Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Framework for Hallucination Detection in Large Language Models

Dheeraj Sundaragiri
CSE
SNIST
Hyderabad, India

Lenkala Manohar Reddy
CSE
SNIST
Hyderabad, India

Pitla Gunavardhan
CSE
SNIST
Hyderabad, India

Bandaru Navahith
CSE
SNIST
Hyderabad, India

***Abstract*—Large Language Models (LLMs) have shown strong performance across a variety of natural language processing tasks, including question answering, summarization, and conversational systems. However, these models often generate hallucinated outputs, where statements are incorrect or unsupported by reliable evidence. This issue limits their reliability in applications that require accurate and trustworthy information.**

To address this challenge, this work proposes a multi-signal framework for hallucination detection that combines retrieval-based grounding, natural language inference (NLI), and semantic similarity analysis. The system employs a hybrid retrieval strategy using FAISS-based dense retrieval and BM25-based sparse retrieval to gather supporting evidence. Generated responses are further processed through a fact atomization stage to extract individual claims, which are then verified against retrieved evidence using an NLI model.

The proposed framework was evaluated on a subset of 1,000 samples from the HaluEval benchmark dataset. Experimental results show that the system achieves an accuracy of 92.4%, a precision of 93.1%, a recall of 91.6%, and an F1-score of 92.34. Compared to simpler single-signal approaches, the multi-signal framework demonstrates improved reliability in identifying hallucinated content.

Overall, the proposed approach provides a scalable and interpretable solution for improving the factual grounding of LLM-based systems.

***Index Terms*—Large Language Models, Hallucination Detection, Retrieval-Augmented Generation, Natural Language Inference, Fact Verification, Multi-Signal Learning**

I. Introduction

Large Language Models (LLMs) such as GPT, LLaMA, and PaLM have demonstrated strong capabilities across many natural language processing tasks, including question answering, summarization, and conversational systems [1]–[3]. Despite these advancements, LLMs are known to produce hallucinated outputs, where generated statements are incorrect, unsupported by evidence, or entirely fabricated [4]. This behavior presents significant challenges for applications that require reliable and trustworthy information, such as healthcare systems, educational tools, and decision-support platforms.

As a result, detecting hallucinated responses has become an increasingly important topic in natural language processing research. Existing approaches typically rely on a single verification signal, such as semantic similarity measures, natural language inference (NLI), or retrieval-based verification techniques [8], [10]. While these methods provide useful indicators of factual consistency, relying on a single signal often fails to capture all types of hallucinated content. In particular, subtle factual inconsistencies or unsupported claims may not be detected when verification is performed using only one type of signal.

To address these limitations, this work proposes a multi-signal hallucination detection framework that integrates several complementary verification mechanisms. The proposed approach combines retrieval-augmented evidence gathering, fine-grained claim extraction through fact atomization, NLI-based verification, and additional consistency indicators such as lexical overlap, entity coverage, numerical consistency, and semantic similarity.

The framework further incorporates a machine learning classifier that learns to combine these verification signals into a unified hallucination detection decision. By integrating multiple signals, the system aims to provide a more robust assessment of whether a generated response is supported by available evidence. The framework is evaluated using the HaluEval dataset, which provides benchmark examples for hallucination detection in question answering scenarios [11].

The main contributions of this work can be summarized as follows:

- A hybrid retrieval-based grounding mechanism that collects supporting evidence from both knowledge bases and external sources.
- A fine-grained hallucination detection approach based on fact atomization and natural language inference verification.
- A multi-signal feature representation that integrates semantic, lexical, and numerical consistency indicators.
- A machine learning classifier that combines multiple verification signals to improve hallucination detection performance.
- An empirical evaluation on the HaluEval dataset demonstrating strong performance, achieving an accuracy of 92.4% and an F1-score of 92.34.

Overall, the proposed framework aims to improve the reliability of LLM-generated responses and contribute toward safer deployment of large language models in real-world applications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

II. Related Work

The problem of hallucination in large language models has received increasing attention as these models become widely used in real-world applications. Hallucinations occur when generated text contains information that is incorrect or unsupported by reliable evidence. Several studies have examined this phenomenon and highlighted the need for reliable detection and mitigation techniques [4], [14]. Early approaches to hallucination detection often relied on probabilistic indicators such as token likelihood, model confidence scores, or generation probabilities. Although these signals can provide some insight into model behavior, they frequently fail to reflect the factual correctness of generated statements [15].

