A Review on Knowledge Graph-Based Retrieval-Augmented Generation for Reliable Scientific Response Generation

Vishal S. Patil[1], Dr. R. R. Keole[2]
[1]*Phd Research Scholar, Department of Computer Science and Engineering, Hvpm College of Engineering and Technology, Amravati, Sant Gadge Baba Amravati University, Amravati*
[2]*Professor And Head, Department of Information Technology, Hvpm College of Engineering and Technology, Amravati, Sant Gadge Baba Amravati University, Amravati*

***Abstract*—The world is rapidly evolving around AI and generative AI. From chat bots that write essays to models that sketch artworks in seconds, we are now living in an age where Large Language Models (LLMs) can read, summarize, and even draft scientific papers in a matter of minutes. Yet, when it comes to science, these LLMs still suffer from major reliability problems: they often produce unsupported claims, omit proper citations, invent references, mix up metadata, and create semantic inconsistencies between the answer and the cited papers. Recent developments such as citation aware generation, GraphRAG, claim verification tools, knowledge graph builders, and RAG evaluation frameworks have started to tackle these issues – but most solutions focus on only one aspect (e.g., improving retrieval, tightening citation integration, detecting hallucinations, or verifying individual claims). What we need is a completely new approach that moves us from a simple "retrieve then generate" workflow to a "verify first" paradigm. This review paper discusses recent developments aimed at improving the trustworthiness of LLM-based scientific writing. It examines RAG, citation-aware generation, GraphRAG, scientific claim verification, knowledge graph construction, and RAG evaluation methods. The review identifies several research gaps, including fragmented solutions that separately address retrieval, citation, hallucination, and verification. Existing studies also provide limited support for claim-level graph traversal, citation reliability assessment, contradiction detection, and auditable reasoning. Therefore, a unified verification framework is required to validate evidence and semantic consistency before generating the final scientific response.**

***Index Terms*—Retrieval Augmented Generation, GraphRAG, Citation Reliability, Response Verification, Claim Verification, Evidence Path Verification, Scientific LLM Systems**

I. INTRODUCTION

Large Language Models (LLMs) are increasingly used for literature review, paper summarization, research-gap identification, and citation-based scientific writing. These tools can reduce the time required for academic work. However, scientific writing requires more than well-formed text. It must include reliable evidence, correct citations, accurate metadata, and consistent interpretation of research findings Retrieval-Augmented Generation (RAG) improves LLM responses by retrieving relevant research papers and passages before answer generation. Although this reduces hallucination, conventional RAG systems still have several limitations. A system may retrieve a relevant paper but generate an unsupported claim, attach the wrong citation, combine results from incompatible studies, or ignore contradictory evidence. Recent methods address these problems through citation-aware generation, citation verification, GraphRAG, scientific knowledge graphs, claim verification, and RAG evaluation. Systems such as ALCE, VeriCite, CiteCheck, SciFact, RAGAS, and ARES contribute to improving citation quality, evidence verification, and response faithfulness. However, most approaches focus on individual problems rather than providing a complete solution. Therefore, scientific LLM systems require an

integrated claim–citation–evidence verification pipeline. Such a system should verify every important claim, check citation correctness, identify supporting and refuting evidence, detect contradictions, and maintain semantic consistency before generating the final response. This review examines recent developments in this area and highlights the need for a graph-augmented, verification-first framework for trustworthy scientific LLM systems.

II. BACKGROUND

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation is a framework where an LLM uses external documents before generating an answer. The basic process is:

1. User asks a question.
2. The system retrieves relevant documents.
3. Retrieved documents are added to the LLM prompt.
4. The LLM generates the final answer.

This approach is better than pure LLM generation because it provides updated and domain-specific knowledge. However, the main weakness is that the system may retrieve relevant-looking documents but still generate unsupported claims.

2.2 GraphRAG

GraphRAG extends RAG by using graph structures. Instead of storing only text chunks, GraphRAG represents entities and relationships in graph form. In scientific literature, the graph may contain papers, authors, citations, methods, datasets, metrics, claims, and results. This helps the system perform multi-hop reasoning and understand relationships among papers.[16].

2.3 Citation-Aware Generation

Citation-aware generation means generating answers with citations. This is important for scientific writing because users need to know which source supports each claim. However, citation-aware generation alone is not sufficient. A citation may be present, but it may not support the exact claim.

