INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 10, October - 2026

DOIs:10.2015/IJIRMF/202610030 --:-- Research Paper / Article / Review

EvidenceGraph: A Claim-Centric Evidence Attribution and Verification Framework for Trustworthy RAG

[1] Lakshmi Muthavarapu , [2] Dr. David Raju Kolluri
[1]UG Scholar , Department of Data Science, Hyderabad Institute of Technology and Management, Hyderabad, India
[2]Associate Professor, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management, Hyderabad, India Email[1]lakshmimuthavarapu123@gmail.com, [2]davidrajuk.cse@hitam.org,

Abstract: *Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems are increasingly used to answer questions from scientific literature. However, generated answers may contain unsupported claims or conflicting information, making verification difficult. This paper presents EvidenceGraph, a framework that links individual claims to relevant evidence and source documents. It combines semantic retrieval using ChromaDB, answer generation, claim extraction, evidence verification, and graph-based relationship analysis. Evaluated on 30 questions from three scientific papers, the system produced 96 claims, of which 62.5% were supported, 20.8% contradicted, and 16.7% had insufficient evidence. Target-paper recall reached 100% at Top-10 retrieval, and all 151 automated tests passed. The results highlight EvidenceGraph's potential for improving evidence traceability and claim-level auditing in scientific RAG systems.*

Key Words: *Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Evidence Attribution, Claim Verification, Semantic Retrieval, Conflict Detection, Source Traceability, Trustworthy AI.*

1. INTRODUCTION:

The rapid development of Large Language Models (LLMs) has transformed the way people access, process, and understand information. These models can generate human-like responses and assist users in tasks such as question answering, summarization, and information retrieval. In scientific research, LLMs can help users explore research papers and obtain relevant information efficiently. However, despite their capabilities, these models may generate inaccurate statements, unsupported claims, or information that does not match the original source. Such issues are particularly important in scientific applications, where the accuracy and traceability of information are essential.

Retrieval-Augmented Generation (RAG) has emerged as an approach to address some of these limitations by combining information retrieval with language generation. Instead of relying only on knowledge learned during training, a RAG system retrieves relevant passages from external documents and provides them as context for generating an answer. This approach enables users to obtain responses grounded in scientific papers and other knowledge sources. However, retrieving relevant documents does not guarantee that every statement in the generated answer is supported by the retrieved evidence. When an answer contains multiple factual claims, identifying the exact passage that supports or contradicts each claim can be difficult. Traditional RAG systems may also provide limited visibility into relationships between evidence passages from different sources.

To address these challenges, claim-level verification has gained importance in developing more transparent and trustworthy AI systems. Breaking a generated response into smaller, independent factual statements allows each claim to be examined separately. By comparing individual claims with retrieved evidence, a system can identify supported claims, contradictory evidence, and cases where the available information is insufficient. In addition, analysing relationships between evidence passages can help reveal repeated information, complementary findings, and potential disagreements. These capabilities are useful for scientific literature analysis, where findings may vary according to research methods, datasets, and experimental conditions.

In this work, EvidenceGraph, a claim-centric evidence attribution and verification framework for trustworthy RAG, is proposed. The system uses semantic embeddings generated by the Sentence-Transformers model *all-MiniLM-L6-v2* and ChromaDB for storing and retrieving relevant research-paper passages.

The retrieved information is used for answer generation, followed by atomic claim extraction, claim-to-evidence verification, source attribution, and evidence relationship analysis. The identified relationships are represented through an interactive graph connecting claims, evidence passages, source documents, and their relationships.

The proposed framework was evaluated using 30 questions from three published scientific papers, resulting in 96 generated claims and 478 evidence chunks. The evaluation found that 62.5% of the claims were supported, 20.8% were contradicted, and 16.7% had insufficient evidence. Target-paper retrieval recall reached 100% at Top-10 retrieval, and the conflict-analysis component detected 94 evidence conflicts under the implemented evaluation. These results demonstrate the potential of EvidenceGraph to improve claim-level auditing, evidence traceability, and transparency in scientific RAG systems.

2. LITERATURE REVIEW:

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation was introduced as a way of combining language generation with information retrieved from external knowledge sources. Lewis et al. introduced RAG for knowledge-intensive language tasks, while REALM also demonstrated the value of retrieving external information during language modelling. Dense Passage Retrieval later showed how semantic representations can be used to retrieve relevant passages for question answering. In a conventional RAG pipeline, relevant document chunks are retrieved and provided to the language model as context. This approach is useful because the model does not have to depend only on information stored in its parameters. However, the retrieved passages are generally treated as one combined context, making detailed claim-level verification difficult.

2.2 Hallucination and Faithfulness

Hallucination remains a major problem in language-model applications. A model may generate information that is not present in the retrieved documents or may combine different pieces of information incorrectly. Previous studies have examined hallucination and factual consistency in natural language generation and RAG systems. Faithfulness is especially important when RAG is used for scientific literature. A response can appear reasonable while still containing a statement that cannot be supported by the retrieved evidence.