Retrieval-based approaches have been proposed to address this limitation by grounding model responses in external knowledge sources. Retrieval-Augmented Generation (RAG) integrates document retrieval mechanisms with generative models so that the model can access relevant information during response generation [5]. Related retrieval-based frameworks such as REALM and Fusion-in-Decoder also aim to improve factual consistency by incorporating external knowledge into the generation process [6], [7]. While retrieval can reduce hallucinations by providing supporting context, errors may still occur when models incorrectly interpret or combine retrieved evidence.

Another widely used strategy for verifying factual consistency involves Natural Language Inference (NLI). NLI models determine whether a hypothesis is entailed, contradicted, or neutral with respect to a given premise [8], [9]. These models have been applied to fact verification tasks where generated claims are compared against supporting evidence [10]. By evaluating logical relationships between claims and evidence, NLI-based approaches provide a structured way to assess whether generated text is supported by reliable information.

More recent research has explored methods that combine multiple verification signals to improve hallucination detection. For example, datasets such as TruthfulQA and HaluEval were introduced to systematically evaluate the factual reliability of language models [11], [12]. Other approaches, such as SelfCheckGPT, attempt to detect hallucinations by analyzing the consistency of model-generated outputs across multiple responses [13]. These studies highlight the importance of using multiple indicators when assessing factual correctness in generated text.

Despite these advances, many existing approaches still rely on a single verification mechanism, which may limit their ability to detect subtle inconsistencies. In contrast, the framework proposed in this work integrates multiple complementary signals, including retrieval-based grounding, natural language inference verification, semantic similarity measures, and machine learning–based classification. By combining these signals, the proposed system aims to provide a more robust and reliable method for detecting hallucinated content in large language model outputs.

III. Proposed Methodology

This study introduces a multi-signal framework for detecting hallucinated responses in large language models. The proposed approach integrates retrieval-based evidence grounding, natural language inference verification, and machine learning–based classification to determine whether a generated response is supported by external knowledge. Instead of relying on a single indicator of factual consistency, the framework combines several complementary signals to improve detection reliability. The overall architecture of the system is illustrated in Fig. 1.

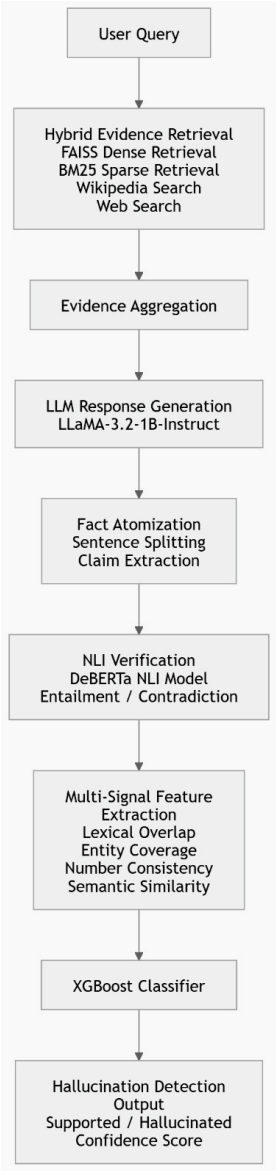

Fig. 1. Overall architecture of the proposed multi-signal hallucination detection system.

The framework is composed of five main components: hybrid evidence retrieval, response generation, fact atomization, NLI-based claim verification, and multi-signal classification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A. Hybrid Evidence Retrieval

The first step of the framework involves retrieving relevant evidence for a given user query. Retrieval-Augmented Generation (RAG) techniques have been shown to improve factual consistency by allowing language models to access external knowledge sources during the generation process [5].

In the proposed system, a hybrid retrieval strategy is employed that combines dense and sparse retrieval approaches. Dense retrieval is implemented using FAISS similarity search over sentence embeddings generated with Sentence-BERT [16]. FAISS allows efficient similarity search across high-dimensional embedding spaces, enabling the system to retrieve semantically related passages from the knowledge base.

Alongside dense retrieval, sparse retrieval is performed using the BM25 ranking algorithm [18]. BM25 measures lexical similarity between the query and candidate documents based on term frequency and inverse document frequency statistics. Combining dense and sparse retrieval improves the robustness of the retrieval stage by capturing both semantic relationships and keyword-level relevance.

In addition to the internal knowledge base, the system can also incorporate information from external sources such as Wikipedia and web search results. The retrieved passages serve as supporting evidence for both response generation and subsequent verification steps.

