–48,50,165].

7.4.1. Vision-Language, Document-Centered, and Remote-Sensing-Grounded RAG

This subgroup explores vision-language RAG, document-centered multimodal retrieval, image-grounded QA, and remote-sensing-grounded systems. In particular, ref. [50] applies visual retrieval augmentation to the recognition of intangible cultural heritage imagery, while the authors in [91] build a vision-language foundation-model-based multimodal RAG framework for remote-sensing lithological recognition. Image-grounded retrieval is applied in [49] to the AI-assisted interpretation of children's drawings, while the authors in [165] develop a dynamic combination of multimodal perception with retrieved manuals for robotics support. Wang et al. [166] build a football-coaching support system on multimodal retrieval over heterogeneous inputs. Multimodal retrieval can also enhance the effectiveness and accuracy of RAG systems for Alzheimer-related support, as presented in [167], or in [168], that merges dermatology vision with retrieval grounding. Furthermore, the authors of [169] effectively integrate multimodal cognitive evidence in medical support, as well as [47], which bridges remote-sensing imagery and domain knowledge through a multimodal dataset and a retrieval-augmented model. The authors in [170] present a reference retrieval system for the restoration of ancient paintings, and visual and textual references for stylistically similar works are retrieved. Relational-reasoning image captioning is addressed in [171] through multiagent retrieval augmentation. These approaches are

significant when relevant information is spread across images, text, document structure, or spatial scenes rather than within textual passages alone [47,49,50,91,165–170].

7.4.2. Multimodal Knowledge Graph and Cross-Modal Alignment

This subgroup connects visual and textual evidence, often by means of multimodal knowledge graphs or cross-modal retrieval structures. In particular, ref. [172] combines multimodal models with graph-based retrieval for accident-video analysis, while a unified multimodal knowledge graph is used in [48] to enhance question answering. Additionally, the authors in [62] develop a robust cross-modal alignment around multimodal document evidence, while [92] expands medical retrieval with multimodal graphs and linked image-text evidence. Finally, in [173], the multimodal perception is explored and combined with graph-enabled RAG for orbital-debris identification, with the authors reporting improved identification accuracy relative to their non-retrieval baseline. These systems therefore tackle not only retrieval but also the semantic compatibility of heterogeneous evidence sources [48,62,92,172,173].

7.4.3. Dual-Channel and Hybrid-Modal Retrieval

This subgroup performs retrieval separately across modalities before combining the resulting evidence, as Guo et al. [46] develop a dual-channel retrieval design for construction-site compliance. Also, the authors in [174] divide and utilize text and image retrieval to enhance the accuracy and robustness of RAG approaches in manufacturing QA. The contribution here lies in architectural control over multimodal access instead of merely adding visual inputs [46,174].

7.5. Governed RAG

The Governed RAG family includes 12 papers and consists of methods that control retrieval based on privacy, security, access, trust, compliance, or robustness constraints instead of relying solely on semantic relevance. Within this family, retrieval functions as a governed interface that needs to meet institutional and operational requirements along with ensuring answer quality [53,93,94,175].

7.5.1. Privacy-Preserving and Encrypted RAG

This subgroup covers encrypted retrieval, privacy-preserving search, perturbation-based confidentiality mechanisms, and protected vector operations. In particular, ref. [53] presents locally private entity perturbation for safer retrieval, while in [52] the authors propose an efficient privacy-preserving retrieval mechanism within the RAG architecture. In addition, Chang et al. [51] integrates encrypted retrieval with secure cloud-based generation. These approaches aim to preserve retrieval utility while reducing information leakage from the retrieval substrate itself.

7.5.2. Secure, Access-Controlled, and Federated RAG

This subgroup focuses on controlled knowledge exposure through secure deployment, access control, or federated retrieval. Authors [93] make a wide exploration of various federated retrieval settings within RAG systems. These approaches restrict not only what is relevant but also what is allowed to be retrieved.

7.5.3. Trust, Provenance, Compliance, and Policy-Aware RAG

This subgroup explores the provenance-aware evidence access, policy-sensitive filtering, accountable retrieval, and compliance-oriented governance rules. For instance, ref. [94]

connects retrieval with environmental or policy-governed decision support, while [79] anchors legal AI in dependable legal sources.

7.5.4. Robustness, Unlearning, and Security-Aware RAG

This subgroup studies adversarial robustness, poisoning resistance, machine unlearning, forgetting, and security-aware retrieval redesign. Specifically, ref. [175] investigates security-aware intrusion detection within RAG systems, and in [176], knowledge-poisoning attacks are analyzed and mitigated in RAG-based LLM systems. The authors in [177] report improved adversarial resilience for semantic caching in terms of privacy and security with RAG utilization, while the authors in [178] explore forgetting and suppression dynamics within machine-unlearning-oriented retrieval safety. Retrieval poisoning is countered in [179] through a push–pull strategy that reshapes the embedding space to favor genuine evidence over adversarially injected passages. For recommendation systems, a purification-based strategy is proposed [180] to attenuate the effect of the perturbed histories, rather than supplementing them with additional evidence.

7.6. Task-Compositional RAG

The Task-Compositional RAG family includes 59 papers and describes systems where RAG serves as one element within a broader task-oriented workflow. In this group, retrieval supports a downstream framework pipeline like a diagnosis, recommendation, report, executable query, plan, or operational decision. Here, retrieval functions as an intermediate grounding step within a larger chain of action or reasoning. Within this family, studies are grouped by composition pattern, that is, by the type of artifact that the workflow produces: a classification or assessment, extracted information, executable code or queries, a recommendation, a report, grounded guidance, or another workflow output.

7.6.1. Classification and Assessment

In this subgroup, the workflow ends in a label, a class or an assessment. Macieira et al. [181] standardize nursing interventions in electronic health records using AI-assisted retrieval, while [54] reports improved precision and trust for a GPT-4 model in emergency radiology diagnosis and classification. Multilabel discipline classification of academic short texts is addressed in [182]. In software and security settings, ref. [83] focuses on software vulnerability assessment, ref. [183] supports smart-contract vulnerability detection, and [184] performs multilevel root-cause analysis of deployed LLM applications, where the system under diagnosis is itself retrieval-augmented. In these studies, retrieval supplies the evidence on which the label is based.

7.6.2. Information Extraction

Here, retrieval supports the extraction of structured information from unstructured sources. In clinical settings, ref. [185] summarizes and extracts key clinical information from electronic health records, ref. [186] extracts clinical information from mesh-implant narratives with a zero-shot approach, and [187] supports clinical entity recognition and code mapping of anatomopathological reports. The authors of [188] use retrieval augmentation to support the causal identification of aviation accidents, semiconductor parameter extraction is grounded in retrieved device documentation in [189], and automated log parsing is addressed in [190].

7.6.3. Code and Query Generation

In this subgroup, the output is executable, such as code, a test case or a query. RAG is used in [191] to automate software test-case generation, while hardware description language

debugging and Verilog bug identification and correction are addressed in [192,193], in each case by retrieving over specifications and prior defect records rather than relying on parametric knowledge of the toolchain. In [194], retrieval over existing vulnerability records supports the generation of training data for vulnerability-detection models. LLM-based querying of building data is improved in [195] by combining chain-of-thought prompting, retrieval augmentation and fine-tuning, and [196] develops a conversational machine-monitoring assistant for CNC environments that is grounded in real-time data retrieval.

7.6.4. Recommendation

The output of this subgroup is a selected or ranked recommendation. An intelligent Chinese patent medicine recommendation framework is designed and tested in [197], a sepsis-oriented knowledge platform for personalized management is constructed in [198], and [199] compares RAG-based LLMs with junior clinicians for guideline-concordant prostate-specific antigen testing. Outside medicine, ref. [56] supports next-POI recommendation through trajectory retrieval, conversational recommendation over event and venue records is addressed in [200], and [201] augments recommendation with retrieval over item descriptions and review text, so that suggestions can be justified from evidence rather than from latent similarity alone.

7.6.5. Report and Document Generation

In this subgroup, retrieval grounds the generation of a longer document. The authors of [95] use retrieval augmentation to improve LLM-generated medication-use instructions, and dietary monitoring and personalized nutritional analysis are supported in [202], where retrieval grounds the interpretation of logged intake. In engineering, ref. [203] applies RAG to oscillation analysis and [204] generates 6G network specifications for use cases. In the built environment, ref. [96] studies the generation of construction-safety information, ref. [55] generates construction claim reports through tree-mapped retrieval, and [205] evaluates urban low-carbon governance. Finally, ref. [206] applies RAG to cyber-incident timeline analysis, and [207] optimizes customer-feedback summarization.

7.6.6. Grounded Guidance and Question Answering

This subgroup comprises assistants built for a defined operational task, whose output is guidance that is read and acted on within that task. In medicine, ref. [208] develops a postoperative virtual assistant, ref. [84] develops an RAG-based LLM for drug-induced liver injury, ref. [209] studies patient-facing informed-consent assistance, ref. [210] analyzes a vaccination-support chat system, ref. [59] develops an assistant for protocol queries in anatomical pathology laboratories, and [211] augments clinical generation with evidence drawn from multiple heterogeneous sources, so that a response can be traced to the record that supports it. In industry and construction, RAG supports smart-manufacturing quality control in [58] and context-aware information retrieval for the Internet of Everything in [212], while knowledge retrieval and question answering for construction management and building regulations are addressed in [213,214], respectively. Knowledge assistants are also proposed for goat farmers in [215] and for tutoring in [216,217], the latter for curriculum-aware bilingual tutoring in low-resource Swahili–English secondary schools. Enterprise and public-sector assistants include enterprise-specific question answering for IT operations and maintenance [57], customs-clearance assistance [218], security event response [219], Vietnamese legal question answering grounded in retrieved statutory text [60], and guidance for refugees grounded in retrieved official documentation [220]. In consumer settings, ref. [221] explains food additives in a consumer-facing application and [81] applies item-based knowledge computing to e-commerce product question answering. Because

the output of these systems is read and acted on by people, their value lies in making the source of each answer traceable.

