

Evidence-Grounded Hallucination Detection and Correction in Retrieval-Augmented Large Language Models Using Claim-Level Verification

Bilal Khan

Assistant Professor, Project Management, Department of Management, India

Abstract - Retrieval-Augmented Generation has become an important approach for improving the factual reliability of large language models by providing external evidence during response generation. However, retrieval augmentation does not guarantee that every generated statement is supported by the retrieved material. Large language models can still produce unsupported, partially supported, or contradictory claims even when relevant evidence is available in the retrieval context. This problem is particularly important for systems expected to provide reliable factual answers because an otherwise accurate response may contain a small number of incorrect claims that are difficult to identify through response-level evaluation. This article examines evidence-grounded hallucination detection and correction using claim-level verification. The study develops a conceptual framework in which generated responses are decomposed into individual factual claims, each claim is matched against retrieved evidence, and unsupported or contradictory claims are subsequently subjected to evidence-based correction. The framework is informed by recent research on retrieval-augmented generation, factuality evaluation, hallucination detection, evidence attribution, and automated verification. Particular attention is given to evidence-bound factual repair approaches that separate semantic localization of factual errors from deterministic realization of evidence-grounded repairs. Existing research including FActScore, SelfCheckGPT, RAGTruth, RAGChecker, RAGAs, ARES, is integrated to establish the theoretical basis of the proposed approach. The analysis indicates that claim-level verification provides a more granular mechanism for identifying factual weaknesses than response-level evaluation, while evidence-constrained correction can reduce the risk of introducing new unsupported information during repair. The proposed framework provides a foundation for developing RAG systems that not only retrieve evidence but also verify, localize, correct, and transparently attribute individual factual claims.

Keywords: Retrieval-augmented generation, hallucination detection, claim-level verification, factuality, evidence grounding, large language models, factual correction, RAG, evidence attribution, trustworthy artificial intelligence.

I. INTRODUCTION

Large language models have demonstrated substantial capabilities in natural language understanding and generation. They can answer questions, summarize documents, generate reports, explain concepts, and perform complex language-based tasks. Despite their fluency, these systems can produce hallucinations in which generated statements are factually incorrect, unsupported, or inconsistent with available evidence. Such behaviour presents a major challenge for applications in which factual reliability is essential.

Retrieval-Augmented Generation was introduced as a means of improving knowledge-intensive language generation by connecting language models to external information sources. Lewis *et al.* (2020) established the retrieval-augmented generation paradigm by combining parametric language models with non-parametric external memory. The approach enables a language model to retrieve relevant information before producing a response and has subsequently become a major architecture for knowledge-grounded language applications.

Retrieval, however, does not eliminate hallucination. A model may retrieve relevant documents but fail to use them correctly. It may combine information from multiple passages incorrectly, introduce details that do not appear in the retrieved evidence, misunderstand the retrieved material, or generate a claim that conflicts with the available sources. Niu *et al.* (2024) demonstrated this problem through RAGTruth, a large hallucination corpus containing naturally generated RAG responses with detailed human annotations. Their research shows that unsupported and contradictory claims can remain present even when retrieval is incorporated into the generation process.

This problem becomes more difficult when responses contain many factual statements. A response may be mostly correct while containing one or two unsupported claims. Response-level labels such as correct or incorrect can therefore conceal important differences in factual reliability. Min *et al.* (2023) introduced FActScore to address this issue by decomposing long-form generations into atomic facts and measuring the proportion supported by reliable sources. Their work established the importance of fine-grained factual evaluation.

Claim-level verification extends this principle by treating individual claims as the fundamental units of factual analysis. Rather than asking whether an entire response is reliable, a verification system asks whether each factual claim is supported by the available evidence, contradicted by the evidence, or insufficiently supported.

II. BACKGROUND OF THE STUDY

The reliability of large language models is closely connected to how they acquire and use factual information. Conventional language models store knowledge within their parameters, but this knowledge may be incomplete, outdated, difficult to inspect, or inconsistently recalled. Retrieval-Augmented Generation addresses some of these limitations by introducing an external knowledge source.

Lewis *et al.* (2020) described RAG as a combination of parametric and non-parametric memory. The retriever provides information from an external collection, while the generator uses the retrieved material to formulate the response. This architecture provides an important mechanism for grounding generation in information outside the model's parameters.