2.4 Response Verification

Response verification means checking the generated answer before showing it to the user. In scientific LLM systems, response verification should check whether each generated claim is supported by valid evidence, whether the citation is reliable, and whether any contradictory evidence exists [12],[14],[15].

2.5 Existing System Flow

The general flow of existing RAG-based scientific LLM systems is shown below.

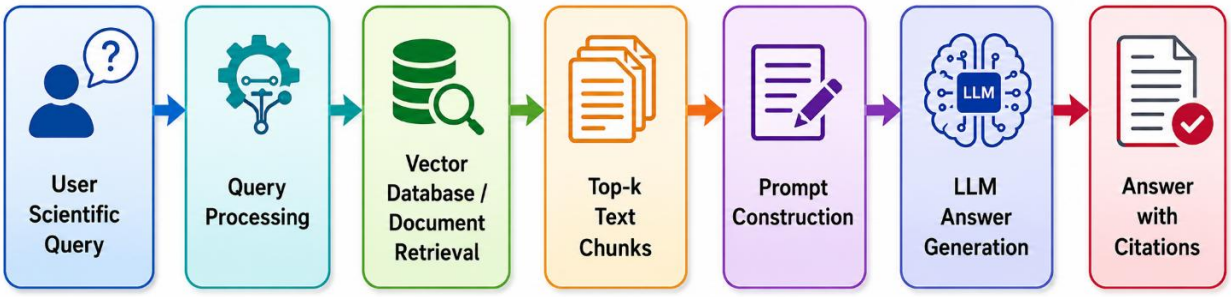

Figure 1. General flow of the existing RAG-based scientific LLM system

In this flow, the system retrieves documents and generates an answer. However, there is no strong verification stage between answer generation and final output. Because of this, the system may generate unsupported statements, weak citations, or contradictory conclusions.

III. LITERATURE REVIEW

3.1 TGRV: Triple Graph-Based Retrieval and Verification for Hallucination Detection in Large Language Models

TGRV is one of the most relevant papers for graph-based response verification. The paper focuses on hallucination detection in open-ended LLM responses [1]. It decomposes generated text into triples and

constructs a response graph. It also constructs an external knowledge graph from retrieved evidence. Then it compares both graphs to check whether the generated response is consistent with external evidence. The important contribution of TGRV is that it treats hallucination detection as a graph verification problem. Instead of checking only sentences, it checks structured entity-relation triples. This is useful because scientific claims often contain relationships between methods, datasets, metrics, and results. However, TGRV is not specifically designed for scientific literature. It does not deeply verify DOI, author, title, venue, citation context, or scholarly metadata. It also does not fully model support, refute, partial support, and contradiction relations among scientific papers. Therefore, TGRV is a strong base paper, but it needs to be extended for scientific LLM systems.

3.2 CG-RAG: Research Question Answering by Citation Graph Retrieval-Augmented LLMs

CG-RAG focuses on research question answering using citation graph-based RAG [2]. It is highly relevant because scientific papers are connected through citations. A paper may support, extend, compare, or challenge another paper. Therefore, citation graph structure can improve scientific evidence retrieval. The paper introduces contextualized citation graph representation and combines sparse lexical retrieval with dense semantic retrieval. This helps the system retrieve papers that are both textually relevant and structurally connected through citations. This is useful for scientific question answering because the answer may depend on a group of related papers, not only one isolated passage.

The limitation is that CG-RAG mainly uses the citation graph for retrieval. It does not fully verify whether the cited paper supports the generated claim. It also does not include DOI validation, citation reliability scoring, or contradiction-aware traversal before answer generation.

3.3 Relink: Constructing Query-Driven Evidence Graph On-the-Fly for GraphRAG

Relink proposes a query-driven evidence graph construction method for GraphRAG [3]. This is important because pre-built knowledge graphs may contain missing links, noisy facts, or irrelevant relationships. In scientific tasks, each user query may require a different evidence path. Therefore, a fixed graph may not be enough. Relink constructs an evidence graph on the fly according to the query. It repairs missing links, removes distractor facts, and creates a more faithful evidence graph. This is useful for scientific LLM systems because generated claims may require claim-specific evidence paths. However, Relink does not focus on scientific citation metadata, DOI validation, citation-context support, or support/refute labels. It provides a strong idea for dynamic evidence graph construction, but it needs to be adapted for scientific claim verification and citation reliability.

