STM JOURNALS

Emerging Trends in
Languages

ISSN: 3049-0693
Volume 3, Issue 2, 2026
July–December
DOI (Journal): 10.37591/ETL

https://journals.stmjournals.com/etl

Review ETL

Semantic Similarity Framework for Automatic Hallucination Detection in Large Language Models

Sujal Gupta[1]*, Hardik Sharma[2], Shivansh Narayan[3], Dhruv Kumar[4], Aayush Bisht[5]

Abstract

Large Language Models can generate fluent, contextually appropriate text across a range of NLP tasks, but they frequently produce outputs that are factually wrong while sounding confident and plausible. This problem, referred to as hallucination, poses serious risks in domains where accuracy matters. We propose a post-processing framework that detects hallucinated responses by comparing them against verified reference text using sentence embeddings. The system computes cosine similarity between the response and reference, and extracts four additional features: named entity mismatches, numerical inconsistencies, keyword overlap, and output perplexity. These features are used to train three classifiers: Logistic Regression, SVM, and Random Forest. Random Forest achieved the best results with 87% accuracy and 0.84 F1-score. The framework requires no access to the language model's internals and works as a standalone verification layer, which makes it applicable to any LLM including API-based services. To evaluate the robustness of the proposed approach, experiments were conducted on a benchmark dataset containing both factual and hallucinated responses across diverse domains. The results demonstrate that combining semantic similarity with linguistic and statistical features significantly improves hallucination detection compared to relying on embedding similarity alone. The proposed framework is computationally efficient, easily deployable, and scalable, making it suitable for integration into real-world applications where reliable and trustworthy AI-generated content is essential.

Keywords: Large language models, hallucination detection, semantic consistency, machine learning, generative AI, AI reliability, natural language processing.

INTRODUCTION

LLMs trained on large text corpora through next-token prediction have become widely used for question answering, translation, summarization, and dialogue. Their adoption in commercial products has grown rapidly, partly because model APIs have made them accessible without requiring in-house ML infrastructure. The outputs these models produce are often impressive in terms of fluency and coherence.

The problem is that LLMs generate text by predicting statistically likely token sequences, not by verifying whether what they say is true. They have no internal fact-checking mechanism. When the statistical patterns in the training data lead to a plausible-sounding but incorrect sequence, the model produces it with the same confidence as a correct answer. Researchers call this hallucination, and it is a well-documented failure mode.

Hallucinations cause serious problems in practice. In healthcare, an incorrect claim about drug interactions could affect patient safety. In legal research, fabricated case citations have already been documented in real court filings. In academic settings, students and researchers who rely on LLM-generated references may

***Author for Correspondence**
Sujal Gupta
E-mail: sujalgupta85777@gmail.com

[1-5]Student, Department of CSE, Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India

Received Date: June 29, 2026
Accepted Date: July 03, 2026
Published Date: July 04, 2026

Citation: Sujal Gupta, Hardik Sharma, Shivansh Narayan, Dhruv Kumar, Aayush Bisht. Semantic Similarity Framework for Automatic Hallucination Detection in Large Language Models. Emerging Trends in Languages. 2026; 3(2): 16–22p.

© STM Journals 2026. All Rights Reserved

unknowingly cite papers that do not exist. The common thread across these examples is that hallucinated text looks identical to correct text on the surface, making it difficult for users to detect without independent verification [1].

Several approaches have been proposed to address this. Retrieval-Augmented Generation (RAG) grounds model outputs in retrieved documents, which helps but does not eliminate hallucination entirely since models can still ignore or misrepresent retrieved content. Uncertainty estimation methods look at token-level probabilities to flag low-confidence outputs, but hallucinated content often has high model confidence. Human review provides high-quality verification but does not scale to production throughput.

We present a different approach: a lightweight detection framework that operates entirely as a post-processing step. Given an LLM response and a verified reference text, the system computes semantic similarity using sentence embeddings and extracts four supplementary features that capture different hallucination patterns. A supervised classifier then predicts whether the response is factually reliable or hallucinated. The system does not require access to the model's internal states or probability distributions, so it can be deployed alongside any LLM, including closed-source API services [2].

LITERATURE REVIEW

Research on hallucination in LLMs has expanded considerably in recent years. Ji et al. published a survey covering hallucination types and detection strategies, which was useful for understanding how the field categorizes these failures. They distinguish between intrinsic hallucinations (contradicting the source material) and extrinsic hallucinations (fabrications that cannot be verified against any reference) [3].

Knowledge-grounded verification is one of the older approaches. The idea is to check model outputs against trusted knowledge bases or databases. It works when the query falls within the coverage of the knowledge source but struggles with topics that are not well represented in the reference corpus. The retrieval step also adds latency, which matters for real-time applications.

Uncertainty estimation takes a different angle by looking at the model's internal confidence signals. The assumption is that low-confidence predictions are more likely to be hallucinated. In practice, this is unreliable because LLMs frequently hallucinate with high confidence. Farquhar et al. proposed semantic entropy as an alternative, measuring uncertainty over meanings rather than individual tokens, which showed improvement over standard token-probability methods.

RAG, introduced by Lewis et al., is probably the most widely adopted intervention. It retrieves relevant documents before generation, giving the model factual material to reference. RAG reduces hallucination but does not solve it. Models sometimes generate content that contradicts the retrieved context. Sun et al. studied this specific failure mode and proposed the ReDeEP framework for detecting cases where models fail to properly use retrieved information.