7.6.7. Control, Optimization, and Other Outputs

The remaining studies use retrieval to support control, optimization or system-level decisions, or they report on the engineering of RAG systems. RAG is applied to resource optimization in 6G and beyond wireless networks in [222], to GenAI-enabled semantic communications in [223], to human–robot interaction in [224], and to adaptive task offloading for edge computing in the Internet of Energy in [225]. Adaptive deception is addressed in [226], where graph-structured retrieval drives protocol-faithful honeypot responses. Finally, ref. [227] reports on the engineering of RAG-based virtual assistants deployed in enterprise practice.

7.7. Answers to the Research Questions

RQ1. The reviewed systems innovate along six distinguishable loci of intervention relative to a baseline retrieve-then-generate pipeline: the retrieval mechanism, the control policy, the knowledge substrate, the evidence modality, a governance constraint, and the role of retrieval within a wider workflow. The decision rules of Section 2 assign each study to exactly one of the six corresponding families.

RQ2. Task-Compositional RAG (59 studies, 30.9%) and Retrieval-Optimized RAG (41, 21.5%) are the largest families, followed by Structured-Memory (34, 17.8%), Orchestrated (27, 14.1%), Multimodal-Grounded (18, 9.4%) and Governed RAG (12, 6.3%) (Table 4). Healthcare and biomedicine (55 studies) and general-purpose methods (41) are the most frequent application domains (Figure 3 and Table 3).

RQ3. The families differ in the failure they address and in where they incur cost (Tables 7 and 8). Retrieval-Optimized and Structured-Memory RAG pay a one-off cost for indexing, adaptation or store construction and a per-query cost for reranking, fusion or graph traversal; Orchestrated RAG pays for extra retrieval rounds and tokens on each request; Governed RAG trades answer utility for protection; and in Task-Compositional RAG the contribution of retrieval is coupled with the rest of the workflow.

RQ4. Across the 191 included studies, downstream task metrics are reported in 134 (70.2%), hallucination, robustness and efficiency metrics in 75 (39.3%), retriever-level metrics in 52 (27.2%), generator-level metrics in 41 (21.5%), human evaluation in 49 (25.7%) and significance testing in 51 (26.7%) (Table 6). Retrieval quality and grounding are therefore measured in a minority of studies, and only 26 (13.6%) measure both, so that the contribution of retrieval to the reported gains is usually not isolated.

RQ5. Each family leaves one open problem (Table 8): whether chunking and embedding choices transfer across corpora (Retrieval-Optimized), when further retrieval is needed and when to stop (Orchestrated), how to keep the store consistent as the corpus changes (Structured-Memory), how to measure non-text grounding directly (Multimodal-Grounded), how much utility is lost relative to an unprotected baseline (Governed), and how much of the reported performance is attributable to retrieval (Task-Compositional). Section 8 derives the corresponding research directions.

8. Limitations, Challenges and Future Directions

Despite the rapid progress of RAG, several core challenges still remain. First, retrieval quality continues to be a major bottleneck. Many systems fail because relevant evidence is segmented poorly, ranked too low, or mixed with distracting context, which directly affects grounding quality and hallucination control [30,32,65]. Future work should therefore

focus on retrieval calibration, including improved chunking, stronger reranking, evidence sufficiency estimation, and domain-adapted retrievers.

Additionally, recent systems indicate that retrieval should not always be treated as a single static step. Query rewriting, routing, iterative retrieval, self-correction, and agentic control can improve performance in difficult settings, but they also increase system complexity and cost [33,35,40,78]. A key direction for future research is to develop more efficient and better-controlled orchestration policies that determine when retrieval is needed, how many retrieval rounds are justified, and how conflicting evidence should be resolved.

The present review synthesizes reported architectural strategies, but does not perform a meta-analysis or a comparative risk-of-bias assessment, and it therefore cannot establish that RAG systematically improves accuracy, robustness, or trustworthiness relative to well-tuned parametric baselines. Establishing that claim would require controlled, benchmark-matched comparisons that the current literature does not consistently provide, and this is itself a substantive research gap.

Retrieval may also lower the quality of output, as it can introduce irrelevant, outdated, or contradictory evidence that confuses the generator. Poisoned or adversarial documents can also be injected into the knowledge base, while sensitive information may be exposed in the retrieved context [51,53,176]. Thus, gains reported in the corpus are conditional on the quality level of the retrieval and should not be generalized.

Certain limitations are imposed by the search strategy, which may constrain the conclusions drawn from this article. The first is that only journal and magazine articles were included in the database, and not conference papers or non-archival repositories. These were used as background references where relevant, but are not included in this database or its classification. Also, the time frame for this review ends on 27 April 2026; therefore, the list should be seen as a snapshot of the dynamic literature in this area rather than a complete record.

Furthermore, the design of external memory remains an open question. Graphs, ontologies, and dual-index systems often strengthen multihop reasoning and relational grounding, but they are expensive to construct, maintain, and update [42,43,45]. Future research should therefore examine dynamic structured memory, hybrid vector-graph stores, and automated methods for keeping external knowledge bases consistent over time.

Multimodal and governed RAG introduce additional deployment challenges, as the main difficulty lies in aligning text with images, video, layout, or spatial evidence in a coherent manner [47,50,62]. In governed systems, retrieval must meet privacy, security, compliance, and robustness requirements in addition to relevance [51,52,79].

Finally, evaluation remains fragmented across the literature. Stronger benchmark design is needed so that retrieval quality, grounding quality, downstream utility, latency, and human trust can be assessed together rather than separately [30,59]. Overall, the main future direction for RAG is the development of knowledge-grounded systems that are adaptive, efficient, structured, multimodal, and ready for deployment.

9. Conclusions

RAG has become a widely adopted means for grounding LLM outputs in external knowledge, and the literature reports gains in factual accuracy, relevance, transparency, and adaptability as compared to standalone parametric generation. In this methodological review, RAG architectures, application domains, and evaluation practices are analyzed on the basis of 191 peer-reviewed journal papers published between January 2024 and 27 April 2026. This survey also introduces a new six-family taxonomy of RAG systems, to showcase that RAG is no longer just a single retrieve-then-generate pipeline but a broader methodological ecosystem where different families address different aspects within the

architecture. This review further demonstrates that RAG evaluation should not depend on a single measure but should combine retriever-level, generator-level, reference-based, downstream-task, and operational or human-centered metrics. Overall, the use of RAG in LLM-based systems is a rapidly expanding trend across many domains. Although key challenges remain, including retrieval reliability, evaluation fragmentation, multimodal alignment, and governance-aware deployment, RAG remains an active and rapidly developing direction for grounding generative AI systems in external knowledge, and the open problems identified above define where its reliability must next be established.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/technologies14110715/s1, Table S1: Paper-level classification of the 191 included studies; Table S2: PRISMA-ScR checklist.

Author Contributions: Conceptualization, V.P.R. and N.V.; methodology, V.P.R. and N.V.; software, V.P.R.; validation, V.P.R. and N.V.; formal analysis, V.P.R. and N.V.; investigation, V.P.R.; resources, N.V.; data curation, V.P.R.; writing—original draft preparation, V.P.R.; writing—review and editing, V.P.R. and N.V.; visualization, V.P.R. and N.V.; supervision, N.V.; project administration, N.V.; funding acquisition, N.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financed by the EU Horizon Europe framework, under grant agreement no. 101177687 (ISABEL project).

Data Availability Statement: This review generated a bibliographic and classification dataset comprising, for each included study, the full citation, the inclusion decision, the assigned primary methodological family, the subcategory, and the application domain. The classification results are reported in aggregate in Table 4, at subcategory level with full reference lists in Table 9, and by application domain in Figure 3. The search strings, databases, filters, and search dates required to reproduce the identification and screening stages are reported in full in Section 2. The paper-level classification of every included study is provided as Supplementary Table S1, and the PRISMA-ScR checklist as Supplementary Table S2.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIGC	Artificial Intelligence Generated Content
CNNs	Convolutional Neural Networks
DF	Document Frequency
EM	Exact Match
FiD	Fusion-in-Decoder
GenAI	Generative Artificial Intelligence
ICC	Intraclass Correlation Coefficient
IDF	Inverse Document Frequency
IR	Information Retrieval
KGs	Knowledge Graphs
LCS	Longest Common Subsequence
LLMs	Large Language Models
MAP	Mean Average Precision
MCC	Matthews Correlation Coefficient
MRR	Mean Reciprocal Rank
nDCG	normalized Discounted Cumulative Gain
NLP	Natural Language processing
RAG	Retrieval-Augmented Generation
RNNs	Recurrent Neural Networks
SOTA	State-of-the-art
TF	Term Frequency