3.4 LLM-Based Corroborating and Refuting Evidence Retrieval for Scientific Claim Verification

This paper introduces CIBER, an LLM-based method for retrieving both corroborating and refuting evidence for scientific claim verification [4]. It is very important because a reliable scientific system should not only search for supporting papers. It should also search for papers that may refute or challenge the generated claim. The paper uses interrogation probes and response consistency to retrieve supporting and refuting evidence. One major strength is that it can work without internal access to the LLM model. This makes it useful for practical systems that use black-box LLM APIs. The limitation is that the retrieved evidence is not organized as a dynamic scientific knowledge graph. The system does not directly provide citation reliability scoring, contradiction severity scoring, or auditable evidence-path reporting. Therefore, this work can be extended by converting support/refute evidence into graph relations.

3.5 Research on the Construction and Application of Retrieval Enhanced Generation Model Based on Knowledge Graph

This paper proposes a knowledge graph-enhanced RAG model that combines document retrieval with graph-based semantic path retrieval [5]. It uses DPR-based text retrieval and GNN-based graph path retrieval. The system applies path attention and knowledge injection to improve generation. This work is useful because it shows that combining unstructured text and structured graph knowledge can improve factual consistency. It also supports the idea that graph paths can provide interpretable reasoning support.

However, the proposed KG-RAG system is not specifically designed for scientific literature verification. It does not include citation reliability scoring, DOI checking, or contradiction-aware scientific evidence paths. It mainly improves retrieval and generation, while response verification remains limited.

3.6 GraphCheck: Breaking Long-Term Text Barriers with Extracted Knowledge Graph-Powered Fact-Checking

GraphCheck focuses on fact-checking long-form text using extracted knowledge graphs [6]. Long documents are difficult for LLMs because relevant evidence may be scattered across many paragraphs. Scientific papers also have this problem because evidence may appear in different sections such as methodology, experiments, results, and discussion. GraphCheck extracts knowledge graphs from claims and grounding documents. It then uses graph neural networks as structured soft prompts for LLM-based verification. This helps the system capture multi-hop reasoning chains. The limitation is that GraphCheck is mainly designed for fact-checking against grounding documents, not scientific citation graphs. It does not include DOI validation, scholarly metadata checking, or pre-generation optimization of scientific responses. Still, it is useful for designing a graph-based response verification engine.

3.7 VeriCite: Towards Reliable Citations in Retrieval-Augmented Generation via Rigorous Verification

VeriCite addresses citation reliability in RAG [7]. It focuses on a major weakness of standard RAG: even when relevant passages are retrieved, generated citations may not support the exact statement. VeriCite introduces a staged framework for answer generation, evidence selection, citation verification, and answer refinement. This paper is important because it treats citation quality as a separate problem. In scientific writing, a citation must not only appear in the answer; it must support the exact claim. However, VeriCite mainly focuses on passage-level citation support. It does not fully model cross-paper contradiction through graph traversal. It also does not deeply validate DOI, publication metadata, source authority, or venue reliability. Therefore, it can be extended with graph-based citation evidence paths.

3.8 Detecting Miscitation on the Scholarly Web through LLM-Augmented Text-Rich Graph Learning

This paper focuses on detecting miscitation in scholarly web text [8]. Miscitation means the cited paper exists, but it is used incorrectly or does not support the claim. This problem is very relevant for scientific LLM systems because generated answers may attach a real paper to an unsupported claim. The paper uses text-rich citation graphs, citation contexts, scholarly metadata, LLM-based evidence-chain reasoning, and GNN-based learning. It is important because it combines citation context and graph learning for citation fidelity. The limitation is that it detects miscitation in scholarly web text, not directly in generated RAG answers. It does not provide a full response verification or semantic consistency optimization pipeline. However, its citation fidelity idea can be used in citation reliability scoring.

3.9 CiteCheck: Retrieval-Grounded Detection of LLM Citation Hallucinations in Scientific Text

CiteCheck focuses on detecting hallucinated citations in scientific text [9]. It checks whether citation metadata such as title, author, year, DOI, and venue are correct. This is very important because LLMs may generate realistic-looking but wrong references. The method parses citation strings into metadata fields and retrieves candidate publications from scholarly sources. Then it compares the generated citation with retrieved publication records and identifies exact matches, minor errors, or major hallucinations.The limitation is that CiteCheck mainly validates citation existence and metadata correctness. It does not fully answer whether the valid citation semantically supports, refutes, or partially supports the generated claim. Therefore, it should be combined with claim verification and graph traversal.