The presence of retrieved evidence, however, does not establish that a generated claim is grounded. A response can contain evidence-supported statements alongside unsupported assertions. It can also contain statements that appear plausible but cannot be established from the retrieved documents.

The distinction between retrieval quality and generation faithfulness has therefore become increasingly important. RAGChecker, developed by Ru *et al.* (2024), provides a fine-grained framework for evaluating both retrieval and generation components. The framework includes diagnostic metrics designed to identify different sources of RAG performance variation and demonstrates the importance of evaluating RAG systems at multiple levels rather than treating them as a single black box.

RAGTruth provides another important contribution by demonstrating that hallucinations can persist in RAG outputs even when the generated responses are conditioned on retrieved evidence. The corpus contains nearly 18,000 naturally generated RAG responses with manual annotations at case and word levels, providing a resource for studying unsupported and contradictory content.

These developments support a shift toward more granular factual analysis. Instead of evaluating only complete responses, researchers increasingly examine sentences, claims, evidence spans, and relationships between generated statements and source material.

III. LITERATURE REVIEW

Research on hallucination detection has focused on identifying factual inconsistency in generated text. SelfCheckGPT, developed by Manakul *et al.* (2023), introduced a sampling-based method for black-box hallucination detection. The approach examines consistency across multiple sampled responses, with greater disagreement interpreted as a potential indication of hallucination. Their experiments demonstrated the ability to identify factual and non-factual sentences without requiring access to model probabilities or an external knowledge database.

Although SelfCheckGPT provides an important detection mechanism, it does not inherently establish whether a specific claim is supported by authoritative external evidence. This limitation becomes significant in RAG systems because the retrieved evidence itself can provide a basis for verification.

FActScore addresses factuality at a more granular level by decomposing generated responses into atomic facts. Min *et al.* (2023) demonstrated that long-form responses often contain a mixture of supported and unsupported information, making response-level factuality judgments insufficient. Their approach measures the proportion of atomic facts supported by reliable knowledge sources.

The development of RAGTruth further established the need for dedicated RAG hallucination datasets. Niu *et al.* (2024) showed that RAG systems can still generate unsupported and contradictory content despite access to retrieved documents. Their corpus provides detailed annotations that allow researchers to examine hallucination at a fine-grained level.

RAGChecker extends this evaluation perspective by examining both retrieval and generation. Ru *et al.* (2024) demonstrated that RAG systems contain multiple interacting components and that system-level evaluation can conceal

important differences in retrieval quality, context utilization, and answer faithfulness.

Automated RAG evaluation has also become an important research direction. Es *et al.* (2024) introduced RAGAs, which evaluates RAG systems across retrieval and generation dimensions without requiring traditional ground-truth reference answers. The framework emphasizes the importance of assessing the relevance of retrieved contexts and the faithfulness of generated responses.

Saad-Falcon *et al.* (2024) developed ARES as an automated evaluation framework for RAG systems. ARES evaluates context relevance, answer faithfulness, and answer relevance and uses prediction-powered inference to reduce dependence on large quantities of human annotation.

These evaluation approaches demonstrate that factual reliability in RAG systems cannot be captured by a single metric. Retrieval relevance, evidence utilization, factual support, and answer relevance represent related but distinct properties.

Verification-based generation provides another direction. Dhuliawala *et al.* (2024) proposed Chain-of-Verification, where a model generates an initial answer, develops verification questions, answers them independently, and then produces a revised response. Their findings show that explicit verification can reduce hallucination.

Recent research has also examined RAG systems in the presence of conflicting evidence. Ge *et al.* (2025) developed the CONFACT dataset to evaluate RAG-based fact-checking when sources disagree. Their findings show that RAG systems can struggle when evidence differs in credibility and content, demonstrating that retrieval alone does not resolve factual conflict.

Russo *et al.* (2025) evaluated RAG-based pipelines for professional fact-checking and found that system performance varies according to retrieval methods, knowledge-base heterogeneity, model size, and the nature of the generated verdict. Their research demonstrates the importance of evaluating RAG systems under realistic fact-checking conditions rather than assuming that retrieved information automatically produces reliable conclusions.

IV. METHODOLOGY

This study adopts a conceptual methodology based on a structured synthesis of research on RAG, hallucination detection, factuality evaluation, claim-level verification, evidence attribution, and factual correction.