2.3 Citation and Evidence Attribution

Citations are commonly used to show where information in an AI-generated response comes from. Research on attributed question answering and citation generation has improved the traceability of generated information. However, attaching one citation to an entire sentence or answer does not always show which individual statement is actually supported by the source. EvidenceGraph addresses this issue by connecting individual claims with the evidence passage used for verification and retaining source and page information.

2.4 Claim Decomposition and Verification

Breaking a generated response into smaller factual units makes verification easier. FActScore demonstrated the use of atomic facts for evaluating factual precision in long-form generation. Other work on factual consistency and verification has used relationships such as entailment, contradiction and neutral evidence. EvidenceGraph applies the same general idea to scientific RAG responses by classifying each generated claim as SUPPORT, CONTRADICT, or INSUFFICIENT based on retrieved passages.

2.5 Conflicting Evidence

Scientific literature does not always provide identical conclusions. Different papers may report different results because of differences in datasets, methods, experiments or research conditions. EvidenceGraph compares retrieved evidence passages and identifies REDUNDANCY, COMPLEMENTARITY and CONTRADICTION. These relationships are represented through the graph so that the user can inspect how evidence passages relate to one another.

Research Gap

Existing RAG systems are effective for retrieving information and generating answers. Citation and verification approaches improve traceability, while conflict-focused methods analyse disagreements between information sources. However, these capabilities are often handled separately. The gap addressed by EvidenceGraph is the combination of claim-level verification, evidence attribution, source tracing and evidence relationship analysis within one framework. The proposed system connects individual claims to their evidence and source documents while also showing relationships between retrieved evidence passages.

3. OBJECTIVES:

• To develop a RAG-based system for answering questions using scientific research papers.

• To divide generated answers into individual atomic claims.

• To associate each claim with relevant evidence from the retrieved research documents.

• To classify claims as SUPPORT, CONTRADICT or INSUFFICIENT.

• To identify relationships such as REDUNDANCY, COMPLEMENTARITY and CONTRADICTION between evidence passages.

• To provide source-level attribution and claim-level confidence information.

• To represent the connections between claims, evidence and sources through an interactive EvidenceGraph.

4. RESEARCH METHOD / METHODOLOGY

4.1 Dataset

The system uses scientific research papers in PDF format as its main knowledge source. For the final real-paper evaluation, three published scientific papers were selected. Ten questions were prepared for each paper, giving a total of 30 research questions. The evaluation generated 96 claims and used 478 evidence chunks. Target document and page information were retained so that retrieval results could be checked against the expected source information.

4.2 Document Processing

The PDF documents are processed using PyMuPDF. The extracted text is cleaned to remove unnecessary line breaks, excessive spaces and other formatting noise. Document names and page numbers are retained during processing so that the original source location is not lost.

4.3 Text Chunking

After cleaning, the document text is divided into smaller overlapping chunks. The implemented configuration uses a 500-character window with 100-character overlap. Each chunk receives an identifier along with its document name, page number and text.

4.4 Embedding Generation

The text chunks are converted into numerical representations using sentence-transformers/all-MiniLM-L6-v2. The model produces 384-dimensional embeddings. These embeddings allow the system to compare a user query with stored document chunks based on semantic similarity.

4.5 Vector Storage and Retrieval

The generated embeddings are stored in ChromaDB along with associated metadata. A user query is encoded using the same embedding model and compared with stored vectors using cosine similarity. The most relevant chunks are retrieved with their document and page information.

4.6 RAG Answer Generation

The retrieved evidence is supplied to the language model as context. The model uses the retrieved information to generate an answer. The generated answer is then passed to claim extraction and evidence checking rather than being treated as automatically verified.

4.7 Atomic Claim Extraction

The generated answer is divided into smaller, self-contained statements called atomic claims. This allows each statement to be examined independently instead of treating the complete response as one unit.

4.8 Claim-to-Evidence Verification

Each atomic claim is compared with retrieved evidence. SUPPORT means the evidence agrees with the claim. CONTRADICT means the evidence contains information opposing the claim. INSUFFICIENT means the retrieved evidence is not enough to confirm or reject the claim.

4.9 Evidence Relationship Analysis

The retrieved evidence passages are compared with each other. Redundancy means similar information is repeated. Complementarity means different passages add useful information together. Contradiction means two evidence passages contain opposing information.

4.10 Source Contribution and Claim-Level Confidence

EvidenceGraph provides source contribution and claim-level confidence information. Source contribution indicates how much a source contributes to the evidence used for answering a query. Claim-level confidence indicates the strength of available evidence for an individual claim.

4.11 EvidenceGraph Construction



Figure 1. Proposed EvidenceGraph framework showing the complete claim-centric evidence attribution and verification pipeline.