References

1. Hindi, M.; Mohammed, L.; Maaz, O.; Alwarafy, A. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access* **2025**, *13*, 46171–46189. [CrossRef]
2. Li, Z.; Wang, Z.; Wang, W.; Hung, K.; Xie, H.; Wang, F.L. Retrieval-augmented generation for educational application: A systematic survey. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100417. [CrossRef]
3. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
4. Yang, W.; Some, L.; Bain, M.; Kang, B. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowl.-Based Syst.* **2025**, *318*, 113503. [CrossRef]
5. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–38. [CrossRef]
6. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
7. Akbar, N.A.; Dembani, R.; Lenzitti, B.; Tegolo, D. RAG-Driven Memory Architectures in Conversational LLMs—A Literature Review With Insights Into Emerging Agriculture Data Sharing. *IEEE Access* **2025**, *13*, 123855–123880. [CrossRef]
8. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997.
9. Miao, Y.; Zhao, Y.; Luo, Y.; Wang, H.; Wu, Y. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. *J. Med. Internet Res.* **2025**, *27*, e80557. [CrossRef] [PubMed]
10. Zheng, X.; Weng, Z.; Lyu, Y.; Jiang, L.; Xue, H.; Ren, B.; Paudel, D.; Sebe, N.; Van Gool, L.; Hu, X. Retrieval Augmented Generation and Understanding in Vision: A Survey and New Outlook. *arXiv* **2025**, arXiv:2503.18016.
11. Huang, Y.; Huang, J.X. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *ACM Comput. Surv.* **2026**, *58*, 300:1–300:38. [CrossRef]
12. Zhu, Z.; Huang, T.; Wang, K.; Ye, J.; Chen, X.; Luo, S. Graph-Based Approaches and Functionalities in Retrieval-Augmented Generation: A Comprehensive Survey. *ACM Comput. Surv.* **2026**, *58*, 261. [CrossRef]
13. Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; Tang, S. Graph Retrieval-Augmented Generation: A Survey. *ACM Trans. Inf. Syst.* **2025**, *44*, 35:1–35:52. [CrossRef]
14. Plaat, A.; van Duijn, M.; van Stein, N.; Preuss, M.; van der Putten, P.; Batenburg, K.J. Agentic Large Language Models, a Survey. *J. Artif. Intell. Res.* **2025**, *84*, 29. [CrossRef]
15. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. [CrossRef] [PubMed]
16. Liu, S.; McCoy, A.B.; Wright, A. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 605–615. [CrossRef] [PubMed]
17. Neha, F.; Bhati, D.; Shukla, D.K. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI* **2025**, *6*, 226. [CrossRef]
18. Brown, A.; Roman, M.; Devereux, B. A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *Big Data Cogn. Comput.* **2025**, *9*, 320. [CrossRef]
19. Yang, R.; Wong, M.Y.H.; Li, H.; Li, X.; Zhu, W.; Liao, J.; Yu, K.; Liew, J.C.K.; Xuan, W.; Chen, Y.; et al. Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations. *Cell Rep. Med.* **2026**, 102927. [CrossRef] [PubMed]
20. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 6491–6501.
21. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. In *Proceedings of the Big Data (BigData 2024)*; Communications in Computer and Information Science; Springer: Singapore, 2025.
22. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
23. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
24. Kitchenham, B.; Brereton, O.P.; Budgen, D.; Turner, M.; Bailey, J.; Linkman, S. Systematic literature reviews in software engineering—A systematic literature review. *Inf. Softw. Technol.* **2009**, *51*, 7–15. [CrossRef]

25. Gomez-Cabello, C.A.; Prabha, S.; Haider, S.A.; Genovese, A.; Collaco, B.G.; Wood, N.G.; Bagaria, S.; Forte, A.J. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering* **2025**, *12*, 1194. [CrossRef] [PubMed]
26. Vake, D.; Vičič, J.; Tošić, A. Bridging the question-answer gap in retrieval-augmented generation: Hypothetical prompt embeddings. *IEEE Access* **2025**, *13*, 129952–129961. [CrossRef]
27. Bornea, A.L.; Ayed, F.; De Domenico, A.; Piovesan, N.; Salem, T.S.; Maatouk, A. Telco-orag: Optimizing retrieval-augmented generation for telecom queries via hybrid retrieval and neural routing. *IEEE J. Sel. Areas Commun.* **2025**, *44*, 2334–2347. [CrossRef]
28. Yang, Q.; Zuo, H.; Su, R.; Su, H.; Zeng, T.; Zhou, H.; Wang, R.; Chen, J.; Lin, Y.; Chen, Z.; et al. Dual retrieving and ranking medical large language model with retrieval augmented generation. *Sci. Rep.* **2025**, *15*, 18062. [CrossRef] [PubMed]
29. Collini, E.; Kurniadi, F.I.; Nesi, P.; Pantaleo, G. Context-aware retrieval augmented generation using similarity validation to handle context inconsistencies in large language models. *IEEE Access* **2025**, *13*, 170065–170080. [CrossRef]
30. Elkiran, H.; Rasheed, J. An Empirical Evaluation of Retrieval, Reranking, and Similarity for a Q&A-Based Retrieval Augmented Generation System. *IEEE Access* **2026**, *14*, 26053–26066. [CrossRef]
31. Krainovskikh, V.; Samigulin, T. Domain-Adapted Retrieval-Augmented Generation for Technical Documentation: Enhancing Reliability and Faithfulness in Technical QA. *IEEE Access* **2026**, *14*, 36016–36024. [CrossRef]
32. Koay, X.K.; Ong, L.Y.; Goh, P.Y. Structure-Aware Chunking for Complex Tables in Retrieval-Augmented Generation Systems. *Emerg. Sci. J.* **2026**, *10*, 184–205. [CrossRef]
33. Vidyarthi, A.; Singh, M.K.; Moirangthem, D.S. SageRAG: Query Rewriting for Retrieval Enhancement and Retrieval-Augmented Generation for Grounded Responses in AI Research Assistance. *Expert Syst. Appl.* **2026**, *308*, 131160. [CrossRef]
34. Huang, Y.; Yang, L.; Yang, X.H.; Xu, X. Retrieval-Augmented Generation for Multi-Hop Question Answering Based on Structured Planning. *ACM Trans. Knowl. Discov. Data* **2026**, *20*, 1–20. [CrossRef]
35. Du, K.; Wang, W.; Zhang, B.; Wang, P.; Zhang, F.; Cao, L.; Guo, Y. IRAGKR: Iterative retrieval augmented generation with fine-grained knowledge refinement. *Neurocomputing* **2025**, *655*, 131282. [CrossRef]
36. Lv, J.; Li, Z.; Gan, J.; Ren, W.; Chen, S.; Wang, J. Dual-strategy retrieval-augmented generation for anomaly recovery in large language models. *Pattern Recognit.* **2026**, *178*, 113430. [CrossRef]
37. Lin, J.; Chen, C.; Wang, T.; Wang, Z.; Cheng, L. A Multi-Agent and synergistic Knowledge Graph retrieval-augmented generation framework for intelligent maintenance. *J. Manuf. Syst.* **2026**, *85*, 557–570. [CrossRef]
38. Han, W.; Xiao, X.; Li, Y.; Wang, J.; Pechenizkiy, M.; Fang, M. Adaptive iterative retrieval for enhanced retrieval-augmented generation. *Neurocomputing* **2025**, *666*, 132272. [CrossRef]
39. Zhan, Z.; Wang, J.; Zhou, S.; Deng, J.; Zhang, R. MMRAG: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1505–1516. [CrossRef] [PubMed]
40. Xiong, Y.; Chen, Y.; Zhang, H. ICR: A Framework for Resolving Knowledge Conflicts in Retrieval-Augmented Generation. *Neurocomputing* **2025**, *664*, 132139. [CrossRef]
41. Siddharth, L.; Luo, J. Retrieval augmented generation using engineering design knowledge. *Knowl.-Based Syst.* **2024**, *303*, 112410. [CrossRef]
42. Xiong, X.; Cai, H.; Yu, H.; Shen, B.; Hu, P. DR-RAG: Domain-Rule-based Retrieval-Augmented Generation for aviation digital model design. *Adv. Eng. Inform.* **2025**, *68*, 103688. [CrossRef]
43. Li, M.; Wang, Z. BuildingGPT: Query building semantic data using large language models and vector-graph retrieval-augmented generation. *Build. Environ.* **2025**, *287*, 113855. [CrossRef]
44. Saleh, A.O.; Tur, G.; Saygin, Y. SG-RAG MOT: SubGraph retrieval augmented generation with merging and ordering triplets for knowledge graph multi-hop question answering. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 74. [CrossRef]
45. Yu, F.; Du, Y. VDM-IOG, a framework of inference on graph in retrieval-augmented generation for vulnerability description mapping. *Cybersecurity* **2026**, *9*, 14. [CrossRef]
46. Guo, K.X.; Wong, P.K.Y.; Cheng, J.C.P.; Chan, C.F.; Leung, P.H.; Tao, X. Enhancing visual-LLM for construction site safety compliance via prompt engineering and Bi-stage retrieval-augmented generation. *Autom. Constr.* **2025**, *179*, 106490. [CrossRef]
47. Wen, C.; Lin, Y.; Qu, X.; Li, N.; Liao, Y.; Li, X.; Lin, H. Remote Sensing Retrieval-Augmented Generation: Bridging remote sensing imagery and comprehensive knowledge with a multimodal dataset and retrieval-augmented generation model. *IEEE Geosci. Remote Sens. Mag.* **2026**, *14*, 85–103. [CrossRef]
48. Ali, Z.; Huang, Y.; Khan, A.; Qi, G.; Zhang, Y.; Feng, J.; Deng, C.; Kefalas, P. Pythia-RAG: Retrieval-Augmented Generation over a Unified Multimodal Knowledge Graph for Enhanced QA. *Knowl.-Based Syst.* **2025**, *335*, 115200. [CrossRef]
49. Aljalal, H.M.; Alshareef, H.H.; Alduraib, N.M.; Hamdoun, H.M.; Alsedais, R. AI-Assisted Interpretation of Children's Drawings via Retrieval-Augmented Generation. *IEEE Access* **2026**, *14*, 25537–25550. [CrossRef]
50. Fan, T.; Wang, H.; Zhao, Y.; Zhou, S.; Shi, B. When MLLMs Meet ICH: A visual retrieval-augmented generation-based method for intangible cultural heritage image recognition-take Shadow Puppetry as a case. *Inf. Process. Manag.* **2026**, *63*, 104481. [CrossRef]