B. Response Generation

Once relevant evidence has been retrieved, the system generates an answer using a large language model. In this implementation, the *LLaMA-3.2-1B-Instruct* model is used as the base generator. Large language models have demonstrated strong capabilities in language understanding and text generation across a wide range of NLP tasks [2].

The retrieved evidence is incorporated into the prompt provided to the language model, creating a context-aware input. By grounding the generation process in external evidence, the model is encouraged to produce responses that align with the retrieved information rather than relying solely on internal parametric knowledge.

C. Fact Atomization

Responses generated by language models may contain several factual statements within a single sentence. To enable more precise verification, the system applies a fact atomization step that decomposes complex sentences into smaller atomic claims.

During this process, the generated response is first segmented into individual sentences. Each sentence is then analyzed to extract subject–predicate–object structures when possible. These extracted units represent atomic factual claims that can be independently evaluated against the retrieved evidence.

By verifying claims at this finer level of granularity, the framework can identify specific inconsistencies that might otherwise be overlooked if the response were evaluated as a single block of text.

D. NLI-Based Claim Verification

Each atomic claim is verified against the retrieved evidence using Natural Language Inference (NLI). NLI models determine the logical relationship between a premise and a hypothesis, classifying it as entailment, contradiction, or neutral [8].

In this framework, a pre-trained DeBERTa-based cross-encoder is used to estimate entailment and contradiction probabilities between claims and candidate evidence sentences. For each claim, the system first identifies the most relevant evidence segments using semantic similarity. The claim–evidence pairs are then evaluated by the NLI model.

Based on the predicted relationships, claims are categorized as supported, contradicted, or unverifiable. Claims with high contradiction probabilities are treated as potential hallucinations.

E. Multi-Signal Feature Extraction

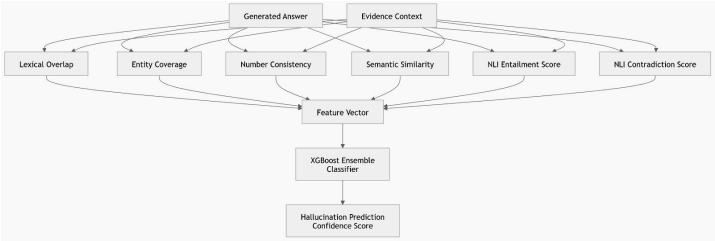

Fig. 2. Multi-signal feature extraction and classification framework used for hallucination detection.

Although NLI verification provides strong evidence for factual consistency, relying on a single signal may not capture every form of hallucination. For this reason, the proposed system extracts several complementary verification signals.

The following features are computed for each generated response:

- **NLI Score:** Measures the level of entailment between the generated claim and supporting evidence.
- **Contradiction Score:** Represents the highest contradiction probability identified by the NLI model.
- **Lexical Overlap:** Calculates the degree of word-level overlap between the generated response and the retrieved evidence.
- **Entity Coverage:** Examines whether named entities mentioned in the generated response are present in the supporting evidence.
- **Number Consistency:** Checks whether numerical values appearing in the generated response match those found in the evidence.
- **Semantic Similarity:** Computes cosine similarity between embeddings of the response and the retrieved evidence.

Together, these features capture both semantic relationships and factual consistency between the generated response and the available evidence.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

F. Machine Learning Classification

The extracted feature set is used to train a machine learning classifier that determines whether a generated response contains hallucinated information. In this work, an XGBoost classifier is employed due to its strong performance on structured feature representations and its ability to model nonlinear interactions between features [19].

The classifier receives the combined feature vector and predicts a binary label indicating whether the response is hallucinated or supported by evidence. In addition to the predicted label, the model also outputs a confidence score reflecting the likelihood of hallucination.

By combining multiple verification signals within a unified classification model, the proposed framework provides a more reliable hallucination detection mechanism than approaches that rely on a single verification signal.

IV. Experimental Setup

This section outlines the experimental configuration used to evaluate the effectiveness of the proposed hallucination detection framework. The experiments aim to measure how well the multi-signal approach identifies hallucinated responses generated by large language models.

A. Dataset

The evaluation was conducted using the *HaluEval* dataset, which was specifically introduced for benchmarking hallucination detection in question answering systems [11]. The dataset consists of question–answer pairs annotated with labels indicating whether the generated response contains hallucinated information.

For the experiments in this study, a subset of 1000 samples from the HaluEval dataset was selected. Each sample includes a user query, a generated answer, and associated evidence that can be used for verification. The dataset also provides ground truth labels indicating whether the answer is factually supported or contains hallucinated content.

