



RETRIEVAL-AUGMENTED GENERATION FOR SCIENTIFIC KNOWLEDGE DISCOVERY: IMPROVING FACTUALITY AND CITATION RELIABILITY

Dr. S. Muthulakshmi, Dr. P. Devi

PG & Research Department of Computer Science, Thiagarajar College, Madurai – 625 009

ABSTRACT

The rapid growth of scientific publications has created a substantial challenge for researchers seeking to identify, synthesize, and verify relevant knowledge. Large Language Models (LLMs) provide powerful capabilities for scientific question answering and literature synthesis, but their reliance on parametric knowledge can result in hallucinated facts, outdated information, incomplete evidence, and unreliable citations. Retrieval-Augmented Generation (RAG) offers a promising solution by connecting generative models to external scientific literature during inference. However, conventional RAG systems may still retrieve irrelevant passages, overlook contradictory evidence, generate unsupported claims, or provide citations that do not adequately support the associated statements. This paper proposes a reliability-oriented RAG framework for scientific knowledge discovery that combines domain-specific retrieval, hybrid semantic and lexical search, cross-encoder reranking, evidence-aware generation, citation verification, and post-generation self-reflection. The framework treats factuality and citation reliability as separate but interconnected objectives. A structured evaluation methodology is proposed using retrieval metrics, answer faithfulness, factual correctness, citation precision, citation recall, citation completeness, and citation faithfulness. Recent research on scientific literature synthesis demonstrates that retrieval-based systems can substantially improve scientific answer quality and citation accuracy compared with language-model-only approaches. The proposed framework provides a practical architecture for developing transparent, evidence-grounded scientific assistants capable of supporting literature review, hypothesis exploration, methodological comparison, and knowledge discovery while reducing hallucination and citation errors.

Keywords: Retrieval-Augmented Generation, Scientific Knowledge Discovery, Large Language Models, Factuality, Citation Reliability, Literature Mining, Explainable AI

INTRODUCTION

Scientific knowledge is expanding at an unprecedented rate. Researchers must continuously examine journal articles, conference papers, preprints, datasets, technical reports, and supplementary information to understand developments within their fields. Conventional literature-review processes depend heavily on keyword searches, manual screening, citation chaining, and researcher expertise. Although these approaches remain essential, the scale and speed of scientific publication make comprehensive knowledge synthesis increasingly difficult.

Large Language Models (LLMs) have introduced a new paradigm for interacting with scientific knowledge. Instead of requiring researchers to manually locate and read multiple documents, an LLM can summarize concepts, compare studies, answer technical questions, and generate literature-based explanations. However, an LLM's internal knowledge is primarily parametric: information is encoded within model parameters during training rather than retrieved directly from current scientific sources. Consequently,



models may generate plausible but incorrect statements, rely on outdated information, or produce citations that do not exist or do not support the associated claims.

Retrieval-Augmented Generation (RAG) addresses part of this problem by retrieving external information and providing it to the language model as contextual evidence. The basic RAG paradigm combines a retriever with a generative model, allowing the system to access information that is not necessarily contained in the model's parameters. Survey research has identified RAG as an important strategy for addressing hallucination and knowledge-staleness problems in LLMs.

Scientific applications require a more stringent form of RAG than general question answering. A scientific answer should not merely be plausible. It should be supported by appropriate evidence, traceable to legitimate sources, and sufficiently precise to distinguish between established findings, interpretations, and hypotheses. Recent scientific literature synthesis research demonstrates the magnitude of this challenge. The OpenScholar study reported substantial citation hallucination by language-model-only systems and showed that specialized retrieval, reranking, self-feedback, and citation verification can substantially improve scientific answer quality.

Therefore, this study focuses on the following research problem:

How can Retrieval-Augmented Generation be designed to improve both factuality and citation reliability in scientific knowledge discovery?

The primary objectives are:

1. To analyze the limitations of conventional LLM-based scientific knowledge discovery.
2. To examine the role of RAG in improving factual grounding.
3. To develop a reliability-oriented architecture combining retrieval, reranking, generation, and citation verification.
4. To define an evaluation methodology for measuring factuality and citation reliability.
5. To identify remaining challenges and future research directions for scientific RAG systems.

LITERATURE REVIEW

2.1 Large Language Models and Scientific Knowledge

LLMs have demonstrated strong capabilities in natural-language understanding, summarization, reasoning, and question answering. However, their knowledge is constrained by the information contained in training data and by the mechanisms through which that knowledge is represented.

