CEGI-S: A Counterfactual Evidence-Gated Intelligence Framework for Reliable and Self-Correcting Large Language Models

Maheshwari Mahesh Vidhate[1], Gayatri Namdev Taur[2], Chaitanya Bhanudas Kale[3], Monika Rupesh Ugle[4]
[1,2,3]*Student, Department of Science and Computer Science, MIT Arts, Commerce and Science College, Alandi (D), Pune, Maharashtra, India*
[4]*Professor, Department of Science and Computer Science, MIT Arts, Commerce and Science College, Alandi (D), Pune, Maharashtra, India*
doi.org/10.64643/ijirt.208549-459

***Abstract*—Large Language Models (LLMs) are increasingly being used for question answering, education, research assistance, document analysis, and other knowledge-based applications. Their ability to generate fluent and meaningful responses makes them useful in many practical situations. However, LLMs can also produce information that is incorrect, incomplete, or not supported by reliable evidence. These unsupported outputs, commonly called hallucinations, can reduce user trust and become a serious problem in applications where factual accuracy is important.**

Retrieval-Augmented Generation (RAG) improves LLM responses by providing external information during generation. However, the retrieval of relevant documents does not necessarily mean that the final answer is fully supported by those documents. Retrieved sources may be incomplete, contradictory, outdated, or only partially relevant to the claims generated by the model. Therefore, a separate verification mechanism is required before the final response is delivered. This paper proposes CEGI-S (Counterfactual Evidence-Gated Intelligence with Support–Conflict Scoring), an evidence-aware framework for improving the reliability of LLM-generated responses. The framework introduces a verification layer between response generation and final answer delivery. It divides the generated response into individual claims, connects the claims with retrieved evidence, checks agreement and conflict among sources, and uses counterfactual challenges to examine the stability of the generated answer.

These signals are combined into a proposed Support–Conflict Evidence Score (SCES). Based on SCES, CEGI-S can select one of three actions: accept the response when sufficient evidence is available, retrieve additional evidence and regenerate the response when the evidence is uncertain or incomplete, or abstain when a reliable answer cannot be established. The proposed framework will be evaluated against conventional LLM and RAG-based approaches using factuality, hallucination rate, evidence support, answer relevance, calibration, abstention quality, latency, and computational cost. The main aim of this research is to investigate whether combining multiple evidence-related signals can provide a practical decision layer for more reliable and responsible generative AI systems.

***Index Terms*—Large Language Models, Generative AI, Retrieval-Augmented Generation, Counterfactual Reasoning, Hallucination Detection, Evidence Verification, Responsible AI, Self-Correcting AI, Intelligent Systems.**

I. INTRODUCTION

Artificial Intelligence has developed from systems designed for specific tasks into systems that can understand and generate natural language. Large Language Models have become an important part of this development. Modern LLMs can answer questions, summarize documents, generate code, assist students, and support researchers. These capabilities have resulted in the rapid adoption of generative AI in many knowledge-intensive applications.

Although LLMs are powerful, they do not always produce factually reliable answers. A model may generate a response that is grammatically correct and convincing even when some of its statements are incorrect or unsupported. This behavior is generally referred to as hallucination. The problem becomes more serious when users depend on the generated information for learning, research, decision-making, or other applications where correctness is important.

Retrieval-Augmented Generation was introduced to provide language models with external information during generation [1]. Instead of depending only on knowledge stored in the model parameters, a RAG system retrieves relevant documents or passages and provides them as additional context. This approach can improve grounding and help the model use information that may not be available in its internal knowledge.

However, retrieval alone does not guarantee a reliable answer. A retrieved passage may be related to the query but may not actually support the claim generated by the model. Different documents may also contain conflicting information. In addition, a model may combine information from several sources and produce a conclusion that is not directly supported by any of them.

Several approaches have been proposed to improve this situation. Self-RAG introduces retrieval and self-reflection into the generation process [2]. Corrective RAG (CRAG) evaluates retrieved information and performs corrective actions when retrieval quality is insufficient [3]. Evaluation frameworks such as ARES and RAGAS provide methods for assessing retrieval relevance, answer faithfulness, and answer quality [4], [5].