3.10 FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation

FaithfulRAG addresses fact-level conflict in RAG [10]. It studies the problem where retrieved context and model parametric knowledge may contain different facts. The paper uses conflict modeling and self-thinking to improve context-faithful generation. This paper is important for semantic consistency. A scientific LLM may combine retrieved evidence, older model knowledge, and conflicting papers. If conflict is

not detected, the final answer may become misleading. The limitation is that FaithfulRAG mainly focuses on conflict between retrieved context and model knowledge. It does not fully model contradictions among scientific papers, datasets, methods, metrics, and results. It also does not include citation reliability or DOI validation. Therefore, it can be extended for scientific contradiction-aware response optimization.

IV. COMPARATIVE ANALYSIS

Sr.	Paper	Main Focus	Key Strength	Major Limitation
1	TGRV [1]	Graph-based hallucination verification	Uses response graph and external knowledge graph for verification	Not specialized for scientific citation metadata and scholarly provenance
2	CG-RAG [2]	Citation graph RAG for research QA	Uses citation graph structure for scientific retrieval	Uses graph mainly for retrieval, not full verification
3	Relink [3]	Query-driven evidence graph construction	Builds evidence graph on the fly	Does not include citation reliability and scientific metadata validation
4	CIBER [4]	Support/refute evidence retrieval	Retrieves both corroborating and refuting evidence	Evidence is not stored as a dynamic scientific KG
5	KG-RAG [5]	Knowledge graph-enhanced RAG	Combines text retrieval and graph path retrieval	Does not include scientific citation verification
6	GraphCheck [6]	Graph-powered fact-checking	Uses extracted KGs for long-form verification	Not focused on scientific citation graphs
7	VeriCite [7]	Citation support verification	Checks whether citations support generated statements	Limited graph-based contradiction modeling
8	Miscitation Detection [8]	Citation fidelity checking	Uses text-rich citation graphs and LLM reasoning	Not directly designed for generated RAG responses
9	CiteCheck [9]	Citation hallucination detection	Checks DOI, title, author, year, and venue metadata	Does not verify claim-level semantic support
10	FaithfulRAG [10]	Fact-level conflict modeling	Improves context-faithful generation	Does not model inter-paper scientific contradiction paths

From this comparison, it is clear that current systems solve important sub-problems, but they do not fully combine all required components in one scientific LLM framework.

V. IDENTIFIED RESEARCH GAPS

Based on the literature review, the following research gaps are identified:

• Existing RAG systems retrieve relevant text chunks but do not fully represent scientific evidence as a graph of papers, claims, methods, datasets, results, and citations.

• Citation-aware generation systems attach citations, but they may not verify whether the citation supports the exact generated claim.

• Citation hallucination detection systems check metadata, but they do not fully connect citation correctness with semantic claim verification.

• GraphRAG systems improve retrieval and multi-hop reasoning, but many of them do not perform mandatory response verification before final answer generation.

• Scientific claim verification systems retrieve support/refute evidence, but they do not always convert evidence into dynamic scientific knowledge graphs.
• Semantic consistency methods detect some conflicts, but they do not fully handle contradiction across papers, datasets, methods, metrics, and results.
• Existing RAG evaluation methods do not fully measure citation reliability, evidence-path correctness, contradiction severity, and reasoning trace correctness.

VI. FUTURE RESEARCH DIRECTIONS

• Represent scientific literature as a graph linking papers, authors, claims, methods, datasets, metrics, results, citations, and support/contradiction relations.
• Verify each claim's evidence path, ensuring a clear chain from claim to paper, citation, method, dataset, and result.
• Develop a citation-reliability score that checks DOI validity, title-author-year match, venue trustworthiness, source authority, citation context, and metadata consistency.
• Provide contradiction-aware responses that present conflicting views and explain discrepancies (e.g., dataset, method, or metric differences).
• Offer an auditable reasoning trace that maps evidence to claims and documents why claims were accepted, revised, suppressed, or flagged.
• Introduce evaluation metrics covering citation reliability, evidence consistency, contradiction-F1, hallucination reduction, answer faithfulness, and reasoning-trace correctness.[17],[18].