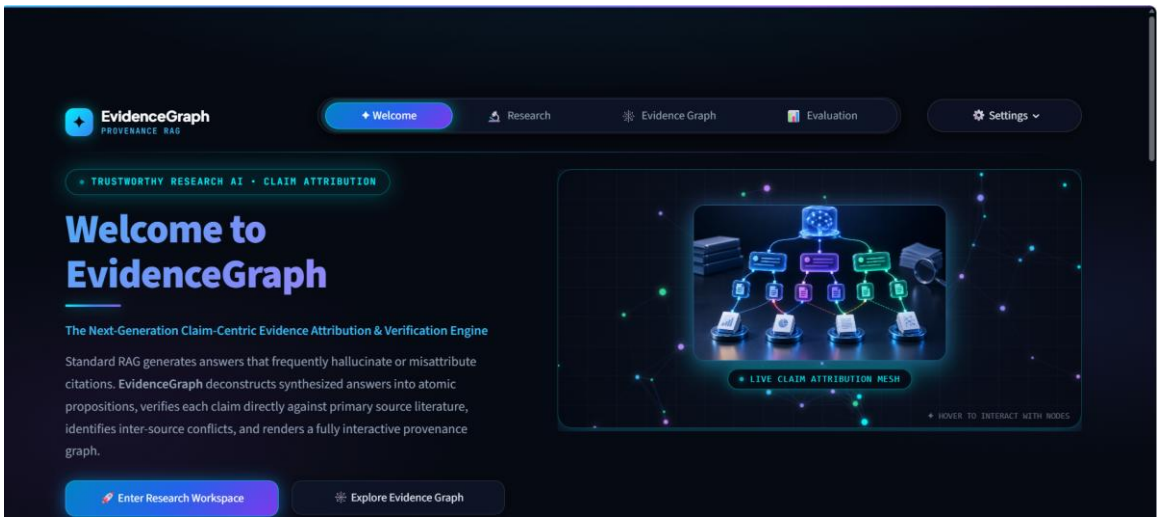


Figure 2. EvidenceGraph home page presenting the project overview and main navigation options.

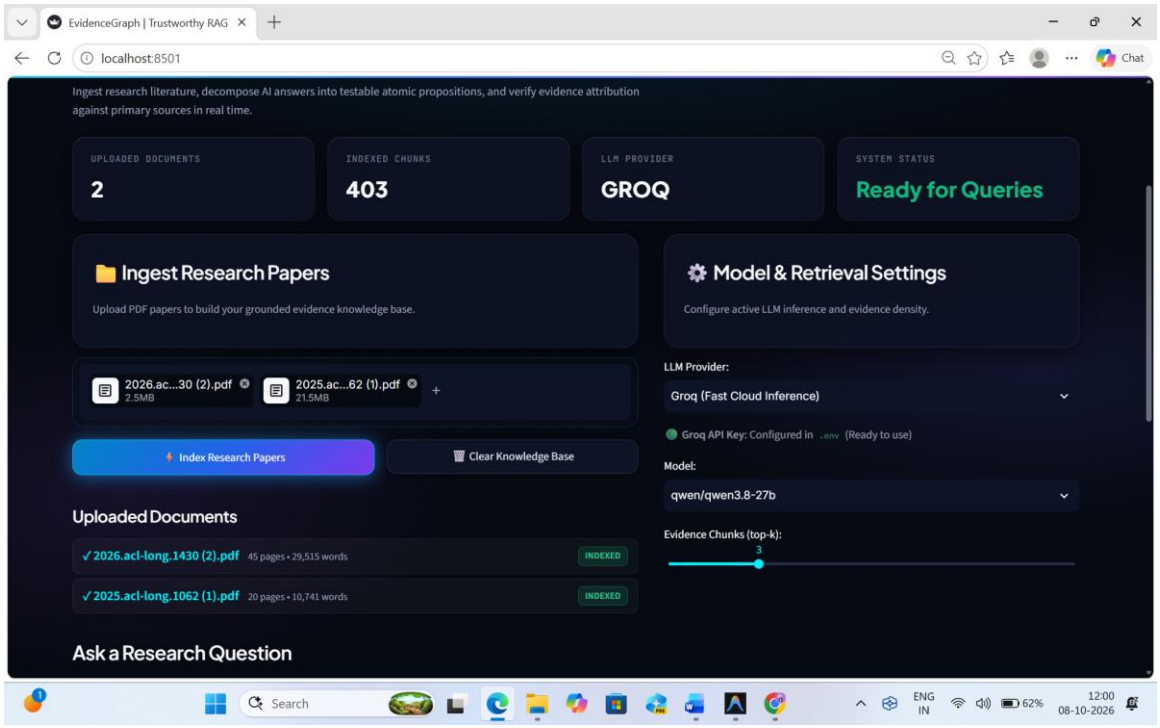


Figure 3. Research workspace showing document ingestion, indexed papers, and model and retrieval settings.

The verified information is organized into an interactive graph. The graph connects claims with evidence and source documents and represents relationships between evidence passages. The main representation is Claim → Evidence → Source → Relationship.