Recent research has also investigated more specific forms of evidence verification and counterfactual reasoning. EAEV uses evidence alignment and counterfactual stability for hallucination detection in RAG systems [11]. CF-RAG investigates counterfactual reasoning and conflicting evidence in retrieval-augmented generation [12]. SURE-RAG studies evidence sufficiency, uncertainty, conflict, and selective answering [13]. These developments show that reliability in RAG systems requires more than simply retrieving documents.

This research proposes CEGI-S as an additional decision layer placed after candidate response generation. The framework focuses on combining several signals instead of relying on a single confidence value.

The overall process is:

Query → Evidence Retrieval → Candidate Answer → Claim Analysis → Evidence Verification → Counterfactual Challenge → SCES Calculation → Decision

The final decision has three possible outcomes:

1. Accept the answer when sufficient evidence supports it.
2. Retrieve and regenerate when the evidence is incomplete, uncertain, or conflicting.
3. Abstain when sufficient reliable evidence cannot be established.

The purpose of CEGI-S is not to claim that hallucinations can be completely eliminated. Instead, the research investigates whether an evidence-based decision mechanism can reduce unsupported responses and improve the reliability of LLM-based systems.

II. PROBLEM STATEMENT

LLMs can produce fluent answers even when the information needed to support those answers is unavailable, incomplete, or contradictory. Although RAG provides external evidence, retrieved information does not automatically guarantee that every generated claim is supported.

The problem addressed in this research is:

How can an LLM-based intelligent system automatically evaluate whether its generated response is sufficiently supported by evidence and decide whether to accept, regenerate, or abstain from providing the response?

The study investigates whether combining evidence support, source consistency, counterfactual stability, uncertainty, and evidence conflict can provide a useful decision mechanism for response verification.

III. RESEARCH GAP

Existing research has addressed different parts of the LLM reliability problem. RAG provides external knowledge for generation [1]. Self-RAG introduces retrieval and self-reflection [2], while CRAG evaluates retrieval quality and performs corrective retrieval when necessary [3]. ARES and RAGAS provide automated approaches for evaluating different aspects of RAG systems [4], [5]

Hallucination and factuality benchmarks such as RAGTruth, FEVER, and TruthfulQA also provide resources for systematic evaluation [6]– [8].

Recent work has moved further toward evidence verification. EAEV studies entity-level evidence

alignment and counterfactual stability for hallucination detection [11]. CF-RAG investigates counterfactual reasoning and conflicting evidence in RAG [12]. SURE-RAG addresses evidence sufficiency, conflict, uncertainty, and selective answering [13].

Therefore, the research gap should not be described as a complete absence of these techniques. Instead, this study focuses on investigating a response-level evidence-gating mechanism that combines multiple verification signals into a single decision score and directly connects the score with three possible actions: accept, regenerate, or abstain.

The research question is:

Can a Support–Conflict Evidence Score that combines evidence support, source consistency, evidence conflict, counterfactual stability, and uncertainty provide a useful mechanism for deciding whether an LLM-generated response should be trusted?

CEGI-S is therefore proposed as a framework that can potentially be added to existing LLM and RAG systems rather than replacing them.

IV. OBJECTIVES

A. Main Objective

To design and evaluate an evidence-gated framework for improving the reliability of LLM-generated responses through claim-level evidence verification and counterfactual checking.

B. Specific Objectives

1. To design the architecture of the proposed CEGI-S framework.
2. To retrieve relevant external evidence for user queries.
3. To generate candidate responses using an LLM and retrieved evidence.
4. To divide generated responses into individual claims.
5. To connect generated claims with relevant supporting or conflicting evidence.
6. To measure consistency among retrieved sources.
7. To identify conflicts between evidence sources.
8. To introduce counterfactual challenges for testing answer stability.
9. To estimate uncertainty associated with generated responses.
10. To formulate the proposed Support–Conflict Evidence Score.
11. To use SCES for selecting accept, regenerate, or abstain decisions.
12. To compare CEGI-S with conventional LLM and RAG-based approaches.
13. To study the trade-off between improved reliability and additional computational cost.

V. PROPOSED CEGI-S FRAMEWORK

CEGI-S consists of seven major stages:

1. Query Understanding
2. Evidence Retrieval
3. Candidate Answer Generation
4. Claim-Level Evidence Mapping
5. Counterfactual Challenge
6. SCES Calculation
7. Decision and Self-Correction

A. Query Understanding

The process begins when the user enters a natural-language query.