For scientific applications, three limitations are particularly important.

First, **knowledge freshness** is problematic. Scientific discoveries, datasets, methods, and corrections are continuously published. A model's static parameters cannot automatically incorporate newly published research.

Second, **hallucination** remains a major concern. A model may generate scientifically plausible statements that are unsupported by evidence.



Third, **citation reliability** is not guaranteed. Generating a reference-looking citation does not ensure that the cited paper exists or that it supports the associated claim.

These limitations motivate the integration of external scientific retrieval mechanisms.

2.2 Retrieval-Augmented Generation

RAG connects a retrieval system with a generative language model. Instead of relying exclusively on the model's internal parameters, the system retrieves relevant external passages and conditions generation on those passages.

A simplified formulation is:

$$R(q) = \{d_1, d_2, \dots, d_k\}$$

where q is the scientific query and $R(q)$ represents the top- k retrieved documents or passages.

The generator then produces:

$$y = G(q, R(q))$$

where y represents the final response.

The quality of the final answer therefore depends on both retrieval and generation. If relevant evidence is not retrieved, even a highly capable generator may produce an incomplete answer. Conversely, retrieving excellent evidence does not guarantee that the generator will faithfully use it.

Recent surveys emphasize that modern RAG research increasingly considers retrieval quality, generation quality, robustness, transparency, privacy, and factuality as interconnected dimensions.

2.3 Scientific RAG

Scientific literature represents a particularly challenging retrieval environment. Scientific documents contain technical terminology, equations, tables, figures, abbreviations, domain-specific entities, and relationships that may not be captured effectively by simple keyword matching.

Scientific RAG systems therefore benefit from:

- domain-specific document collections;
- semantic embeddings;
- metadata filtering;
- hybrid retrieval;
- passage-level indexing;
- citation-aware generation;
- evidence verification;
- contradiction detection.

PaperQA and related systems demonstrated the potential of retrieval-based agents for answering questions from scientific documents while providing citations. More recent scientific RAG systems have expanded these ideas through larger corpora, improved retrievers, rerankers, and iterative verification.

The OpenScholar framework is particularly relevant because it was designed specifically for scientific literature synthesis. It uses a large scientific datastore, trained retrieval components, reranking, self-feedback, and citation verification. The reported results show that retrieval-supported systems can outperform non-retrieval baselines on scientific synthesis tasks.



2.4 Factuality and Faithfulness

Factuality and faithfulness are related but distinct.

Factuality asks:

Is the generated statement actually true?

Faithfulness asks:

Is the generated statement supported by the retrieved evidence?

A response may be factually correct but poorly grounded in the retrieved passages. Conversely, a response can faithfully summarize a retrieved passage that itself contains an error or outdated information.

RAGAs introduced a reference-free evaluation framework for RAG systems that separately evaluates retrieval relevance, generation faithfulness, and answer quality.

This distinction is particularly important for scientific knowledge discovery because researchers need both accurate information and transparent evidence.

2.5 Citation Reliability

Citation reliability should not be reduced to whether a citation exists.

A reliable scientific citation should satisfy several conditions:

1. The cited document should actually exist.
2. The source should be relevant to the claim.
3. The source should contain evidence supporting the claim.
4. The citation should be placed close to the claim it supports.
5. The generated statement should not exaggerate the evidence.
6. Important claims should not remain unsupported.

Recent research has further distinguished **citation correctness** from **citation faithfulness**. Citation correctness concerns whether a source supports a statement, whereas citation faithfulness examines whether the model genuinely relied on that source rather than merely attaching a plausible citation after generating a statement. This distinction is important because a citation can appear correct while the model's reasoning was not actually grounded in the cited evidence.

2.6 Research Gap

Existing RAG research has produced significant improvements in information retrieval and factual generation. However, several gaps remain:

- Retrieval quality is frequently evaluated separately from generation quality.
- Citation correctness and citation faithfulness are not always evaluated together.
- Scientific documents require domain-aware retrieval strategies.
- Contradictory findings across papers remain difficult to synthesize.
- Citation completeness is often overlooked.
- Retrieval systems may propagate errors contained in source documents.
- Long-form scientific answers create additional attribution challenges.

The proposed framework addresses these limitations by treating retrieval, generation, factuality, and citation verification as a unified pipeline.