Scientific Research Papers → PDF Processing → Text Chunking → Embedding Generation → Vector Storage (ChromaDB) → User Query → Top-K Retrieval → RAG Generation → Claim Extraction → Verification → Relationship Analysis → EvidenceGraph

5. FINDINGS :

5.1 Evaluation Dataset

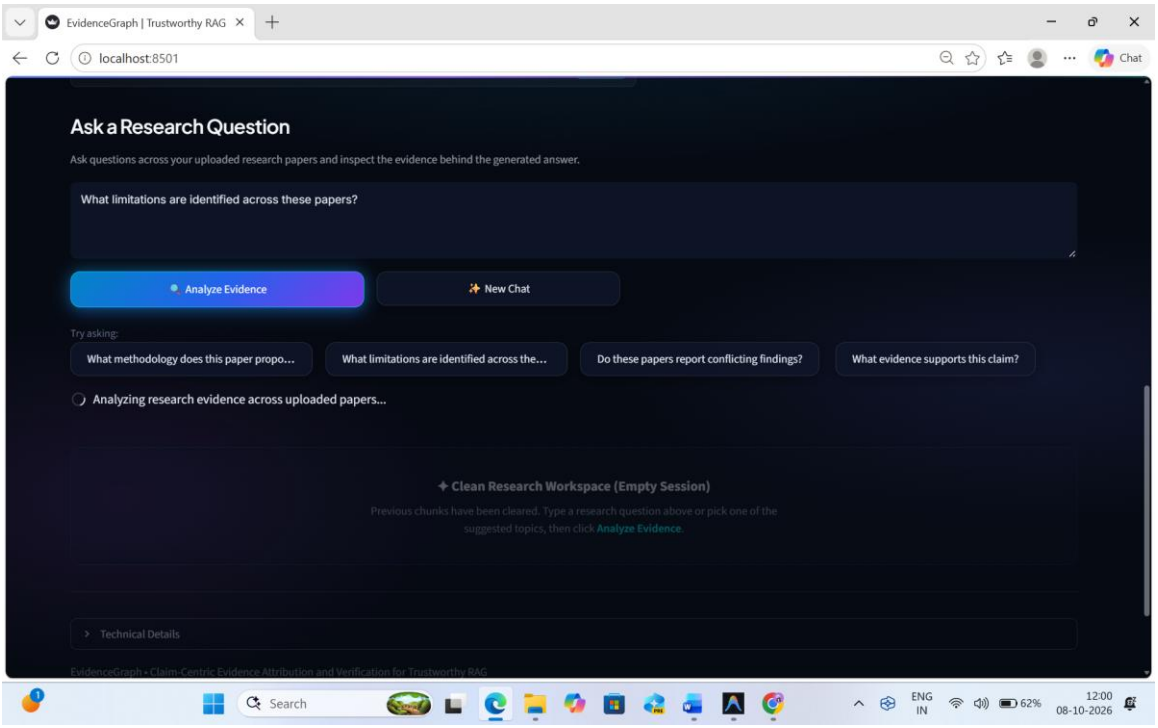


Figure 4. Research-question interface used to query the uploaded scientific papers.

Parameter	Value
Published scientific papers	3
Research questions	30
Questions per paper	10
Generated claims	96
Evidence chunks	478