51. Chang, Y.; Luan, T.H.; Wang, S.; Wang, Y. SafeRAG: Secure Cloud-Based Retrieval-Augmented Generation for LLM-Empowered Voice Assistants. *IEEE Trans. Netw. Sci. Eng.* **2026**, *13*, 6211–6224. [CrossRef]
52. Zhou, J.; Wu, J. Efficient Vector-Multiplicative Privacy-Preserving Retrieval-Augmented Generation for Large Language Models. *IEEE Trans. Dependable Secur. Comput.* **2026**, *23*, 6947–6963. [CrossRef]
53. He, L.; Tang, P.; Zhang, Y.; Zhou, P.; Su, S. Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation. *Inf. Process. Manag.* **2025**, *62*, 104150. [CrossRef]
54. Fink, A.; Nattenmüller, J.; Rau, S.; Rau, A.; Tran, H.; Bamberg, F.; Reisert, M.; Kotter, E.; Diallo, T.; Russe, M.F. Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: A proof-of-concept study. *Eur. Radiol.* **2025**, *35*, 5091–5098. [CrossRef] [PubMed]
55. Zhu, W.; Li, X.; Wang, L.; Wang, J.; Wei, Y. TM-RAG: A Tree-Mapped Retrieval-Augmented generation framework for construction claim report generation. *Adv. Eng. Inform.* **2026**, *69*, 104092. [CrossRef]
56. Lin, Z.; Zhang, X.; Zhao, K.; Lu, X.; Zhang, Y. STrajRAG: Supervised trajectory retrieval augmented generation for next POI recommendation with travel semantics. *Inf. Process. Manag.* **2025**, *62*, 104235. [CrossRef]
57. Jiang, Z.; Zhang, T.; Bai, S.; Lin, L.; Zhang, H.; Xun, Y.; Ren, J.; Si, W.; Zhang, S. Towards Enterprise-Specific Question-Answering for IT Operations and Maintenance based on Retrieval-Augmented Generation Mechanism. *Expert Syst. Appl.* **2025**, *307*, 130961. [CrossRef]
58. Álvaro, J.A.H.; Barreda, J.G. An advanced retrieval-augmented generation system for manufacturing quality control. *Adv. Eng. Inform.* **2025**, *64*, 103007. [CrossRef]
59. Pires, D.; Perezhohin, Y.; Castelli, M. Retrieval-Augmented Generation Assistant for Anatomical Pathology Laboratories. *Emerg. Sci. J.* **2025**, *9*, 3116–3128. [CrossRef]
60. Le, H.; Luu, N.; Nguyen, T.; Dao, T.; Dinh, S. Optimizing Answer Generator in Vietnamese Legal Question Answering Systems Using Language Models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–17. [CrossRef]
61. Gao, X.; Chang, X. A Retrieval-Augmented Generation Framework Based on a Knowledge Graph of Cybersecurity Vulnerabilities in Power Networks. *IEEE Access* **2025**, *13*, 186693–186710. [CrossRef]
62. How, S.K.; Ong, L.Y.; Leow, M.C. Enhance Multimodal Retrieval-Augmented Generation Using Multimodal Knowledge Graph. *Emerg. Sci. J.* **2025**, *9*, 3349–3361. [CrossRef]
63. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6769–6781.
64. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 874–880.
65. Saha, B.; Saha, U.; Malik, M.Z. QuIM-RAG: Advancing retrieval-augmented generation with inverted question matching for enhanced QA performance. *IEEE Access* **2024**, *12*, 185401–185410. [CrossRef]
66. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1762–1777.
67. Jiang, Z.; Xu, F.F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; Neubig, G. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7969–7992.
68. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
69. Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884.
70. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
71. Hao, J.X.; Chen, L.; Meng, L. Advancing large language models with enhanced retrieval-augmented generation: Evidence from biological UAV swarm control. *Drones* **2025**, *9*, 361. [CrossRef]
72. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 39–48.
73. Wan, Y.; Chen, Z.; Liu, Y.; Chen, C.; Packianather, M. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Adv. Eng. Inform.* **2025**, *65*, 103212. [CrossRef]
74. Wang, Z.; Liu, Z.; Lu, W.; Jia, L. Improving knowledge management in building engineering with hybrid retrieval-augmented generation framework. *J. Build. Eng.* **2025**, *103*, 112189. [CrossRef]

75. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
76. Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; Manning, C.D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
77. El Bazzi, M.S.; Taha, A.; Ait Ider, A. Advancing Multilingual Retrieval-Augmented Generation for Reliable Medication Counseling. *IEEE Access* **2025**, *13*, 215550–215564. [CrossRef]
78. Flores, N.; Ramirez, C.; Mauricio, D. Expert Multi-Agent Conversational System Using Retrieval-Augmented Generation and Dynamic Text-to-SQL for Government Transparency. *IEEE Access* **2025**, *13*, 198178–198200. [CrossRef]
79. Gupta, S.; Agrawal, V.; Negi, Y.; Karunanithi, S.; Balakrishnan, A. Reducing Hallucinations in Legal AI: A Retrieval Augmented Generation Based Model for Accurate Legal Guidance. *IEEE Access* **2026**, *14*, 44451–44463. [CrossRef]
80. Es, S.; James, J.; Espinosa Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 150–158.
81. Xu, C.; Kang, Y.; Feng, Q.; Hua, J.; Li, P.; Wu, F.; Wei, H.; Chen, X.; Huang, S.J.; Chen, S. ItemRAG: Retrieval-augmented generation with item-based knowledge computing for e-commerce product question answering. *Big Data Min. Anal.* **2026**, *9*, 407–424. [CrossRef]
82. Hanmante, S.; Patil, S.; Shahade, A.K. A multi module ai system for intelligent health insurance support using retrieval augmented generation. *Sci. Rep.* **2025**, *16*, 1403. [CrossRef] [PubMed]
83. Chen, Z.; Chen, X.; Li, Z.; Xue, J.; Gao, C. ReVul-CoT: Towards Effective Software Vulnerability Assessment with Retrieval-Augmented Generation and Chain-of-Thought Prompting. *J. Syst. Softw.* **2026**, *237*, 112855. [CrossRef]
84. Rao, A.; Cholankeril, G.; Flores, A.; Sood, G.; White, T.E.; Kanwal, F.; El-Serag, H.B. Development of retrieval-augmented generation–based large language model for drug-induced liver injury using Livertox data. *Hepatol. Commun.* **2026**, *10*, e00895. [CrossRef] [PubMed]
85. Pasaribu, J.; Yudistira, N.; Mahmudy, W.F. Tabular data classification and regression: Xgboost or deep learning with retrieval-augmented generation. *IEEE Access* **2024**, *12*, 191719–191732. [CrossRef]
86. Malik, M.S.; Le, V.T.; Ou, Y.Y. Enhancing the classification of metal-binding residue in proteins with retrieval-augmented generation, protein language models, and deep learning. *Eng. Appl. Artif. Intell.* **2026**, *171*, 114330. [CrossRef]
87. Malik, M.S.; Hussain, M.; Ho, Q.T.; Le, V.T.; Ou, Y.Y. RAG-msCNN: Predicting protein-DNA binding sites by integration of retrieval-augmented generation (RAG) with protein language models and the multi-scale separable convolutional neural network. *Biomed. Signal Process. Control* **2026**, *115*, 109407. [CrossRef]
88. Chen, Z.; Min, E.; Zhao, X.; Li, Y.; Jia, X.; Liao, J.; Li, J.; Wang, S.; Hu, B.; Yin, D. A Question answering Dataset for temporal-Sensitive Retrieval-augmented Generation. *Sci. Data* **2025**, *12*, 1855. [CrossRef] [PubMed]
89. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
90. Alawwad, H.A.; Alhothali, A.; Naseem, U.; Alkhathlan, A.; Jamal, A. Enhancing textual textbook question answering with large language models and retrieval augmented generation. *Pattern Recognit.* **2025**, *162*, 111332. [CrossRef]
91. Chen, X.; Liu, Y.; Han, W.; Zheng, X.; Wang, S.; Wang, J.; Wang, L. A vision-language foundation model-based multi-modal retrieval-augmented generation framework for remote sensing lithological recognition. *ISPRS J. Photogramm. Remote Sens.* **2025**, *225*, 328–340. [CrossRef]
92. Siragusa, I.; Contino, S.; Ciura, M.L.; Alicata, R.; Pirrone, R. MedPix 2.0: A comprehensive multimodal biomedical data set for advanced AI applications with retrieval augmented generation and knowledge graphs. *Data Sci. Eng.* **2025**, *11*, 395–411. [CrossRef]
93. Fu, T.; Hu, J.; Min, G.; Khowaja, S.A.; Singh, K.; Dev, K. Federated Retrieval-Augmented Generation-Based LLM for Enhanced Cyber Threat Detection in the Internet-of-Energy. *IEEE Netw.* **2025**, *40*, 13–19. [CrossRef]
94. Zhou, Z.; Jiang, T.; Jin, Q. Integrating fine-tuning and retrieval-augmented generation to address the application challenges of pollution control guidelines. *J. Environ. Manag.* **2026**, *400*, 128718. [CrossRef] [PubMed]
95. de Jesus, D.D.R.; de Souza Júnior, A.P.; de Albergaria, E.T.; Pagano, A.S.; de Oliveira, I.J.R.; dos Santos Dias, C.; Lage, E.M.; de Oliveira, F.R.; Oliveira, J.A.; de Carvalho Gomes, I.; et al. Enhanced LLM-supported instructions for medication use through retrieval-augmented generation. *Comput. Biol. Med.* **2025**, *198*, 111135. [CrossRef] [PubMed]
96. Uhm, M.; Kim, J.; Ahn, S.; Jeong, H.; Kim, H. Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. *Autom. Constr.* **2025**, *170*, 105926. [CrossRef]
97. Song, M. Defining the problem: The impact of OCR quality on retrieval-augmented generation performance and strategies for improvement. *Inf. Process. Manag.* **2026**, *63*, 104368. [CrossRef]