PROPOSED METHODOLOGY

3.1 Research Design

This study proposes a conceptual and experimentally testable architecture for reliable scientific RAG. The framework consists of six major stages:

1. Scientific corpus construction.
2. Hybrid retrieval.
3. Evidence reranking.
4. Evidence-grounded generation.
5. Citation verification.
6. Self-reflection and final answer refinement.

The system can be represented as:

Query→Retrieval→Reranking→Evidence→Generation→Citation Verification→Final Answer
Query \rightarrow Retrieval \rightarrow Reranking \rightarrow Evidence \rightarrow Generation \rightarrow Citation \ Verification \rightarrow Final \ Answer

3.2 Scientific Corpus Construction

A domain-specific scientific corpus should be constructed from reliable sources such as:

- PubMed;
- PubMed Central;
- arXiv;
- Crossref metadata;
- Semantic Scholar;
- institutional repositories;
- open-access journal collections.

Each document should contain metadata such as:

Field	Description
Title	Scientific paper title
Authors	Paper authors
Year	Publication year
Abstract	Scientific abstract
Full Text	Available article content
DOI	Persistent identifier
Journal	Publication venue
Keywords	Scientific concepts
Sections	Introduction, Methods, Results, etc.

Source validation is essential because retrieval from unreliable or duplicated documents can directly affect the generated answer.



3.3 Document Processing and Chunking

Scientific documents should not be divided into arbitrary fixed-length chunks alone. Section-aware segmentation is preferable.

A document can be represented as:

$$D=\{S_1, S_2, \dots, S_n\} D=\{S_1, S_2, \dots, S_n\}$$

where each S_i represents a semantic section or passage.

Examples include:

- Abstract;
- Introduction;
- Methodology;
- Results;
- Discussion;
- Conclusion;
- References.

Metadata should remain attached to every chunk so that the generator can later construct accurate citations.

3.4 Hybrid Retrieval

A hybrid retrieval strategy is proposed by combining:

Lexical Retrieval

BM25 can identify exact terminology, names, abbreviations, and technical phrases.

Semantic Retrieval

Dense embeddings identify semantically similar passages even when different terminology is used.

The combined retrieval score can be expressed as:

$$\text{Score}(d, q) = \alpha \text{SBM25}(d, q) + (1 - \alpha) \text{SDense}(d, q) \quad \text{Score}(d, q) = \alpha S_{\text{BM25}}(d, q) + (1 - \alpha) S_{\text{Dense}}(d, q)$$

where:

- $\text{SBM25}_{\text{BM25}}$ = lexical similarity;
- $\text{SDense}_{\text{Dense}}$ = embedding similarity;
- α = weighting parameter.

Hybrid retrieval is particularly useful in scientific domains because terminology can be both highly specific and semantically variable.

3.5 Cross-Encoder Reranking

The initial retrieval stage may produce 20–50 candidate passages. A cross-encoder reranker can then evaluate the query and each candidate jointly.

$$R^* = \text{TopK}(\text{CrossEncoder}(q, d_i)) \quad R^* = \text{operatorname{TopK}}(\text{CrossEncoder}(q, d_i))$$

The purpose is to improve precision by removing passages that are semantically similar but scientifically irrelevant.

Reranking has become an important component in advanced scientific RAG pipelines. OpenScholar, for example, integrates specialized retrieval and reranking to improve scientific literature synthesis.

3.6 Evidence-Aware Generation

The generator should receive:

- original question;



- retrieved evidence;
- source metadata;
- citation identifiers;
- instructions to avoid unsupported claims.

A recommended generation constraint is:

Generate only claims supported by retrieved evidence. If sufficient evidence is unavailable, explicitly indicate uncertainty or insufficient evidence.

This changes the system objective from unrestricted generation to evidence-grounded synthesis.

A simplified objective is:

$$P(y|q,E)P(y|q,E)$$

where EE represents retrieved evidence.

3.7 Citation Mapping

Each generated claim should be mapped to one or more evidence passages.

For claim cic_i , the system determines:

$$C_i = \{e_1, e_2, \dots, e_m\} \quad C_i = \{e_1, e_2, \dots, e_m\}$$

where C_i represents the citation set supporting claim cic_i .

The system should avoid assigning one citation to a long paragraph containing multiple independent claims unless the source genuinely supports all of them.

3.8 Citation Verification

Citation verification is performed at three levels.

Level 1: Existence Verification

Check whether the cited article exists using DOI or bibliographic metadata.

Level 2: Relevance Verification

Determine whether the cited document is relevant to the claim.