5.2 Claim-Level Evidence Verification

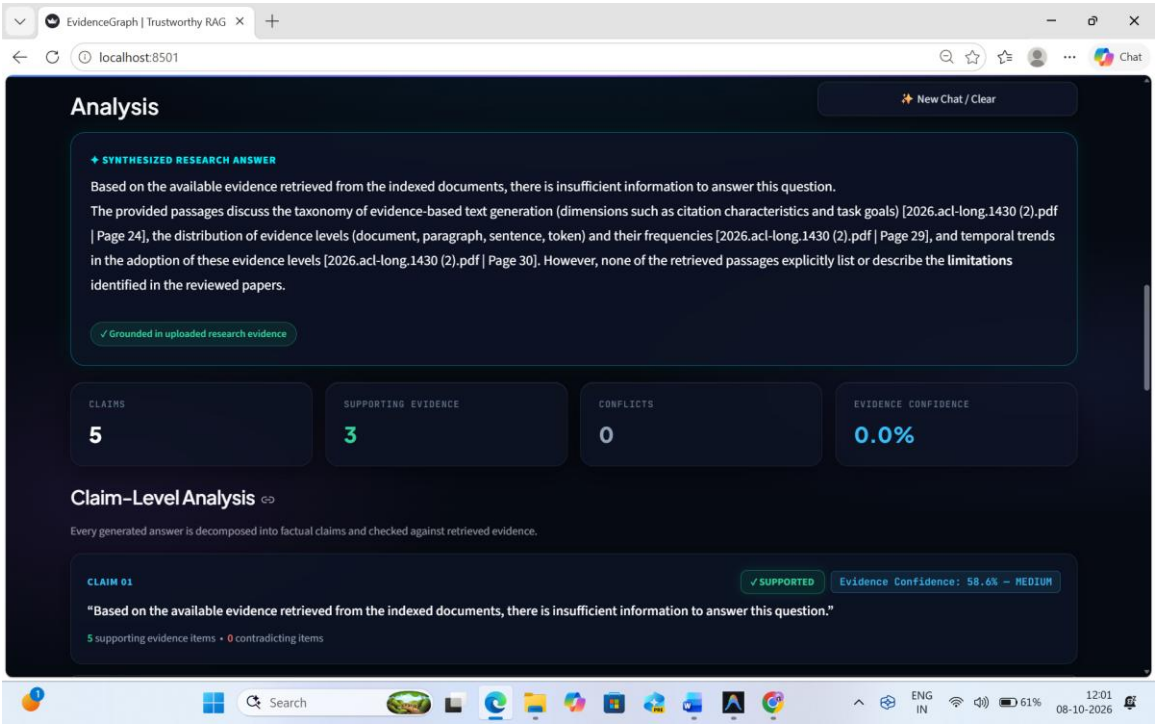
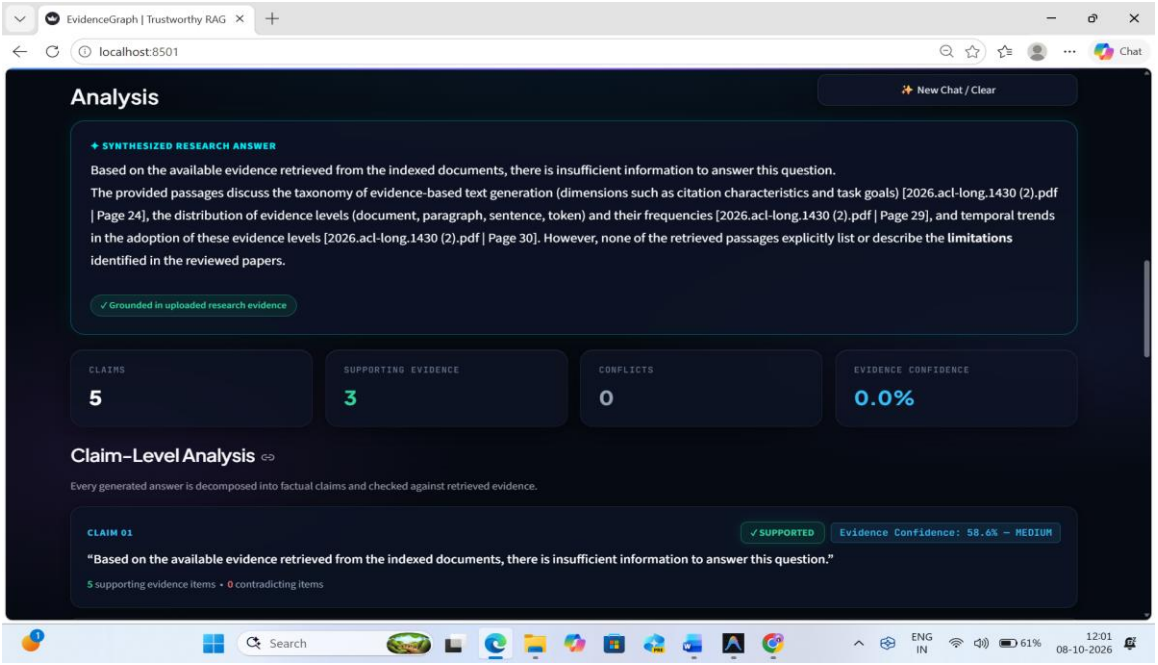


Figure 5. Analysis view showing the generated research answer and claim-level summary.



Out of the 96 generated claims, 60 were supported (62.5%), 20 were contradicted (20.8%), and 16 had insufficient evidence (16.7%). The 36 unsupported claims consist of the contradicted and insufficient-evidence cases.

Verification Category	Count	Percentage
Supported	60	62.5%
Contradicted	20	20.8%
Insufficient	16	16.7%
Unsupported	36	37.5%
Total	96	100%

The research-question and answer interface is shown in Figure 4, while the claim-level analysis is presented in Figures 5 and 6. The resulting evidence graph is shown in Figure 7.

5.3 Retrieval Results

Top-K	Exact Precision	Exact Recall	Target-Paper Recall
1	16.7%	16.7%	86.7%
3	13.3%	26.7%	96.7%
5	10.7%	33.3%	96.7%
10	8.0%	43.3%	100.0%

The system retrieved the correct target paper more consistently than the exact supporting page. Target-paper recall increased from 86.7% at Top-1 to 100% at Top-10.