B. Knowledge Base

To enable evidence-based verification, a knowledge base constructed from the Simple Wikipedia corpus was used. The corpus was preprocessed and divided into overlapping textual segments to improve retrieval effectiveness.

Each segment was encoded using Sentence-BERT embeddings [16], which capture semantic relationships between sentences. These embeddings were indexed using FAISS to enable efficient similarity search across the document collection [17].

In addition to the local knowledge base, the system can retrieve supplementary information from external sources such as Wikipedia pages and web search results when additional evidence is required.

C. Model Configuration

The system employs the *LLaMA-3.2-1B-Instruct* model as the base language model for response generation. The model was deployed using 4-bit quantization to reduce memory usage while maintaining efficient inference.

For semantic retrieval and similarity computations, the *all-MiniLM-L6-v2* SentenceTransformer model was used to generate sentence embeddings. These embeddings allow the system to measure semantic similarity between generated responses and retrieved evidence.

The NLI verification component uses a pre-trained DeBERTa-based cross-encoder model to evaluate the relationship between claims and supporting evidence. The model produces probabilities representing entailment, contradiction, and neutral relationships between claim–evidence pairs.

D. Evaluation Metrics

The performance of the hallucination detection framework was evaluated using standard classification metrics:

- **Accuracy:** The proportion of correct predictions across all evaluated samples.
- **Precision:** The proportion of predicted hallucinated responses that are correctly identified.
- **Recall:** The proportion of actual hallucinated responses that are successfully detected.
- **F1 Score:** The harmonic mean of precision and recall.

These metrics provide a balanced evaluation of the system's ability to identify hallucinated responses while minimizing both false positives and false negatives.

E. Implementation Details

All experiments were implemented using Python and PyTorch within a Google Colab environment equipped with an NVIDIA Tesla T4 GPU. The system integrates several widely used open-source libraries, including Transformers, SentenceTransformers, FAISS, and XGBoost.

The hallucination detection classifier was trained using the XGBoost algorithm, which is widely used for structured prediction tasks due to its efficiency and strong predictive performance [19]. Feature vectors extracted from the verification pipeline served as input to the classifier during training and evaluation.

V. Results and Discussion

This section presents the evaluation results of the proposed multi-signal hallucination detection framework on the HaluEval dataset. The experiments aim to assess the effectiveness of combining multiple verification signals for identifying hallucinated responses generated by large language models.

A. Quantitative Results

Table I summarizes the performance of the proposed system across standard classification metrics.

The results indicate that the proposed framework performs reliably in detecting hallucinated responses. The precision score shows that most responses identified as hallucinations are correctly classified, while the recall score indicates that the system successfully detects a large proportion of hallucinated outputs present in the dataset. The balanced F1 score

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
PERFORMANCE RESULTS ON HALUEVAL DATASET

Metric	Score
Accuracy	92.4%
Precision	93.1%
Recall	91.6%
F1 Score	92.34

further demonstrates that the model maintains a good trade-off between precision and recall.

B. Comparison with Baseline

To better understand the impact of the proposed multi-signal approach, its performance was compared with a simpler baseline configuration that relied primarily on a single verification signal.

TABLE II
COMPARISON BETWEEN BASELINE AND MULTI-SIGNAL APPROACH

Method	Accuracy	F1 Score
Single-Signal Baseline	70.1%	69.4
Proposed Multi-Signal Framework	92.4%	92.34

As shown in Table II, the multi-signal framework significantly outperforms the baseline approach. The improvement suggests that relying on a single verification signal is often insufficient for detecting complex hallucination patterns. By integrating multiple complementary indicators, the system is better able to identify inconsistencies between generated responses and supporting evidence.

C. Impact of Multi-Signal Verification

The improved performance of the proposed framework can largely be attributed to the integration of several verification signals. Traditional hallucination detection methods often rely on individual indicators such as semantic similarity or NLI-based verification. While these methods provide useful signals, they may fail to capture certain types of factual inconsistencies.

In contrast, the proposed framework combines multiple indicators, including lexical overlap, entity coverage, semantic similarity, and contradiction detection. These signals collectively provide a more comprehensive representation of the relationship between generated responses and retrieved evidence, allowing the system to detect inconsistencies more effectively.

D. Feature Importance Analysis

The machine learning classifier plays a central role in combining the different verification signals. Analysis of feature importance shows that NLI-based signals, particularly entailment and contradiction probabilities, contribute most strongly to the final predictions.