The proposed methodology treats a generated response as a collection of factual claims rather than as a single indivisible object. The response is first analyzed to identify statements that make factual assertions. Claims that are purely stylistic or conversational can be separated from claims that require factual verification.

Each factual claim is then compared with the retrieved evidence. The purpose of this stage is to determine whether the evidence supports the claim, contradicts it, or fails to provide sufficient information for verification Agha *et al.* (2025).

The framework distinguishes between retrieval failure and generation failure. Retrieval failure occurs when the information required to verify a claim is absent or inadequately retrieved. Generation failure occurs when sufficient evidence is available but the model produces a statement that is unsupported or inconsistent with that evidence.

This distinction is important because the appropriate correction depends on the type of failure. If the relevant evidence was not retrieved, additional retrieval may be necessary. If the evidence was retrieved but the claim was generated incorrectly, the system can proceed directly to evidence-grounded repair.

The correction stage follows the principle of separating semantic localization from factual realization proposed by Rajendran and Singaravelu (2026). The system first identifies which evidence supports the intended factual content and then restricts the final correction to information that can be directly justified by the selected evidence.

The final response is reconstructed after verification and correction. Claims that cannot be adequately supported are not treated as factual simply because they appear plausible. The system can instead preserve uncertainty or decline to assert information that lacks sufficient evidence.

V. RESULTS

The conceptual analysis indicates that claim-level verification provides a more granular mechanism for hallucination detection than response-level evaluation. By decomposing responses into individual factual claims, the system can identify localized errors while preserving information that is already supported by evidence.

The analysis further indicates that retrieval alone is insufficient for factual reliability. Evidence must be explicitly connected to the claims generated from it. This finding is

consistent with RAGTruth, which demonstrates that unsupported and contradictory content remains possible in retrieval-augmented responses.

A further finding is that evidence-bound correction can address an important weakness of conventional hallucination mitigation. A system that freely rewrites an incorrect statement may generate another unsupported statement. Separating evidence localization from deterministic realization provides a mechanism for limiting the content of corrections to information that can be grounded in the retrieved evidence.

The analysis also indicates that conflicting evidence represents a significant challenge for automated verification. Source credibility, evidence quality, retrieval ranking, and disagreement between documents can all influence the final decision.

Another finding is that reliable verification requires evaluation at multiple levels. Retrieval relevance, evidence support, claim correctness, and final answer faithfulness represent different dimensions of system performance.

Finally, the analysis indicates that correction should be treated as a separate stage from detection. A system may successfully identify a hallucination but still fail to repair it correctly. Reliable RAG therefore requires a pipeline in which detection, evidence localization, correction, and post-correction verification are explicitly connected.

VI. DISCUSSION

The findings support a transition from retrieval-augmented generation toward evidence-verified generation. RAG has substantially improved the ability of language models to access external information, but retrieval alone does not guarantee factual reliability.

The literature demonstrates that hallucination can persist at multiple levels. SelfCheckGPT shows that generated statements can be inconsistent across samples, FActScore demonstrates that long-form responses contain mixtures of supported and unsupported atomic facts, and RAGTruth demonstrates that unsupported claims can remain present in RAG outputs.

These findings establish the need for more granular verification.

Claim-level verification provides an appropriate foundation because factual errors can be isolated without discarding entire responses. This is particularly useful for

long-form answers in which most statements may be supported while a smaller number require correction.

The evidence-bound repair approach of Rajendran and Singaravelu (2026) is particularly significant because it addresses the difference between finding the correct evidence and authoring a corrected statement. Their separation of semantic localization from deterministic realization provides a mechanism for reducing the generative model's freedom during the correction stage.

The approach also complements Chain-of-Verification. CoVe demonstrates the value of structured verification, while evidence-bound repair places stronger restrictions on how the corrected information is realized. Together, these directions suggest that reliable factual correction requires both verification and controlled generation.

The findings concerning conflicting evidence are equally important. Ge *et al.* (2025) demonstrate that RAG-based fact-checking can struggle when evidence differs in credibility. A claim-level verifier must therefore consider not only whether evidence exists but also whether the evidence is sufficiently authoritative to justify the claim.