5.4 Claim-Level Semantic Evaluation

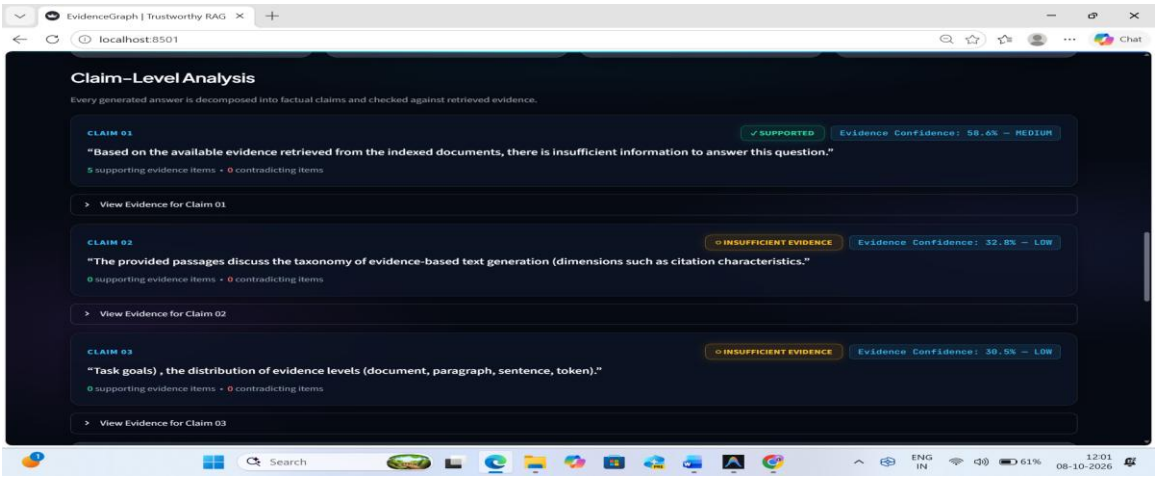


Figure 6. Claim-level analysis interface displaying claim classifications and evidence-confidence indicators.

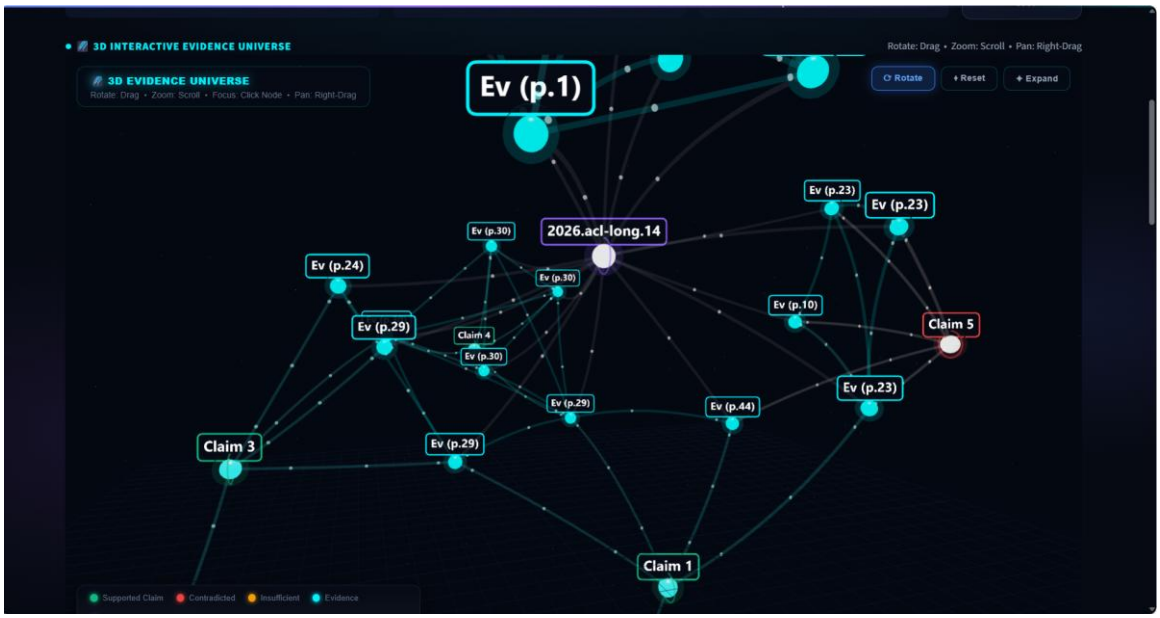


Figure 7. Interactive 3D evidence universe showing claims, evidence passages, and source documents as connected nodes.

- Token claim accuracy: 2.3%
- Average semantic similarity: 32.4%
- Semantic match rate: 5.0%

These values measure the generated claims and should not be treated as direct proof of factual truth.