VII. CONCLUSION

Scientific LLM systems are becoming important tools for literature review, research question answering, citation-based writing, and academic support. However, current systems are still limited by hallucinated claims, weak citations, unsupported statements, incorrect metadata, and contradictions across scientific papers.

This review studied ten recent papers related to GraphRAG, citation-aware generation, citation reliability, scientific claim verification, evidence graph construction, and response verification. The review shows that existing methods provide useful solutions, but they mostly solve separate parts of the problem. Some papers improve graph retrieval, some verify citations, some retrieve support/refute evidence, and some detect factual conflict. However, no single system fully combines scientific knowledge graph construction, claim-level graph traversal, citation reliability, contradiction detection, semantic consistency optimization, and auditable reasoning trace generation. Therefore, future research should move from simple retrieve-and-generate systems toward verification-first scientific LLM systems.

REFERENCES

[1] Y. Zhang, Y. Zeng, Y. Zhu, M. Song, W. Wang, and Z. Ou, "TGRV: Triple Graph-based Retrieval and Verification for Hallucination Detection in Large Language Models," in Proc. IEEE DSIS, 2025, doi: 10.1109/DSIS67228.2025.11390158.
[2] Y. Hu, Z. Lei, Z. Dai, A. Zhang, A. Angirekula, Z. Zhang, and L. Zhao, "CG-RAG: Research Question Answering by Citation Graph Retrieval-Augmented LLMs," arXiv preprint arXiv:2501.15067, 2025.
[3] M. Huang, C. Bu, Y. He, X. Zhuo, and X. Wu, "Relink: Constructing Query-Driven Evidence Graph On-the-Fly for GraphRAG," arXiv preprint arXiv:2601.07192, 2026.
[4] S. Wang, J. R. Foulds, M. O. Gani, and S. Pan, "LLM-based Corroborating and Refuting Evidence Retrieval for Scientific Claim Verification," arXiv preprint arXiv:2503.07937, 2025.
[5] S. Wang, H. Yang, and W. Liu, "Research on the Construction and Application of Retrieval Enhanced Generation (RAG) Model Based on Knowledge Graph," Scientific Reports, vol. 15, Art. no. 40425, 2025, doi: 10.1038/s41598-025-21222-z.
[6] Y. Chen, H. Liu, Y. Liu, R. Yang, H. Yuan, Y. Fu, P. Zhou, Q. Chen, J. Caverlee, and I. Li, "GraphCheck: Breaking Long-Term Text Barriers with Extracted Knowledge Graph-Powered Fact-Checking," arXiv preprint arXiv:2502.16514, 2025.
[7] H. Qian, Y. Fan, J. Guo, R. Zhang, Q. Chen, D. Yin, and X. Cheng, "VeriCite: Towards Reliable Citations in Retrieval-Augmented Generation via

Rigorous Verification," in Proc. SIGIR-AP / ACM, 2025, doi: 10.1145/3767695.3769505.

[8] H. Wu, H. Xiang, J. Gao, X. Zhao, D. Wu, and J. Li, "Detecting Miscitation on the Scholarly Web through LLM-Augmented Text-Rich Graph Learning," in Proc. ACM Web Conference, 2026, doi: 10.1145/3774904.3792568.

[9] K. Khajavi, S. Sadeghi, R. Adhikari, and A. Tessier, "CiteCheck: Retrieval-Grounded Detection of LLM Citation Hallucinations in Scientific Text," arXiv preprint arXiv:2605.27700, 2026.

[10] Q. Zhang, Z. Xiang, Y. Xiao, L. Wang, J. Li, X. Wang, and J. Su, "FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation," arXiv preprint arXiv:2506.08938, 2025.

[11] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large Language Models to Generate Text with Citations," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.

[12] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.

[13] I.-C. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, and P. Liu, "FacTool: Factuality Detection in Generative AI—A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios," arXiv preprint arXiv:2307.13528, 2023.

[14] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or Fiction: Verifying Scientific Claims," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[15] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-Verification Reduces Hallucination in Large Language Models," in Findings of the Association for Computational Linguistics (ACL), 2024.

[16] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From Local to Global: A GraphRAG Approach to Query-Focused Summarization," arXiv preprint arXiv:2404.16130, 2024.

[17] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proc. European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL), 2024.

[18] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in Proc. North American Chapter of the Association for Computational Linguistics (NAACL), 2024.