Level 3: Entailment Verification

Determine whether the evidence actually supports the claim.

The verification process can be represented as:

$$V(c,e) = \text{Entailment}(c,e) \times \text{Relevance}(e,q) \quad V(c,e) = \text{Entailment}(c,e) \times \text{Relevance}(e,q)$$

A citation is accepted only when both relevance and evidence support exceed predefined thresholds.

3.9 Self-Reflection

The final stage uses a verification model or the same LLM to critique the generated answer.

The system asks:

1. Which claims require citations?
2. Does every major claim have evidence?
3. Does each citation support the corresponding claim?
4. Are any statements stronger than the retrieved evidence?
5. Are contradictory findings represented?
6. Are citations real and correctly formatted?

Self-RAG demonstrates the value of adaptive retrieval and self-reflection for improving factuality and citation accuracy.



EVALUATION FRAMEWORK

The proposed system should be evaluated using multiple dimensions rather than a single accuracy score.

4.1 Retrieval Metrics

Recall@K

$\text{Recall@K} = \frac{\text{Relevant documents retrieved in top K}}{\text{Total relevant documents}}$

Precision@K

$\text{Precision@K} = \frac{\text{Relevant documents in top K}}{K}$

Mean Reciprocal Rank

MRR measures how highly the first relevant document appears in the ranking.

4.2 Generation Metrics

The following dimensions should be evaluated:

Metric	Purpose
Factual Accuracy	Measures correctness of claims
Faithfulness	Measures grounding in retrieved evidence
Answer Relevance	Measures whether the answer addresses the question
Completeness	Measures coverage of important evidence
Coherence	Measures logical organization
Citation Precision	Measures whether citations support claims
Citation Recall	Measures whether citation-worthy claims are cited
Citation Faithfulness	Measures genuine evidence attribution

RAGAs provides a useful automated evaluation foundation for assessing retrieval relevance, faithfulness, and answer quality.

4.3 Proposed Experimental Comparison

A future implementation can compare the following configurations:

Model	Retrieval	Reranking	Verification	Expected Evaluation Focus
LLM-only	No	No	No	Baseline factuality
Basic RAG	Dense	No	No	Retrieval benefit
Hybrid RAG	BM25 + Dense	No	No	Retrieval improvement
Hybrid + Reranking	BM25 + Dense	Yes	No	Evidence precision
Verified RAG	BM25 + Dense	Yes	Yes	Citation reliability
Proposed RAG	Hybrid	Yes	Yes + Self-reflection	Overall reliability

These configurations allow researchers to conduct ablation studies and determine which components contribute most strongly to scientific reliability.



RESULTS

Because this paper proposes a reliability-oriented framework rather than reporting a newly executed benchmark, numerical experimental results are not fabricated. Instead, the Results section synthesizes evidence from recent scientific RAG research and identifies expected effects of the proposed architecture.

5.1 Retrieval Improves Scientific Knowledge Access

Existing studies indicate that retrieval can substantially improve access to current and domain-specific knowledge. Scientific RAG systems can retrieve passages that are absent from the model's parametric memory, enabling answers based on recent publications.

The OpenScholar study provides particularly strong evidence for this approach. Its scientific datastore contained approximately 45 million papers, and the system combined retrieval, reranking, and iterative feedback for literature synthesis. The study reported improvements over several non-specialized and retrieval-based baselines on scientific synthesis tasks.

5.2 Retrieval Alone Does Not Guarantee Reliability

An important finding from the literature is that RAG is not automatically reliable.

If the retriever selects irrelevant or contradictory passages, the generator can produce an answer that is grounded in inappropriate evidence.

Therefore:

Reliable RAG $\neq$ Retrieval + Generation
Reliable \ RAG $\neq$ Retrieval + Generation

Instead:

Reliable RAG = Retrieval + Reranking + Grounded Generation + Verification
Reliable \ RAG = Retrieval + Reranking + Grounded \ Generation + Verification

This observation is consistent with broader RAG surveys emphasizing that retrieval improvements do not necessarily translate directly into improvements in final generated answers.

5.3 Citation Reliability Requires Independent Evaluation

Citation existence is insufficient as an evaluation metric.

A generated answer may contain a legitimate DOI while incorrectly attributing a finding to that paper. Consequently, citation evaluation should measure both citation correctness and citation faithfulness.

Recent work identifying post-rationalized citations demonstrates that citation attribution itself can be unreliable even when sources appear legitimate.