98. Alabbasi, N.; Erak, O.; Alhussein, O.; Lotfi, I.; Muhaidat, S.; Debbah, M. Teleoracle: Fine-tuned retrieval-augmented generation with long-context support for networks. *IEEE Internet Things J.* **2025**, *12*, 13170–13182. [CrossRef]
99. Pesl, R.D.; Mathew, J.G.; Mecella, M.; Aiello, M. Retrieval-augmented generation for service discovery: Chunking strategies and benchmarking. *IEEE Trans. Serv. Comput.* **2026**, *19*, 1520–1534. [CrossRef]
100. Zhang, P.; Peng, X.; Han, R.; Chen, T.; Ma, J. Rag2Mol: Structure-based drug design based on retrieval augmented generation. *Briefings Bioinform.* **2025**, *26*, bbaf265. [CrossRef] [PubMed]
101. Wei, S.S.; Jhang, W.E.; Liu, Y.C.; Chuang, C.C.; Ou, Y.Y. RTK_RAG: Leveraging Retrieval Augmented Generation with Multi-Window Convolutional Neural Networks for Superior ATP Binding Site Prediction in Receptor Tyrosine Kinases. *J. Chem. Inf. Model.* **2025**, *65*, 7277–7284. [CrossRef] [PubMed]
102. Chen, G.; Yu, W.; Lu, X.; Zhang, X.; Meng, E.; Sha, L. Unlocking multi-view insights in knowledge-dense retrieval-augmented generation. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 4430–4439. [CrossRef]
103. Li, R.; Mao, S.; Zhu, C.; Yang, Y.; Tan, C.; Li, L.; Mu, X.; Liu, H.; Yang, Y. Enhancing Pulmonary Disease Prediction Using Large Language Models With Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study Based on Radiology Report. *J. Med. Internet Res.* **2025**, *27*, e72638. [CrossRef] [PubMed]
104. Wang, C.; Chai, S.; Xu, T.; Adil, M.; Qiu, T. CP-RAG: Mitigating Distracting Content in Retrieval-Augmented Generation for Industrial Knowledge Question Answering. *IEEE Internet Things J.* **2026**, *13*, 15056–15066. [CrossRef]
105. Kelly, P.; Schild, J.; Jafari, A. FolkRAG: A retrieval-augmented generation system for cultural heritage materials. *Neural Comput. Appl.* **2025**, *37*, 20281–20297. [CrossRef]
106. Li, Y.; Du, X.; Wang, Y.; Chen, X.; Zhou, Z.; Lian, J.; Chuang, Y.W.; Hong, P.; Hou, P.C.; Zhou, L. Ai-assisted literature screening: A hybrid approach using large language models and retrieval-augmented generation. *Int. J. Med. Inform.* **2025**, *207*, 106205. [CrossRef] [PubMed]
107. Bhayana, R.; Fawzy, A.; Deng, Y.; Bleakney, R.R.; Krishna, S. Retrieval-augmented generation for large language models in radiology: Another leap forward in board examination performance. *Radiology* **2024**, *313*, e241489. [CrossRef] [PubMed]
108. Pingua, B.; Sahoo, A.; Kandpal, M.; Murmu, D.; Rautaray, J.; Barik, R.K.; Saikia, M.J. Medical llms: Fine-tuning vs. retrieval-augmented generation. *Bioengineering* **2025**, *12*, 687. [CrossRef] [PubMed]
109. Wang, D.; Liang, J.; Ye, J.; Li, J.; Li, J.; Zhang, Q.; Hu, Q.; Pan, C.; Wang, D.; Liu, Z.; et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: Comparative study. *J. Med. Internet Res.* **2024**, *26*, e58041. [CrossRef] [PubMed]
110. Parameswaran, V.; Bernard, J.; Bernard, A.; Deo, N.; Tsung, S.; Lyytinen, K.; Sharp, C.; Rodriguez, F.; Maron, D.J.; Dash, R. Evaluating large language models and retrieval-augmented generation enhancement for delivering guideline-adherent nutrition information for cardiovascular disease prevention: Cross-sectional study. *J. Med. Internet Res.* **2025**, *27*, e78625. [CrossRef] [PubMed]
111. Liu, B.; Li, W.; Zhao, R.; Klauer, R.R.; Hansen, A.; Miller, N.; Chen, Y.; Blenner, M.A.; Tang, Y.J. Leveraging Retrieval-Augmented Generation to Accelerate Discoveries on Mealworm Larvae and Plastic Degradation. *Environ. Sci. Technol.* **2025**, *59*, 27437–27448. [CrossRef] [PubMed]
112. Bora, A.; Cuayáhuitl, H. Systematic analysis of retrieval-augmented generation-based LLMs for medical chatbot applications. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2355–2374. [CrossRef]
113. Lakatos, R.; Pollner, P.; Hajdu, A.; Joo, T. Investigating the performance of retrieval-augmented generation and domain-specific fine-tuning for the development of AI-driven knowledge-based systems. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 15. [CrossRef]
114. Perkins, G.; Anderson, N.W.; Spies, N.C. Retrieval-augmented generation salvages poor performance from large language models in answering microbiology-specific multiple-choice questions. *J. Clin. Microbiol.* **2025**, *63*, e01624-24. [CrossRef] [PubMed]
115. Rewthamrongsris, P.; Thongchotchat, V.; Burapacheep, J.; Trachoo, V.; Khurshid, Z.; Porntaveetus, T. Evaluating Retrieval-Augmented Generation-Large Language Models for Infective Endocarditis Prophylaxis: Clinical Accuracy and Efficiency. *Int. Dent. J.* **2026**, *76*, 109344. [CrossRef] [PubMed]
116. Lee, J.; Ahn, S.; Kim, D.; Kim, D. Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. *Autom. Constr.* **2024**, *168*, 105846. [CrossRef]
117. Xu, R.; Hong, Y.; Zhang, F.; Xu, H. Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. *Sci. Rep.* **2024**, *14*, 30794. [CrossRef] [PubMed]
118. Elmitwalli, S.; Mehegan, J.; Braznell, S.; Gallagher, A. Scalable evaluation framework for retrieval augmented generation in tobacco research using large Language models. *Sci. Rep.* **2025**, *15*, 22760. [CrossRef] [PubMed]
119. Katsis, Y.; Rosenthal, S.; Fadnis, K.; Gunasekara, C.; Lee, Y.S.; Popa, L.; Shah, V.; Zhu, H.; Contractor, D.; Danilevsky, M. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Trans. Assoc. Comput. Linguist.* **2025**, *13*, 784–808. [CrossRef]
120. Gao, Y.; Xiong, Y.; Wu, W.; Li, B.; Zhong, Y.; Wang, H. U-NIAH: Unified RAG and LLM Evaluation for Long Context Needle-in-a-Haystack. *ACM Trans. Inf. Syst.* **2026**, *44*, 1–30. [CrossRef]