5.5 Conflict Analysis

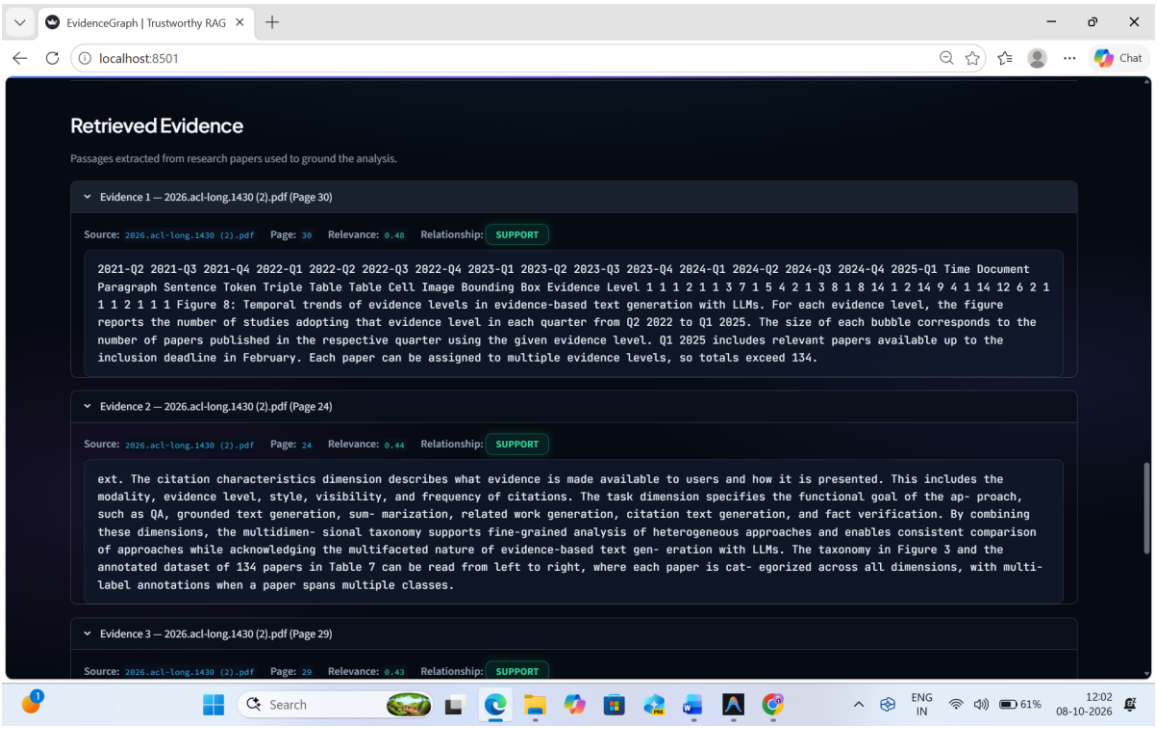
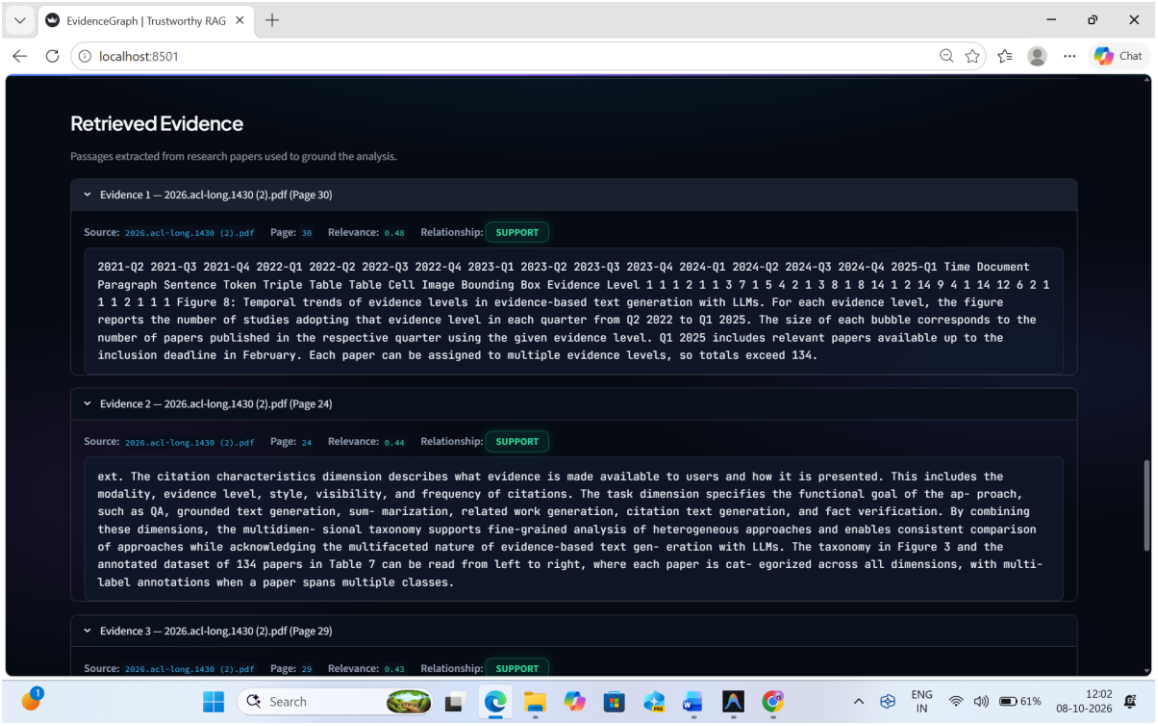


Figure 8. Retrieved evidence view showing source documents, page references, relevance, and evidence details.