For example:

“Does edge computing reduce latency compared with centralized cloud processing?”

The query analysis stage identifies the important concepts and determines the type of information required to answer the question.

The system may identify concepts such as edge computing, centralized cloud processing, network latency, and comparative performance.

The purpose of this stage is to prepare the query for effective evidence retrieval.

B. Evidence Retrieval

The system retrieves relevant passages from an external document collection.

A hybrid retrieval strategy may combine keyword-based and semantic retrieval techniques, such as:

- BM25 keyword retrieval
- Dense semantic retrieval
- Vector similarity search

Let the retrieved evidence be represented as:

$E = \{e_1, e_2, ..., e_k\}$

where each e_i represents a retrieved evidence passage.

The retrieved passages are ranked according to their relevance to the query.
Dense retrieval methods such as HyDE can also be considered for improving retrieval when direct query matching is insufficient [10].
However, CEGI-S does not treat retrieval relevance as proof of correctness. The retrieved information is considered potential evidence and is checked against the claims generated by the model.

C. Candidate Answer Generation
The LLM receives the user query and retrieved evidence and generates an initial response.
The candidate response can be represented as:
A = Generate (Q, E)
where:
- Q represents the user query,
- E represents the retrieved evidence, and
- A represents the candidate answer.

The candidate answer is not immediately returned to the user. It first passes through the verification stages.

VI. CLAIM-LEVEL EVIDENCE MAPPING

The candidate answer is divided into individual claims:
$A \rightarrow \{C_1, C_2, ..., C_n\}$
Each claim is compared with the retrieved evidence.
The system determines whether each claim is:
- Supported by the evidence,
- Contradicted by the evidence,
- Partially supported, or
- Not sufficiently supported.

This claim-level approach is useful because a single response may contain several correct statements and one unsupported statement. Evaluating the complete response using only one confidence value may hide such differences.
For each claim, the framework records the relationship between the claim and available evidence. These relationships are later used to calculate the support, consistency, and conflict components of SCES.

VII. COUNTERFACTUAL CHALLENGE

The counterfactual stage is used to test whether the generated conclusion remains stable when important assumptions are challenged.

For example, suppose the generated response states:
"Edge computing reduces latency because processing is performed closer to the user."

A counterfactual challenge could ask:
"Would the same conclusion remain valid if the edge server were overloaded or if the connection to the edge server were poor?"
The objective is not to intentionally create a false answer. The objective is to examine whether the original conclusion depends on assumptions that are not sufficiently supported by the available evidence.
The original response and the response generated under the counterfactual condition are compared.

The resulting measure is represented as:
K = Counterfactual Stability
A higher K indicates that the main conclusion remains reasonably stable under appropriate counterfactual challenges. A lower K indicates that the conclusion may be sensitive to unsupported assumptions.

VIII. SUPPORT–CONFLICT EVIDENCE SCORE

The main decision component of CEGI-S is the proposed Support–Conflict Evidence Score (SCES).

The score combines five signals:
- S = Evidence Support
- C = Cross-Source Consistency
- K = Counterfactual Stability
- U = Uncertainty
- F = Evidence Conflict

A normalized formulation is proposed as:
$SCES = w_1S + w_2C + w_3K - w_4U - w_5F$
where:
- w_1, w_2, w_3, w_4, and w_5 are weights,
- each component is normalized to a common scale.

The positive components increase the reliability score, while uncertainty and evidence conflict reduce it.
The weights should not be selected only on intuition. They should be tuned using a validation dataset and evaluated on separate test data.

The proposed formulation is intended as a starting point for experimentation rather than a claim that these exact weights are optimal.

IX. DECISION AND SELF-CORRECTION

After calculating SCES, the framework selects one of three actions.

A. Accept

If the score is above the acceptance threshold:

$SCES \geq T_a$

the response is considered sufficiently supported and is returned to the user.

B. Retrieve and Regenerate

If:

$T_r \leq SCES < T_a$

the evidence is considered uncertain or incomplete. The system performs additional retrieval and generates a new candidate response.

C. Abstain

If:

$SCES < T_r$

and additional retrieval does not provide adequate evidence, the system abstains instead of providing an unsupported answer.

This three-way decision is an important part of CEGI-S because the system is not forced to generate an answer when sufficient evidence is unavailable.