Semantic similarity and entity coverage also play an important role by identifying cases where the generated response introduces entities or concepts that are not present in the supporting evidence. These signals help the classifier detect hallucinations that might not be captured by NLI scores alone.

E. Discussion

Overall, the experimental results indicate that combining retrieval-based grounding with multi-signal verification provides a practical approach for improving hallucination detection. The hybrid retrieval mechanism ensures that relevant evidence is available for verification, while the NLI component enables fine-grained reasoning over claim–evidence relationships.

The use of a machine learning classifier further improves the robustness of the system by learning how different verification signals interact. This allows the framework to adapt to different types of hallucination scenarios and improves overall detection performance.

VI. FUTURE WORK

Although the proposed multi-signal framework demonstrates strong performance for hallucination detection, several directions remain for further improvement. One potential extension involves evaluating the framework on larger and more diverse datasets to better understand its generalization capabilities across different domains and task settings.

Another promising direction is the integration of stronger language models and retrieval mechanisms. Recent developments in retrieval-augmented generation and large-scale embedding models may provide more accurate evidence retrieval, which could further improve the reliability of the verification process.

Future work may also explore more advanced reasoning methods for claim verification. For example, multi-hop reasoning and structured knowledge sources such as knowledge graphs could enable deeper verification of complex factual claims that require multiple pieces of supporting evidence.

In addition, the current framework focuses primarily on question answering scenarios. Extending the approach to other generative tasks, such as summarization, dialogue systems, and content generation, could provide broader insights into hallucination detection across different applications.

Finally, incorporating lightweight real-time detection mechanisms may allow the framework to be deployed in practical systems where hallucination detection must occur during generation rather than after response production.

VII. CONCLUSION

Hallucination remains one of the major challenges in the practical deployment of large language models. When generated responses contain unsupported or fabricated information, the reliability and trustworthiness of AI systems can be significantly affected, particularly in applications that require accurate knowledge and factual consistency.

This work presented a multi-signal framework for hallucination detection that integrates hybrid evidence retrieval, fact atomization, natural language inference verification, and machine learning–based classification. The proposed approach

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

combines several complementary verification signals, including lexical overlap, entity coverage, numerical consistency, semantic similarity, and contradiction detection, to evaluate whether generated responses are supported by external evidence.

Experimental evaluation on the HaluEval dataset shows that the proposed framework achieves strong detection performance, obtaining an accuracy of 92.4% and an F1-score of 92.34. The results indicate that combining multiple verification signals provides a more reliable mechanism for detecting hallucinated content compared to approaches that rely on a single signal.

Overall, the proposed framework contributes toward improving the reliability of LLM-generated outputs and provides a practical foundation for building more trustworthy language model applications.

REFERENCES

[1] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020.
[2] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," 2023.
[3] A. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways," 2022.
[4] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, 2023.
[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[6] K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-training," ICML, 2020.
[7] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models," ACL, 2021.
[8] S. Bowman et al., "A Large Annotated Corpus for Learning Natural Language Inference," EMNLP, 2015.
[9] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Natural Language Inference," NAACL, 2018.
[10] J. Thorne et al., "FEVER: A Large-scale Dataset for Fact Extraction and Verification," NAACL, 2018.
[11] P. Minervini et al., "HaluEval: A Large-Scale Hallucination Evaluation Benchmark," 2023.
[12] S. Lin et al., "TruthfulQA: Measuring How Models Mimic Human Falsehoods," ACL, 2021.
[13] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection," EMNLP, 2023.
[14] J. Maynez et al., "On Faithfulness and Factuality in Abstractive Summarization," ACL, 2020.
[15] Z. Jiang et al., "How Can We Know What Language Models Know?" TACL, 2021.
[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks," EMNLP, 2019.
[17] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with FAISS," IEEE Transactions on Big Data, 2019.
[18] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in IR, 2009.
[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD, 2016.
[20] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
[21] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.
[22] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[23] A. Radford et al., "Language Models are Unsupervised Multitask Learners," OpenAI, 2019.
[24] F. Petroni et al., "Language Models as Knowledge Bases?" EMNLP, 2019.
[25] P. Laban et al., "Summarization Benchmarks for Hallucination Detection," 2023.
[26] N. Dziri et al., "Faithfulness in Natural Language Generation," TACL, 2022.
[27] O. Honovich et al., "QA-based Fact Verification," ACL, 2022.
[28] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," 2023.