The full EvidenceGraph system identified 94 evidence conflicts, while the baseline recorded zero conflicts under the implemented conflict-detection evaluation. The 94 detections corresponded to 80 unique chunk pairs. The audit found that some detections were related to rhetorical negation rather than direct factual contradiction. Therefore, these cases should be understood as detected evidence relationships requiring analysis, rather than 94 confirmed factual errors.

5.6 Ablation and Software Testing

The ablation study showed that the raw claim metrics remained the same across the baseline and full system in the implemented benchmark. The main additional capability appeared in the evidence-auditing components, especially conflict detection. The system was tested using 151 automated test cases. All 151 tests passed, with zero failed and zero skipped tests.

6. DISCUSSION / ANALYSIS

The results show that EvidenceGraph can retrieve the correct target research paper effectively at the document level. Target-paper recall reached 86.7% at Top-1 and increased to 100% at Top-10. Exact source-and-page retrieval was more difficult, with exact recall reaching 43.3% at Top-10.

The claim-level verification results provide more detail than judging an entire answer as correct or incorrect. Among the 96 claims, 60 were supported, 20 were contradicted and 16 had insufficient evidence.

The conflict-analysis component identified 94 evidence conflicts, but the audit showed that not every detected relationship represented a direct factual contradiction. Some cases were associated with rhetorical negation, so the underlying evidence should be inspected.

The ablation study shows that EvidenceGraph did not increase raw claim accuracy or semantic similarity in this benchmark. Its main value is the additional visibility into claim support, evidence provenance and relationships between retrieved passages.

7. CONCLUSION

This paper presented EvidenceGraph, a claim-centric framework for evidence attribution and verification in scientific Retrieval-Augmented Generation systems. The framework combines semantic retrieval, answer generation, atomic claim extraction, claim-to-evidence verification, evidence relationship analysis and source attribution.

The system was evaluated using 30 questions from three published scientific papers. The evaluation produced 96 generated claims and 478 evidence chunks. Among the generated claims, 62.5% were supported, 20.8% were contradicted and 16.7% had insufficient evidence. Target-paper recall reached 100% at Top-10 retrieval, and the conflict-analysis component identified 94 evidence conflicts under the implemented evaluation.

The system also passed all 151 automated tests. These results show that EvidenceGraph provides a structured way to

inspect individual claims, trace evidence to its source and examine relationships between retrieved evidence passages.

The results do not show that EvidenceGraph improves raw generation accuracy. The main contribution is claim-level auditing, evidence tracing and relationship analysis around the generated response.

8. LIMITATIONS

• The real-paper evaluation uses only three published scientific papers and 30 questions. A larger benchmark would provide a broader evaluation.

• Exact source-and-page retrieval remains challenging even when the target paper is retrieved.

• Claim verification depends on the quality and coverage of retrieved evidence.

• Conflict detection may identify rhetorical negation or textual patterns that are not direct factual contradictions.

• Semantic similarity and lexical claim metrics measure textual similarity and should not be treated as direct measurements of factual truth.

9. FUTURE SCOPE

• Increase the evaluation dataset with more scientific papers and research areas.

• Improve page-level retrieval so exact evidence passages are located more reliably.

• Improve conflict classification to distinguish direct contradictions from rhetorical negation.

• Extend the framework to tables, figures and charts in research papers.

• Support multimodal evidence verification for text and visual content.

• Improve adaptive retrieval based on claim complexity.

• Evaluate the framework on larger collections and more diverse scientific literature.

REFERENCES

1. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS 2020).

2. Guu, K., et al. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. International Conference on Machine Learning (ICML).

3. Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP 2020.

4. Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38.

5. Huang, L., et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Computing Surveys.

6. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. Proceedings of EMNLP 2023.

7. Bohnet, B., et al. (2022). Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. arXiv preprint arXiv:2212.08037.

8. Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. Proceedings of EMNLP 2023.

9. Honovich, O., et al. (2022). TRUE: Re-evaluating Factual Consistency Evaluation Approaches of Abstractive Summarization. Transactions of the Association for Computational Linguistics, 10, 160–175.

10. Thorne, J., et al. (2018). FEVER: A Large-scale Dataset for Fact Extraction and VERification. Proceedings of NAACL-HLT 2018, 809–819.

11. Chen, Y., et al. (2024). Benchmarking and Analyzing Knowledge Conflicts in Retrieval-Augmented Generation. Findings of the Association for Computational Linguistics.