X. PROPOSED ALGORITHM

Algorithm 1: CEGI-S

Input: User query Q

Output: Verified answer A or abstention

1. Receive user query Q.
2. Analyze the information requirements of Q.
3. Retrieve the top-k relevant evidence passages.
4. Generate a candidate answer A.
5. Divide A into claims $C_1, C_2, ..., C_n$.
6. Map each claim to relevant evidence.
7. Calculate evidence support S.
8. Measure consistency among supporting sources.
9. Identify conflicting evidence F.
10. Generate suitable counterfactual challenges.
11. Compare the original and counterfactual responses.
12. Calculate counterfactual stability K.
13. Estimate uncertainty U.
14. Calculate SCES.
15. Compare SCES with the decision thresholds.
16. Accept the response if the score is sufficiently high.
17. Retrieve additional evidence and regenerate if the score is intermediate.
18. Abstain if sufficient evidence cannot be established.
19. Return the final response or abstention.

XI. RESEARCH METHODOLOG

The study will follow an experimental methodology to determine whether CEGI-S improves the reliability of LLM-generated responses.

A. Baseline Systems

CEGI-S will be compared with the following systems:

1. Conventional LLM without external retrieval.
2. Standard RAG.
3. Self-RAG or a comparable self-reflective method.
4. Corrective RAG or a comparable corrective retrieval method.
5. Proposed CEGI-S.

The same or comparable model and retrieval conditions should be maintained where possible so that the comparison is meaningful.

B. Datasets

The study can use established datasets including

- RAGTruth for hallucination-oriented evaluation [6].
- FEVER for fact verification [7].
- TruthfulQA for evaluating truthful answering [8].

These datasets cover different aspects of factual reliability and can help evaluate the proposed framework from multiple perspectives.

C. Evaluation Metrics

1) Factuality

Measures whether the information generated by the system is factually correct.

2) Hallucination Rate

Measures the frequency of unsupported or incorrect claims.

3) Evidence Support

Measures how strongly the generated claims are supported by retrieved evidence.

4) Answer Relevance

Measures whether the generated response directly addresses the user's query.

5) Calibration
Measures the relationship between system confidence and actual correctness.

6) Abstention Quality
Measures whether the system abstains appropriately when reliable evidence is unavailable.

7) Latency
Measures the additional time required by verification and regeneration.

8) Computational Cost
Measures additional retrieval operations, LLM calls, token usage, and other computational requirements.
RAG evaluation frameworks such as ARES and RAGAS can provide useful reference points when selecting evaluation procedures [4], [5].

XII. EXPECTED RESULTS

Since CEGI-S is proposed as a framework requiring experimental validation, numerical results should not be invented before implementation.
The expected evaluation is that CEGI-S may reduce unsupported responses compared with a basic LLM and standard RAG system because it adds claim-level evidence verification and a final evidence-based decision stage.

However, this improvement may come with additional computational cost because the framework requires extra verification steps and possibly additional LLM calls.

The final experimental study should report actual values for:

- Factuality,
- Hallucination rate,
- Evidence support,
- Relevance,
- Calibration,
- Abstention quality,
- Latency, and
- Computational cost.

The results should be presented using tables and graphs after the system has been implemented and tested.

XIII. DISCUSSION

CEGI-S approaches LLM reliability from a decision-making perspective. The main idea is that retrieving a relevant document should not automatically be treated as sufficient evidence for accepting a generated answer. The claim-level verification stage allows the system to examine individual statements instead of evaluating the response only as a whole. This can help identify unsupported claims that might otherwise be hidden inside an otherwise reasonable response.
Evidence conflict is another important part of the proposed framework. Different sources may provide different conclusions, especially when information is incomplete or when sources are based on different conditions. CEGI-S treats unresolved conflict as a reason to reduce confidence rather than automatically selecting one source.
The counterfactual stage provides an additional challenge to the candidate response. If a response changes significantly when a relevant assumption is modified, this may indicate that the original conclusion depends on conditions that require further verification.
At the same time, the framework introduces additional computational requirements. Claim decomposition, evidence verification, counterfactual generation, and additional retrieval can increase latency and resource consumption. Therefore, reliability should be evaluated together with efficiency.
Recent research confirms that evidence verification, counterfactual reasoning, and selective answering are active areas of RAG research. EAEV focuses on evidence-aligned entity verification and counterfactual stability, CF-RAG investigates counterfactual reasoning and conflicting evidence, and SURE-RAG studies evidence sufficiency and selective decisions.
For this reason, the contribution of CEGI-S should be established through controlled experiments and comparison with appropriate recent baselines rather than through a claim of completely new individual components.