121. Zhao, S.; Shao, Y.; Huang, Y.; Song, J.; Wang, Z.; Wan, C.; Ma, L. Understanding the Fundamental Design Decisions of Retrieval-Augmented Generation Systems. *ACM Trans. Softw. Eng. Methodol.* **2026**. [CrossRef]
122. Jiang, P.; Han, J.; Ahmed-Kristensen, S. A two-stage retrieval-augmented generation framework for producing sustainable product design guidelines. *Sustain. Futur.* **2025**, *10*, 101094. [CrossRef]
123. Cheng, Y.; Zhao, H.; Xiang, D.; Zhang, Z.; Liu, G.; Liu, Y.; Zhao, J.; Cai, X. Power system operational reliability evaluation with retrieval-augmented generation enhanced large language model. *Energy AI* **2026**, *24*, 100688. [CrossRef]
124. Cui, L.; Liu, Y.; Ouyang, C.; Yu, Y.; Zhang, J.; Wan, Y.; Yang, F. Bailicai: A domain-optimized retrieval-augmented generation framework for medical applications. *Big Data Min. Anal.* **2026**, *9*, 376–392. [CrossRef]
125. Chen, H.K.; Wang, H.H. An Innovative Retrieval-Augmented Generation Framework for Stage-Specific Knowledge Translation in Biomimicry Design. *Biomimetics* **2025**, *10*, 626. [CrossRef] [PubMed]
126. Zhang, Z.; Han, Y.; Li, R.; Li, X. HiMRAG: Hierarchical Multimodal Retrieval-Augmented Generation for Robot Task Planning. *IEEE Robot. Autom. Lett.* **2026**, *11*, 3883–3890. [CrossRef]
127. Zhang, R.; Du, H.; Liu, Y.; Niyato, D.; Kang, J.; Sun, S.; Shen, X.; Poor, H.V. Interactive AI with retrieval-augmented generation for next generation networking. *IEEE Netw.* **2024**, *38*, 414–424. [CrossRef]
128. Mo, F.; Gao, Y.; Wu, Z.; Liu, X.; Chen, P.; Li, Z.; Wang, Z.; Li, X.; Jiang, M.; Nie, J.Y. Leveraging historical information to boost retrieval-augmented generation in conversations. *Inf. Process. Manag.* **2026**, *63*, 104449. [CrossRef]
129. Fu, Y.; Liu, D.; Zhang, B.; Jiang, Z.; Mei, H.; Guan, J. Cue RAG: Dynamic multi-output cue memory under H framework for retrieval-augmented generation. *Neurocomputing* **2025**, *639*, 130235. [CrossRef]
130. Al-Maliki, S.; Gharaibeh, A.; Rahouti, M.; Amin, M.R.; Abdallah, M.; Qadir, J.; Al-Fuqaha, A. Budget-Constrained Online Retrieval-Augmented Generation: The Chunk-as-a-Service Model. *IEEE Trans. Artif. Intell.* **2026**, *7*, 4822–4836. [CrossRef]
131. Das, S.; Ge, Y.; Guo, Y.; Rajwal, S.; Hairston, J.; Powell, J.; Walker, D.; Peddireddy, S.; Lakamana, S.; Bozkurt, S.; et al. Two-layer retrieval-augmented generation framework for low-resource medical question answering using reddit data: Proof-of-concept study. *J. Med. Internet Res.* **2025**, *27*, e66220. [CrossRef] [PubMed]
132. Chen, Z.; Hao, J.; Sun, H.; Zhao, L.; Li, J.; Qian, Q.; Peng, Q.; Wang, X.; Cong, S.; Shen, L.; et al. MedScaleRE-PF: A prompt-based framework with retrieval-augmented generation, chain-of-thought, and self-verification for scale-specific relation extraction in Chinese medical literature. *Inf. Process. Manag.* **2025**, *62*, 104278. [CrossRef]
133. Li, Y.; Ke, W.; Liu, J.; Wang, P.; Liu, J.; He, Y. Towards evidence-aware retrieval-augmented generation via self-corrective chain-of-thought. *Inf. Process. Manag.* **2026**, *63*, 104369. [CrossRef]
134. Zhou, M.; Tang, J.; Zeng, W.; Zhao, X. INKER: Adaptive dynamic retrieval augmented generation with internal-external knowledge integration. *Inf. Process. Manag.* **2026**, *63*, 104534. [CrossRef]
135. Fotopoulou, E.; Zafeiropoulos, A.; Arkadopoulou, E.; Androna, C.M.; Mandilara, I.; Papavassiliou, S. Expert Agents for Social and Emotional Learning: An Agentic Graph Retrieval Augmented Generation approach. *IEEE Access* **2026**, *14*, 37565–37581. [CrossRef]
136. Jahanbakhsh, N.; Vega-Barbas, M.; Pau, I.; Elvira-Martín, L.; Moosavi, H.; García-Vázquez, C. Leveraging Retrieval-Augmented Generation for Automated Smart Home Orchestration. *Future Internet* **2025**, *17*, 198. [CrossRef]
137. Krotkov, N.A.; Sbytov, D.A.; Chakhoyan, A.A.; Kornienko, P.I.; Starikova, A.A.; Stepanov, M.G.; Piven, A.O.; Aliev, T.A.; Orlova, T.; Rafayelyan, M.S.; et al. Nanostructured material design via a retrieval-augmented generation (rag) approach: Bridging laboratory practice and scientific literature. *J. Chem. Inf. Model.* **2025**, *65*, 11064–11078. [CrossRef] [PubMed]
138. Foundjem, A.; Nganyewou Tidjon, L.; Da Silva, L.; Khomh, F. Multi-Agent AI Framework for Threat Mitigation and Resilience in Machine Learning Systems. *ACM Trans. Softw. Eng. Methodol.* **2026**. [CrossRef]
139. Ludwig, H.; Schmidt, T.; Kühn, M. An ontology-based retrieval augmented generation procedure for a voice-controlled maintenance assistant. *Comput. Ind.* **2025**, *169*, 104289. [CrossRef]
140. Takahara, W.; Yamaguchi, Y.; Ogano, M.; Kakami, F.; Harashima, Y.; Takayama, T.; Takasuka, S.; Kudo, A.; Fujii, M. Materials Dual-Source Knowledge Retrieval-Augmented Generation for Local Large Language Models in Photocatalysts. *J. Chem. Inf. Model.* **2025**, *65*, 13098–13114. [CrossRef] [PubMed]
141. Zhang, J.; Zhang, W.; Wei, Q.; Zhong, H.; Miao, Y. GRAG-ProSafe QAS: Graph retrieval-augmented generation for process production safety intelligent question and answer system. *Expert Syst. Appl.* **2025**, *298*, 129947. [CrossRef]
142. Cui, H.; Zheng, P.; Ren, M.; Yan, Y. An LLM-based cross-domain knowledge retrieval augmented generation method for bio-inspired solution design. *Adv. Eng. Inform.* **2026**, *69*, 104017. [CrossRef]
143. Liang, J.; Hou, S.; Jiao, H.; Qing, Y.; Zhao, A.; Shen, Z.; Xiang, L.; Wu, H. GeoGraphRAG: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104712. [CrossRef]
144. Bahr, L.; Wehner, C.; Wewerka, J.; Bittencourt, J.; Schmid, U.; Daub, R. Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. *J. Ind. Inf. Integr.* **2025**, *45*, 100807. [CrossRef]

145. Niederhaus, M.; Migenda, N.; Weller, J.; Kohlhase, M.; Schenck, W. Integrating Graph Retrieval-Augmented Generation into Prescriptive Recommender Systems. *Big Data Cogn. Comput.* **2025**, *9*, 261. [CrossRef]
146. Tsampos, I.; Marakakis, E. DietQA: A Comprehensive Framework for Personalized Multi-Diet Recipe Retrieval Using Knowledge Graphs, Retrieval-Augmented Generation, and Large Language Models. *Computers* **2025**, *14*, 412. [CrossRef]
147. Yuan, D.; Zhou, H.; Liu, X.; Chen, H.; Xin, Y.; Zhang, J. Enhancing Large Language Models (LLMs) for Telecom Using Dynamic Knowledge Graphs and Explainable Retrieval-Augmented Generation. *IEEE Wirel. Commun.* **2026**, *33*, 284–292. [CrossRef]
148. Ghosh, S.; Mittal, G. Advancing engineering research through context-aware and knowledge graph–based retrieval-augmented generation. *Front. Artif. Intell.* **2025**, *8*, 1697169. [CrossRef] [PubMed]
149. Yao, L.; Ren, F.; Du, K.; Du, Q. From knowledge graph construction to retrieval-augmented generation: A framework for comprehensive earthquake emergency support. *Geo-Spat. Inf. Sci.* **2026**, *29*, 509–529. [CrossRef]
150. Yang, W.C.; Li, X.; Chang, C.Y.; Sinha Roy, D. Hierarchical knowledge graph-based QA systems with retrieval-augmented generation. *Enterp. Inf. Syst.* **2026**, *20*, 2580477. [CrossRef]
151. Wang, Q.; Wu, B.; Xu, M.; Liu, X.; Mao, Y.; Zhou, Z.; Yan, G. HyperSynergyX: Synergistic Drug Combination Prediction via Hypergraph Modeling and Knowledge Graph-Enhanced Retrieval-Augmented Generation. *IEEE J. Biomed. Health Inform.* **2026**. [CrossRef] [PubMed]
152. Hou, Y.; Bishop, J.R.; Liu, H.; Zhang, R. Improving dietary supplement information retrieval: Development of a retrieval-augmented generation system with large language models. *J. Med. Internet Res.* **2025**, *27*, e67677. [CrossRef] [PubMed]
153. Shen, Y.; Xu, D.; Li, X.; Fu, L.; Liu, H. MGPRAG: Enhancing Medical Large Language Models via Precision Retrieval-Augmented Generation. *IEEE Access* **2025**, *13*, 183225–183241. [CrossRef]
154. Liu, S.; Wright, A.P.; McCoy, A.B.; Huang, S.S.; Steitz, B.; Wright, A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1032–1039. [CrossRef] [PubMed]
155. Xu, L.; Lu, L.; Liu, M.; Song, C.; Wu, L. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Herit. Sci.* **2024**, *12*, 118. [CrossRef]
156. Wang, S.; Yang, H.; Bai, G. Construction of intelligent decision support systems through integration of retrieval-augmented generation and knowledge graphs. *Sci. Rep.* **2025**, *15*, 35462. [CrossRef] [PubMed]
157. Xiong, Y.; Zhang, R.; Liu, Y.; Niyato, D.; Xiong, Z.; Liang, Y.C.; Mao, S. When knowledge graph meets retrieval augmented generation for wireless networks: A tutorial and case study. *IEEE Wirel. Commun.* **2025**, *33*, 134–142. [CrossRef]
158. Du, X.; Zheng, G.; Wang, K.; Zou, Y.; Wang, Y.; Deng, W.; Feng, J.; Liu, M.; Chen, B.; Peng, X.; et al. Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG. *ACM Trans. Softw. Eng. Methodol.* **2026**, *35*, 1–25. [CrossRef]
159. Lu, Z.; Xu, H.; Chen, A.; Tang, S.; Zhang, J.; Feng, Y.; Pan, W.; Huang, J. HSG-RAG: Hierarchical Knowledge Base Construction for Embedded System Development. *ACM Trans. Des. Autom. Electron. Syst.* **2025**, *30*, 1–21. [CrossRef]
160. Zhang, W.; Dai, G.; Luo, D.; Wang, Y.; Ye, C. From Hallucination to Certainty: Meta-Knowledge Guided Self-Correcting Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2026**, *17*, 1–31. [CrossRef]
161. Shi, D.; Li, J.; Meyer, O.; Bauernhansl, T. Enhancing retrieval-augmented generation for interoperable industrial knowledge representation and inference toward cognitive digital twins. *Comput. Ind.* **2025**, *171*, 104330. [CrossRef]
162. Garcia, B.T.; Westerfield, L.; Yelemali, P.; Gogate, N.; Rivera-Munoz, E.A.; Du, H.; Dawood, M.; Jolly, A.; Lupski, J.R.; Posey, J.E. Improving automated deep phenotyping through large language models using retrieval-augmented generation. *Genome Med.* **2025**, *17*, 91. [CrossRef] [PubMed]
163. Shang, Y.; Ke, Z.; Lin, P.; Ren, Q.; Zhang, W.; Wang, X.; Li, X.; Gong, F.; Wang, S.; Wang, B.; et al. Empowering knowledge graphs with hybrid retrieval-augmented generation for the intelligent mix scheme of mass concrete. *Case Stud. Constr. Mater.* **2025**, *23*, e04979. [CrossRef]
164. Yang, J.S.; Zeng, Z.; Shen, Z. Neural-Symbolic Dual-Indexing Architectures for Scalable Retrieval-Augmented Generation. *IEEE Access* **2025**, *13*, 210507–210519. [CrossRef]
165. Wang, W.; Wang, P.; Chen, Y.; Zhang, X.; Song, P.; Cherkasov, O.; Fan, X. RAGLRO: Retrieval-Augmented Generation With Large Language Models for Robotic Operations. *CAAI Trans. Intell. Technol.* **2026**, *11*, 385–395. [CrossRef]
166. Wang, Z.; Wang, H.; Xiong, G.; Lin, R. FC-RAG: Enhancing Football Coaching with Multimodal Retrieval-Augmented Generation. *IEEE Trans. Learn. Technol.* **2026**, *19*, 117–126. [CrossRef]
167. Lahiri, A.K.; Hu, Q.V. AlzheimerRAG: Multimodal retrieval-augmented generation for clinical use cases. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 89. [CrossRef]
168. Olca, E.A. Professor X: Diagnosis and Treatment of Dermatological Diseases by Integration of Visual Diagnosis and Retrieval-Augmented Generation (RAG) Technologies. *IEEE Access* **2025**, *13*, 201246–201263. [CrossRef]
169. Kunekar, P.; Mankar, S.; Cholke, P.; Kulkarni, A.; Nooji, P.; Gadhave, R. MedIntelliCare: Neurodynamic-inspired AI for medical decision support by integrating retrieval-augmented generation with multimodal cognitive processing. *Cogn. Neurodynamics* **2026**, *20*, 61. [CrossRef] [PubMed]