The evaluation literature further indicates that no single metric can fully characterize RAG reliability. RAGAs, ARES, and RAGChecker each emphasize different dimensions of retrieval and generation quality. Their combined lessons suggest that claim-level verification should be evaluated using multiple complementary criteria rather than a single factuality score.

VII. CONCLUSION

Retrieval-Augmented Generation provides large language models with access to external information, but retrieval alone does not eliminate hallucination. Generated responses can still contain unsupported, partially supported, or contradictory claims even when relevant evidence is present in the retrieval context.

Evidence-grounded claim-level verification provides a promising approach to this problem because it treats individual factual claims as the fundamental units of analysis. This allows systems to identify localized factual errors, connect claims to supporting evidence, distinguish retrieval failure from generation failure, and preserve information that is already adequately supported.

The literature reviewed in this study demonstrates the development of increasingly granular approaches to factuality. SelfCheckGPT established a sampling-based method for

hallucination detection, FActScore demonstrated the importance of atomic factual evaluation, RAGTruth provided a large corpus for analyzing hallucinations within RAG systems, and RAGChecker, RAGAs, and ARES introduced increasingly sophisticated methods for evaluating retrieval and generation quality.

The correction problem requires an additional layer of control. Chain-of-Verification demonstrates that structured verification can reduce hallucination, while recent research provides a direct framework for evidence-bound factual repair by separating semantic localization from deterministic realization. This separation is particularly valuable because a system that freely generates a correction can itself introduce unsupported information.

The resulting framework positions factual reliability as a continuous process involving retrieval, claim identification, evidence attribution, verification, correction, and post-correction validation. Such a process is more demanding than conventional RAG generation, but it provides a stronger basis for trustworthy language-model applications.

Future research should validate this framework experimentally across diverse domains, retrieval environments, languages, and evidence conditions. Greater emphasis should also be placed on conflicting evidence, temporal validity, source credibility, and automated verification reliability. The continued development of claim-level verification and evidence-bound correction can contribute to RAG systems that are not only more informative but also more transparent, auditable, and resistant to factual hallucination.

REFERENCES

- [1] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578. <https://doi.org/10.18653/v1/2024.findings-acl.212>
- [2] 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>
- [3] Ge, Z., Wu, Y., Chin, D. W. K., Lee, R. K.-W., & Cao, R. (2025). Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. *In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence* (pp. 9656–9664). <https://doi.org/10.24963/ijcai.2025/1073>
- [4] Godbole, A., & Jia, R. (2025). Verify with caution: The pitfalls of relying on imperfect factuality metrics. *Findings of the Association for Computational Linguistics: ACL 2025*, 22889–22912. <https://doi.org/10.18653/v1/2025.findings-acl.1175>
- [5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *In Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474).
- [6] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- [7] Agha, C. W., Mokwenye, I. I., Atikpo, E., & Nwulu, N. (2025). Gumbel Distribution-Based Flood Frequency Analysis for Coastal Flood Risk Assessment in Ugborodo Community, Niger Delta, Nigeria. *Environmental Technology and Science Journal*, 16(1), 133-143. <https://doi.org/10.4314/etsj.v16i1.13>
- [8] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- [9] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K.-S., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10862–10878). *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2024.acl-long.585>
- [10] Rajendran, S., & Singaravelu, D. (2026). Evidence-bound factual repair in retrieval-augmented LLM answers: Separating semantic localization from deterministic realization (Version 1) [Preprint].

Preprints.org.

<https://doi.org/10.20944/preprints202609.0490.v1>

- [11] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv*. <https://arxiv.org/abs/2408.08067>
- [12] Russo, D., Menini, S., Staiano, J., & Guerini, M. (2025). Face the facts! Evaluating RAG-based pipelines for professional fact-checking. In *Proceedings of the 18th International Natural*

Language Generation Conference (pp. 846–865). Association for Computational Linguistics. <https://aclanthology.org/2025.inlg-main.50/>

- [13] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers). Association for Computational Linguistics. <https://aclanthology.org/2024.naacl-long.20/>

Citation of this Article:

Bilal Khan. (2026). Evidence-Grounded Hallucination Detection and Correction in Retrieval-Augmented Large Language Models Using Claim-Level Verification. *International Current Journal of Engineering and Science (ICJES)*, 5(9), 19-24. Article DOI: <https://doi.org/10.47001/ICJES/2026.509003>