XIV. LIMITATIONS

The proposed framework has several expected limitations.
First, its performance depends on the quality of the retrieved evidence. If the required information is

absent from the knowledge source, CEGI-S may not be able to verify the answer.

Second, claim decomposition can introduce errors. If the generated answer is divided incorrectly, the subsequent evidence verification may also be affected.

Third, counterfactual challenge generation may require additional model calls. This can increase response time and computational cost.

Fourth, the SCES weights and decision thresholds require experimental tuning. Poorly selected thresholds may result in excessive regeneration or unnecessary abstention.

Finally, CEGI-S cannot guarantee that every accepted answer is correct. It is intended as a reliability mechanism rather than a complete replacement for human verification in high-risk applications.

XV. FUTURE SCOPE

Future research can extend CEGI-S in several directions.

First, the SCES weights could be learned automatically from experimental data instead of being manually selected.

Second, the framework could be tested with different LLMs, retrieval models, and knowledge sources to determine whether its performance remains consistent across different environments.

Third, graph-based evidence representation could be explored to represent relationships between claims, entities, and supporting sources.

Fourth, the counterfactual module could be improved to generate domain-specific challenges for areas such as education, scientific research, and enterprise knowledge systems.

Fifth, human-in-the-loop verification could be introduced. In such a system, CEGI-S could automatically answer high-confidence queries while sending uncertain cases to a human reviewer.

Finally, future work could investigate lightweight versions of the framework that provide improved reliability without significantly increasing response latency.

XVI. CONCLUSION

This paper presented CEGI-S (Counterfactual Evidence-Gated Intelligence with Support–Conflict Scoring), a proposed framework for improving the reliability of Large Language Model responses.

The framework is based on the idea that retrieving information is not sufficient by itself to guarantee a reliable answer. A generated response should also be checked against the evidence used to support it. CEGI-S therefore introduces claim-level evidence mapping, source consistency analysis, conflict detection, counterfactual checking, and uncertainty estimation before a final response is delivered.

These signals are combined through the proposed Support–Conflict Evidence Score (SCES). Depending on the score, the system can accept the response, retrieve additional information and regenerate the answer, or abstain when sufficient evidence cannot be established.

The proposed research will compare CEGI-S with conventional LLM, standard RAG, and selected self-corrective or corrective approaches. The evaluation will consider factuality, hallucination rate, evidence support, answer relevance, calibration, abstention quality, latency, and computational cost.

The objective of CEGI-S is not to claim that hallucinations can be completely removed. Instead, it investigates whether an evidence-based decision layer can help an LLM recognize when its response is adequately supported and when it should reconsider or abstain.

The proposed framework can contribute to the broader development of reliable and responsible generative AI by making evidence verification an explicit part of the response-generation process.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Y. Lee, D. Kügelgen, J. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[2] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to retrieve, generate, and critique through self-reflection,” in Proc. Int. Conf. Learn. Representations (ICLR), 2024.

[3] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, “Corrective retrieval augmented generation,” arXiv preprint arXiv:2401.15884, 2024.

[4] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, “ARES: An automated evaluation

framework for retrieval-augmented generation systems," in Proc. NAACL-HLT, 2024.

[5] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: System Demonstrations, 2024, pp. 150–158.

[6] Z. Niu et al., "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics, 2024.

[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and VERification," in Proc. NAACL-HLT, 2018, pp. 809–819.

[8] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 3214–3252.

[9] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.

[10] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2023, pp. 1762–1777.

[11] R. Jia, Z. Fang, M. Wu, J. Lu, and Y. Zhang, "Evidence-aligned entity verification for hallucination detection in retrieval-augmented generation," in Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 29538–29554, doi: 10.18653/v1/2026.findings-acl.1477.

[12] H. Qin, C. Wei, Y. Chen, and Y. Wang, "Counterfactual reasoning for retrieval-augmented generation," in Proc. Int. Conf. Learn. Representations (ICLR), 2026.

[13] J. Qiu, Z. Han, and C. Huang, "SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation," *arXiv preprint arXiv:2605.03534*, 2026.