170. Xu, Z.; Zhang, X.; Yang, Y.; Chen, W.; Liu, J.; Xu, T.; Wang, Z.; El Saddik, A. MuralAgent: Enhancing Ancient Mural Outpainting with RAG-Based Texts and Multimodal Integration. *ACM Trans. Multimed. Comput. Commun. Appl.* **2025**, *21*, 1–17. [CrossRef]
171. Jiang, A.; Wang, D.; Peng, C.; Wang, M. Relational Reasoning Image Captioning via Multi-Agent Retrieval-Augmented Generation. *Knowl.-Based Syst.* **2025**, *333*, 114977. [CrossRef]
172. Uhm, M.; Kim, J.; Lee, G. Automated analysis of construction safety accident videos using a large multimodal model and graph retrieval-augmented generation. *Autom. Constr.* **2025**, *177*, 106363. [CrossRef]
173. Roll, D.S.; Kurt, Z.; Li, Y.; Woo, W.L. Augmenting orbital debris identification with neo4j-enabled graph-based retrieval-augmented generation for multimodal large language models. *Sensors* **2025**, *25*, 3352. [CrossRef] [PubMed]
174. Li, Y.; Dong, Z.; Ameri, F.; Zhang, J. ManuRAG: Multi-modal retrieval augmented generation for manufacturing question answering. *J. Intell. Manuf.* **2026**. [CrossRef]
175. Li, X.; Zheng, Z.; Zhao, M.; Zhao, Y.; Shi, L.; Wang, B. RLFE-IDS: A framework of intrusion detection system based on retrieval augmented generation and large language model. *Comput. Netw.* **2025**, *268*, 111341. [CrossRef]
176. Zhao, T.; Chen, J.; Ru, Y.; Zhu, H.; Hu, N.; Liu, J.; Lin, Q. Exploring knowledge poisoning attacks to retrieval-augmented generation. *Inf. Fusion* **2025**, 103900. [CrossRef]
177. Afiffy, M.; Fakhr, M.W.; Maghraby, F.A. Enhancing adversarial resilience in semantic caching for secure retrieval augmented generation systems. *Sci. Rep.* **2026**, *16*, 5936. [CrossRef] [PubMed]
178. Wang, S.; Zhu, T.; Ye, D.; Zhou, W. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *IEEE Trans. Dependable Secur. Comput.* **2025**, *23*, 2030–2045. [CrossRef]
179. He, L.; Zhang, X.; Liu, Q.; Li, C.; Liu, Z.; Zhou, P.; Su, S. Push and Pull: Defending against Retrieval Poisoning Attacks via Embedding Space Reshaping. *ACM Trans. Inf. Syst.* **2026**, *44*, 1–31. [CrossRef]
180. Ning, L.; Fan, W.; Li, Q. Retrieval-Augmented Purifier for Robust LLM-Empowered Recommendation. *ACM Trans. Inf. Syst.* **2026**, *44*, 1–42. [CrossRef]
181. Macieira, T.G.; Bjarnadottir, R.I.; de Oliveira Salgado, P.; Baranwal, A.; Semenov, A.; Priola, K.B.; Pestana, P.; Okafor, S.; Mena, N.; Montoya, N.; et al. Making nursing visible: AI-assisted standardization of electronic health record interventions using generative pre-trained transformer models and retrieval-augmented generation process. *Nurs. Outlook* **2025**, *73*, 102494. [CrossRef] [PubMed]
182. Shang, D.; Duan, Y.; Bai, P.; Xie, J. Exploring retrieval-augmented generation for multi-label discipline classification of academic short texts. *Scientometrics* **2025**, *131*, 2373–2399. [CrossRef]
183. Tan, T.; Chen, X. DeepDesc: Integrating retrieval-augmented generation with large language models for smart contract vulnerability detection. *Empir. Softw. Eng.* **2026**, *31*, 32. [CrossRef]
184. Tan, G.; He, Z.; Li, M.; Huang, H.; Wang, Y.; Chen, P.; Casale, G.; Zhang, C. LLMRCA: Multilevel Root Cause Analysis for LLM Applications Using Multimodal Observability Data. *ACM Trans. Softw. Eng. Methodol.* **2026**. [CrossRef] [PubMed]
185. Alkhalaf, M.; Yu, P.; Yin, M.; Deng, C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J. Biomed. Inform.* **2024**, *156*, 104662. [CrossRef] [PubMed]
186. Bala, I.; Sharma, E.; Mitchell, L. Clinical insight from mesh implant narratives using zero-shot Retrieval-Augmented Generation approach. *Artif. Intell. Surg.* **2025**, *5*, 476–489. [CrossRef]
187. Abdaoui, H.; Barki, C.; Dergaa, I.; Tlili, K.; Ceylan, H.İ.; Bragazzi, N.L.; de Giorgio, A.; Ben Salah, R.; Rahmouni, H.B. Accurate Clinical Entity Recognition and Code Mapping of Anatomopathological Reports Using BioClinicalBERT Enhanced by Retrieval-Augmented Generation: A Hybrid Deep Learning Approach. *Bioengineering* **2025**, *13*, 30. [CrossRef] [PubMed]
188. Ren, T.; Zhang, Z.; Jia, B.; Zhang, S. Retrieval-Augmented Generation-aided causal identification of aviation accidents: A large language model methodology. *Expert Syst. Appl.* **2025**, *278*, 127306. [CrossRef]
189. Wei, Y.; Huang, L.; Feng, Q.; Chen, Z.; Yan, J.; Lin, T.J.; Huang, Z.; Ren, K.; Xing, W.; He, L. ModelGen: Automating Semiconductor Parameter Extraction with Large Language Model Agents. *ACM Trans. Des. Autom. Electron. Syst.* **2025**, *30*, 1–26. [CrossRef]
190. Ma, Z.; Kim, D.J.; Chen, T.H.P. Unsupervised, Accurate, and Efficient Log Parsing Using Smaller Open-Source Large Language Models. *ACM Trans. Softw. Eng. Methodol.* **2026**, *35*, 1–28. [CrossRef]
191. Sisomboon, W.; Kaewyotha, J.; Songpan, W. Automated Software Test Case Generation Using Directional Partially Weighted Ensemble Large Language Models with Retrieval-Augmented Generation (RAG). *IEEE Access* **2026**, *14*, 31124–31145. [CrossRef]
192. Yao, X.; Li, H.; Chan, T.H.; Xiao, W.; Yuan, M.; Huang, Y.; Chen, L.; Yu, B. HDLdebugger: Streamlining HDL debugging with Large Language Models. *ACM Trans. Des. Autom. Electron. Syst.* **2025**, *30*, 1–26. [CrossRef]
193. Qayyum, K.; Jha, C.K.; Ahmadi-Pour, S.; Hassan, M.; Drechsler, R. LLM-assisted Bug Identification and Correction for Verilog HDL. *ACM Trans. Des. Autom. Electron. Syst.* **2025**, *30*, 1–28. [CrossRef]
194. Daneshvar, S.S.; Nong, Y.; Yang, X.; Wang, S.; Cai, H. VulScribeR: Exploring RAG-based Vulnerability Augmentation with LLMs. *ACM Trans. Softw. Eng. Methodol.* **2026**, *35*, 1–26. [CrossRef]
195. Li, M.; Hu, Z.; Mohebi, P.; Li, S.; Wang, Z. Enhancing LLM-based building data query with chain-of-thought, retrieval-augmented generation, and fine-tuning. *Autom. Constr.* **2026**, *182*, 106738. [CrossRef]