5.4 Expected Effect of the Proposed Architecture

Based on existing literature, the proposed architecture is expected to provide the following improvements:

Component	Primary Contribution
Scientific corpus	Current and domain-specific evidence
Hybrid retrieval	Higher recall and terminology coverage
Reranking	Higher evidence precision
Evidence-aware generation	Reduced unsupported statements
Citation mapping	Improved claim-source alignment



Component	Primary Contribution
Citation verification	Reduced incorrect citations
Self-reflection	Detection of unsupported claims
Contradiction checking	Better treatment of conflicting studies

These expected improvements should be validated experimentally using scientific question-answering and literature-synthesis benchmarks.

DISCUSSION

The findings from existing research suggest that RAG represents an important architectural shift from knowledge generation based exclusively on model parameters toward evidence-grounded generation. However, scientific knowledge discovery imposes stronger requirements than conventional question answering.

First, **retrieval quality is fundamental**. A generator cannot reliably use evidence that was never retrieved. Scientific retrieval must therefore consider terminology, publication metadata, document sections, and semantic relationships.

Second, **retrieval quality and factuality are not equivalent**. A system may retrieve highly relevant papers but still misinterpret their conclusions. Evidence-aware generation and entailment checking are therefore necessary.

Third, **citation reliability must be treated as a first-class objective**. Traditional generation evaluation often concentrates on answer similarity or correctness. Scientific systems require additional measures for citation precision, recall, completeness, and faithfulness.

Fourth, **contradictory evidence requires explicit handling**. Scientific literature frequently contains different results because of variations in datasets, experimental conditions, sample populations, methodologies, and statistical assumptions. A reliable system should not automatically select one study. Instead, it should identify the disagreement and explain the relevant conditions.

Fifth, **source quality must be considered**. A RAG system can faithfully retrieve a low-quality source. Therefore, evidence reliability depends partly on corpus curation, metadata validation, publication status, and source provenance.

Sixth, **human oversight remains important**. RAG can accelerate literature discovery but should not automatically replace expert interpretation, especially in biomedical, clinical, engineering safety, or policy-sensitive contexts.

Recent scientific literature synthesis research provides evidence that specialized retrieval systems combined with self-feedback and citation verification can achieve substantial improvements over general-purpose LLM approaches. OpenScholar's evaluation across thousands of scientific queries illustrates the potential of specialized scientific RAG.

PRACTICAL APPLICATIONS

The proposed framework can support several scientific workflows.

7.1 Literature Review

Researchers can ask questions such as:



What are the major deep-learning approaches for early cancer detection published since 2020?
The system can retrieve relevant studies and organize findings by methodology, dataset, performance, and limitations.

7.2 Research Gap Identification

A RAG system can compare multiple papers and identify areas where:

- evidence is limited;
- methodologies have not been compared;
- datasets remain underused;
- contradictory findings exist;
- future research has been proposed repeatedly.

7.3 Hypothesis Generation

Evidence-grounded retrieval can help researchers identify relationships across disciplines and formulate hypotheses while clearly distinguishing established evidence from speculative suggestions.

7.4 Methodology Comparison

Researchers can compare algorithms, datasets, experimental protocols, evaluation metrics, and limitations across publications.

7.5 Scientific Question Answering

Domain-specific RAG can provide answers with traceable source passages, allowing researchers to inspect the evidence behind individual claims.

LIMITATIONS

Several limitations remain.

First, retrieval quality depends on the underlying scientific corpus. Papers that are unavailable, incorrectly indexed, or behind inaccessible sources may be excluded.

Second, scientific documents contain complex structures such as mathematical equations, tables, images, supplementary information, and references. Text-only retrieval may therefore lose important information.

Third, citation verification remains computationally expensive for long answers containing many claims.

Fourth, an LLM-based verifier may itself make incorrect judgments. Automated verification should therefore not be assumed to be perfectly reliable.

Fifth, scientific literature contains conflicting and evolving evidence. A system must avoid presenting uncertain findings as established facts.

Finally, evaluation benchmarks may not fully represent real-world scientific research tasks. Expert evaluation remains necessary for high-stakes scientific applications.

FUTURE RESEARCH DIRECTIONS

Future research should investigate several directions.

9.1 Multimodal Scientific RAG



Future systems should retrieve and reason over:

- text;
- tables;
- equations;
- figures;
- experimental images;
- supplementary datasets.