196. Jeon, J.; Sim, Y.; Lee, H.; Han, C.; Yun, D.; Kim, E.; Nagendra, S.L.; Jun, M.B.; Kim, Y.; Lee, S.W.; et al. ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. *J. Manuf. Syst.* **2025**, *79*, 504–514. [CrossRef]
197. Qin, S.; Wang, Y.; Cui, T.; Ma, J.; Zhou, X.; Guo, X.; Zhang, C.; Zhou, C.; Rongjuan, G.; Li, H. Intelligent Chinese patent medicine (CPM) recommendation framework: Integrating large language models, retrieval-augmented generation, and the largest CPM dataset. *Pharmacol. Res.* **2025**, *219*, 107883. [CrossRef] [PubMed]
198. Zhang, C.; Yang, H.; Liu, X.; Wu, R.; Zong, H.; Wu, E.; Zhou, Y.; Li, J.; Shen, B. A Knowledge-Enhanced Platform (MetaSepsisKnowHub) for Retrieval Augmented Generation–Based Sepsis Heterogeneity and Personalized Management: Development Study. *J. Med. Internet Res.* **2025**, *27*, e67201. [CrossRef] [PubMed]
199. Tung, J.Y.M.; Le, Q.; Yao, J.; Huang, Y.; Lim, D.Y.Z.; Sng, G.G.R.; Lau, R.S.E.; Tan, Y.G.; Chen, K.; Tay, K.J.; et al. Performance of Retrieval-Augmented Generation Large Language Models in Guideline-Concordant Prostate-Specific Antigen Testing: Comparative Study With Junior Clinicians. *J. Med. Internet Res.* **2025**, *27*, e78393. [CrossRef] [PubMed]
200. Kunstmann, H.; Ollier, J.; Persson, J.; von Wangenheim, F. EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context. *ACM Trans. Recomm. Syst.* **2026**. [CrossRef]
201. Wei, C.; Duan, K.; Zhuo, S.; Wang, H.; Huang, S.; Liu, J. Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model. *J. Artif. Int. Res.* **2025**, *82*, 1147–1173. [CrossRef]
202. Jiang, Z.; Zhao, R.; Lin, L.; Yu, Y.; Chen, H.; Zhang, X.; Xu, X.; Wang, Y.; Ma, X.; Ngai, E.C.H. DietGlance: Dietary Monitoring and Personalized Analysis at a Glance with Knowledge-Empowered AI Assistant. *ACM Trans. Comput. Healthc.* **2026**, *7*, 1–46. [CrossRef]
203. Singh, A.; Rousta, M.M.; Huang, B. Oscillation analysis using Retrieval-Augmented Generation. *Comput. Chem. Eng.* **2025**, *206*, 109489. [CrossRef]
204. Tang, Y.; Guo, W. Automatic retrieval-augmented generation of 6G network specifications for use cases. *IEEE Commun. Mag.* **2024**, *63*, 95–102. [CrossRef]
205. Cao, Z.; Shen, L.; Xu, X.; Guo, Y.; Pan, B.; Bao, H. A Retrieval-Augmented Generation (RAG) Based Framework for Evaluating Urban Low-Carbon Governance and Its Implications for Sustainable Development. *Sustainability* **2026**, *18*, 2143. [CrossRef]
206. Loumachi, F.Y.; Ghanem, M.C.; Ferrag, M.A. Advancing cyber incident timeline analysis through retrieval-augmented generation and large language models. *Computers* **2025**, *14*, 67. [CrossRef]
207. Praneeth, B.; Mohana; Nattem, E.C.; Jetti, K.; Kavyashree, B.K.; Rakshitha, D.; Kumar, P.R.; Sreelakshmi, K. Optimization of Customer Feedback Summarization Using Large Language Models (LLM) and Advanced Retrieval-Augmented Generation. *IEEE Access* **2025**, *13*, 124319–124332. [CrossRef]
208. Haider, S.A.; Prabha, S.; Gomez Cabello, C.A.; Genovese, A.; Collaco, B.; Wood, N.; London, J.; Bagaria, S.; Tao, C.; Forte, A.J. The Development and Evaluation of a Retrieval-Augmented Generation Large Language Model Virtual Assistant for Postoperative Instructions. *Bioengineering* **2025**, *12*, 1219. [CrossRef] [PubMed]
209. Donner, S.; Knauer, P.; Kienzle, A.; Dinneen, J.; Burger, J.; Perka, C.; Donner, S. Exploring Patient Perspectives, Engagement, and Output Quality in Doctor-Supervised Use of Artificial Intelligence During Informed Consent Consultation With ChatGPT and Retrieval Augmented Generation (RAG): Quantitative Exploratory Study. *J. Med. Internet Res.* **2025**, *27*, e73717. [CrossRef] [PubMed]
210. Wang, Z.; Tsang, T.H.; Yu, F.Y.; Fang, Y.; Chen, S.; Sun, F.; Mo, P.K.; Wong, K.C. Effectiveness, Usability, and Acceptability of ChatGPT With Retrieval-Augmented Generation (SIV-ChatGPT) in Increasing Seasonal Influenza Vaccination Uptake Among Older Adults: Quasi-Experimental Study. *J. Med. Internet Res.* **2025**, *27*, e76849. [CrossRef] [PubMed]
211. Neupane, S.; Mitra, S.; Mittal, S.; Gaur, M.; Golilarz, N.A.; Rahimi, S.; Amirlatifi, A. MedInsight: A Multi-Source Context Augmentation Framework for Generating Patient-Centric Medical Responses Using Large Language Models. *ACM Trans. Comput. Healthc.* **2025**, *6*, 1–19. [CrossRef]
212. Arazzi, M.; Marconi Sciarroni, M.; Nocera, A.; Storti, E. RAG-IoE: IoT context-aware information retrieval with Large Language Models in Industry 5.0. *ACM Trans. Internet Things* **2025**, *6*, 1–31. [CrossRef]
213. Wu, C.; Ding, W.; Jin, Q.; Jiang, J.; Jiang, R.; Xiao, Q.; Liao, L.; Li, X. Retrieval augmented generation-driven information retrieval and question answering in construction management. *Adv. Eng. Inform.* **2025**, *65*, 103158. [CrossRef]
214. Jiang, Z.; Chen, G.; Xu, Z. Building regulation question-answering system using retrieval-augmented generation with dual-stage fine-tuned large language model. *Adv. Eng. Inform.* **2026**, *69*, 104089. [CrossRef]
215. Han, N.; Liu, D.; Norton, T. Towards an AI-based knowledge assistant for goat farmers based on retrieval-augmented generation. *Comput. Electron. Agric.* **2026**, *245*, 111542. [CrossRef]
216. Lee, Y. Developing a computer-based tutor utilizing generative artificial intelligence (GAI) and retrieval-augmented generation (RAG). *Educ. Inf. Technol.* **2025**, *30*, 7841–7862. [CrossRef]

217. Rugemalila, I.E.; Cai, W.; Zhang, X.; Liu, C.; Wang, B. Curriculum-Aware Retrieval-Augmented Generation for Bilingual Tutoring in Low-Resource Swahili–English Secondary Schools. *Technologies* **2026**, *14*, 129. [CrossRef]
218. Hu, R.; Liu, S.; Qi, P.; Liu, J.; Li, F. ICCA-RAG: Intelligent customs clearance assistant using retrieval-augmented generation (RAG). *IEEE Access* **2025**, *13*, 39711–39726. [CrossRef]
219. Ismail; Kurnia, R.; Widyatama, F.; Wibawa, I.M.; Brata, Z.A.; Ukasyah; Nelistiani, G.A.; Kim, H. Enhancing Security Operations Center: Wazuh Security Event Response with Retrieval-Augmented-Generation-Driven Copilot. *Sensors* **2025**, *25*, 870. [CrossRef] [PubMed]
220. Sprenkamp, K.; Eckhardt, S.; Zavolokina, L.; Schwabe, G. From Information-Seeking to Information-Asking: Designing RefuGPT, a Chatbot for Ukrainian Refugees in Switzerland. *Digit. Gov. Res. Pract.* **2025**, *6*, 1–21. [CrossRef]
221. Feng, Y.; Wang, Y.; Wang, X.; Zhao, B.; Bi, J.; Han, S.; Luo, Y. Food additive lens: An on-device AI application for real-time science-based consumer education on food additives using retrieval-augmented generation. *Digit. Discov.* **2026**, *5*, 1172–1190. [CrossRef]
222. Zeeshan, H.M.A.; Umer, M.; Akbar, M.; Kaushik, A.; Jamshed, M.A.; Jung, H.; Hassan, S.A. LLM-based retrieval-augmented generation: A novel framework for resource optimization in 6G and beyond wireless networks. *IEEE Commun. Mag.* **2025**, *63*, 60–67. [CrossRef]
223. Tang, S.; Zhang, R.; Yan, Y.; Yang, Q.; Niyato, D.; Wang, X.; Mao, S. Retrieval-augmented generation for GenAI-enabled semantic communications. *IEEE Wirel. Commun.* **2025**, *33*, 259–268. [CrossRef]
224. Fazlollahtabar, H. Human-robot interaction using retrieval-augmented generation and fine-tuning with transformer neural networks in industry 5.0. *Sci. Rep.* **2025**, *15*, 29233. [CrossRef] [PubMed]
225. Zhou, X.; Guan, X.; Wang, N.; Chen, H.; Ohtsuki, T.; Zhang, Y. Adaptive Task Offloading for Edge Computing in Internet of Energy via Retrieval-Augmented Generation. *IEEE Netw.* **2026**, *40*, 219–228. [CrossRef]
226. Chamotra, S.; Barbhuiya, F. SAGE: An Adaptive IoT Honeypot with FSM-Driven Protocol Emulation and GraphRAG-Powered Response Generation. *ACM Trans. Internet Things* **2026**, *7*, 1–34. [CrossRef]
227. Yang, R.; Fu, M.; Tantithamthavorn, C.; Arora, C.; Vandenhurk, L.; Chua, J. RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice. *J. Syst. Softw.* **2025**, *226*, 112436. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.