9.2 Knowledge-Graph-Enhanced RAG

Scientific knowledge graphs could connect:

Author→Paper→Method→Dataset→Result→Citation
Author \rightarrow Paper \rightarrow Method
\rightarrow Dataset \rightarrow Result \rightarrow Citation

This would improve structured reasoning and relationship discovery.

9.3 Agentic Scientific RAG

Agentic systems could autonomously perform:

1. query decomposition;
2. literature search;
3. evidence filtering;
4. contradiction detection;
5. synthesis;
6. citation verification;
7. iterative refinement.

9.4 Citation Faithfulness

Future benchmarks should explicitly distinguish:

- citation correctness;
- citation completeness;
- citation relevance;
- citation faithfulness.

9.5 Uncertainty-Aware Generation

Instead of forcing an answer, systems should be capable of producing:

- high-confidence conclusions;
- moderate-confidence interpretations;
- insufficient-evidence responses.

9.6 Continuous Scientific Knowledge Updating

Scientific RAG systems should support continuously updated indexes so that newly published research can be incorporated without retraining the entire language model.



CONCLUSION

Retrieval-Augmented Generation provides a promising foundation for reliable scientific knowledge discovery by connecting LLMs with external scientific literature. Nevertheless, conventional RAG does not automatically eliminate hallucination or guarantee accurate citations. Scientific applications require a more comprehensive architecture in which retrieval, reranking, evidence-grounded generation, citation mapping, verification, and self-reflection operate together.

This paper proposed a reliability-oriented scientific RAG framework that combines hybrid retrieval, cross-encoder reranking, evidence-aware generation, citation verification, and self-reflection. The literature indicates that these components address complementary failure modes: retrieval improves access to current knowledge, reranking improves evidence selection, grounded generation reduces unsupported claims, citation verification improves attribution, and self-reflection enables iterative correction.

The central conclusion is that **scientific RAG should be evaluated not merely by whether it retrieves relevant documents or generates fluent answers, but by whether each important scientific claim can be traced to appropriate evidence**. Future systems should therefore treat factuality, citation reliability, uncertainty, provenance, and expert verification as integral components of scientific AI.

A reliable scientific knowledge-discovery system should ultimately provide not only an answer, but also a transparent evidence trail showing **where the answer came from, what evidence supports it, and where uncertainty remains**.

REFERENCES

1. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2310.11511>
2. Brown, A., Roman, M., & Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *Big Data and Cognitive Computing*, 9(12), 320. <https://doi.org/10.3390/bdcc9120320>
3. Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). 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* (pp. 150–158). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
4. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Meng, W., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
5. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *European Conference on Information Retrieval*, 874–880. https://doi.org/10.1007/978-3-030-72240-1_81
6. Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., & White, A. D. (2023). PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv*. <https://doi.org/10.48550/arXiv.2312.07559>



7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://doi.org/10.48550/arXiv.2005.11401>
8. Li, C., & Flanigan, J. (2025). RAC: Efficient LLM factuality correction with retrieval augmentation. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 25145–25159. <https://doi.org/10.18653/v1/2025.findings-emnlp.1370>
9. Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., Zeng, Q., Dong, L., Xia, Y., Kenthapadi, K., Rossi, R., Dernoncourt, F., Tanjim, M. M., Ahmed, N., Liu, X., Fan, W., Blasch, E., Wang, Y., Jiang, M., & Derr, T. (2025). Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2502.06872>
10. Ovadia, M., Tannenbaum, A., & others. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2409.10102>
11. Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnampati, M., Rodriques, S. G., & White, A. D. (2024). Language agents achieve superhuman synthesis of scientific knowledge. *arXiv*. <https://doi.org/10.48550/arXiv.2409.13740>
12. Wallat, J., Heuss, M., de Rijke, M., & Anand, A. (2025). Correctness is not faithfulness in retrieval augmented generation attributions. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. <https://doi.org/10.48550/arXiv.2406.07326>
13. Zhang, S., et al. (2024). Corrective retrieval augmented generation. *arXiv*. <https://doi.org/10.48550/arXiv.2401.15884>
14. Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Dou, Z., Ho, T.-S., & Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2409.10102>
15. White, A. D., Cox, S., Lála, J., O'Donoghue, O., Shtedritski, A., Rodriques, S. G., & others. (2025). Synthesizing scientific literature with retrieval-augmented language models. *Nature*. <https://doi.org/10.1038/s41586-025-10072-4>