Maynez et al. studied hallucination in abstractive summarization, a setting where the source document is explicitly provided. Even with the factual content available, models frequently generated unsupported claims. This is a useful finding because it shows hallucination is not just a result of missing information. It appears to be a more fundamental tendency in how these models generate text.

Manakul et al. proposed SelfCheckGPT, which samples multiple responses to the same prompt and measures consistency across them. The logic is that factual content should be consistent across samples while hallucinated content should vary. Yang et al. used metamorphic relations, essentially systematic query transformations, to test whether model outputs remain logically consistent under rephrasing [4].

© STM Journals 2026. All Rights Reserved

Our work builds on semantic similarity methods, which compare generated text against reference material in an embedding space. Sentence-BERT enables this by producing dense sentence representations where semantically similar texts are mapped to nearby vectors. We extend this approach by combining the similarity score with additional features targeting specific hallucination patterns and training a classifier on the combined feature set. In our review of the existing work, we found that while individual hallucination signals have been studied, fewer papers have combined multiple complementary features within a single classification framework designed as a generic post-processing layer [5].

PROBLEM STATEMENT

Given an LLM-generated response G and a verified reference text R for the same query, the task is to classify G as either factually consistent with R (correct) or containing hallucinated content (hallucinated). This is a binary classification problem.

The task is difficult for several reasons. Hallucinated text is not distinguishable from correct text based on surface features like fluency, grammar, or length. A response that states an incorrect year or attributes a discovery to the wrong person reads the same as one that gets these details right. Hallucination also varies in granularity: some responses are entirely fabricated, while others contain a single wrong detail in otherwise accurate text. Additionally, LLMs assign high probability to hallucinated sequences, so the model's own confidence is not a reliable indicator of correctness.

The consequences of undetected hallucination are significant in high-stakes domains. Incorrect medical guidance fabricated legal citations, and nonexistent academic references all represent concrete harms that have been documented in real-world use. Automated detection is needed because manual review cannot keep pace with the volume of text that deployed LLMs generate.

Our objective is a detection system that operates on the output text only, without requiring access to model internals or retrieval infrastructure. The system should work as a modular post-processing component that can be added to any LLM deployment [6].

PROPOSED METHODOLOGY

Our approach has five main steps: collecting a labeled dataset, preprocessing the text, computing semantic similarity, extracting additional features, and training classifiers.

Data Collection

We assembled a dataset of prompt-response pairs spanning multiple domains. Each entry contains the input query, the LLM-generated response, and a verified reference answer sourced from peer-reviewed publications, encyclopedias, and curated databases. Every response was manually labeled as correct or hallucinated by comparing its factual claims against the reference. This labeling step required reading both texts carefully and checking each claim individually, which was time-consuming but necessary for reliable ground truth [7].

Text Preprocessing

Before computing similarity, we ran all text through a standard preprocessing pipeline: lowercasing, removing punctuation and special characters, tokenizing into words, removing stop words, and applying lemmatization to reduce words to their base forms. These are standard NLP preprocessing steps. Their purpose is to remove surface-level variation (capitalization, word inflections, formatting differences) so that the downstream similarity computation focuses on content rather than presentation [8].

Semantic Similarity Analysis

The core of our approach is the comparison between the generated response and the reference text

© STM Journals 2026. All Rights Reserved

in a semantic embedding space. We encode both texts into fixed-dimensional vectors using a pretrained sentence embedding model. Texts with similar meaning end up close together in this space regardless of whether they use the same words.

We measure the distance between the two vectors using cosine similarity, which produces a score between -1 and 1. Higher values indicate greater semantic alignment. A factually correct response should score high because it conveys the same information as the reference. A hallucinated response should score lower because the claims it makes differ from the reference, even if the text is fluent and well-structured [9].

Feature Extraction

Semantic similarity captures the overall alignment between response and reference, but it can miss localized errors. A response that gets most things right but invents one author name or shifts a date by two years might still have a reasonably high similarity score. To catch these cases, we extract four additional features.

The first is the semantic similarity score itself. The second is a named entity mismatch count: we extract entities (people, organizations, locations) from both texts using named entity recognition and count how many entities in the response do not appear in the reference. LLMs commonly fabricate entities, especially researcher names and institutional affiliations, so this feature targets a frequent hallucination pattern.

The third feature is a numerical inconsistency count. We identify numbers in both texts (dates, percentages, quantities) and check for discrepancies. Hallucinated responses often contain numbers that are close to correct but slightly off, which is hard to catch through semantic similarity alone. The fourth is a keyword overlap ratio measuring the proportion of content words in the response that also appears in the reference. The fifth is a perplexity score over the generated text, which indicates how surprising the output is under a language model distribution. Higher perplexity can sometimes indicate generation under low model certainty [10].

Classification

The five features are combined into a single feature vector for each response. We trained three classifiers: Logistic Regression as a linear baseline, SVM with an RBF kernel for nonlinear boundaries, and Random Forest as an ensemble method. We chose these three because they cover a range of model complexity without introducing the overhead of a deep learning classifier, which would be unnecessary given that the feature space is only five-dimensional.

All classifiers were trained on a held-out training partition and evaluated on a separate test set. Hyperparameters were selected through cross-validation on the training data. Each classifier outputs a binary label: correct or hallucinated [11].

Evaluation Metrics

We report accuracy, precision, recall, and F1-score. Accuracy is the overall fraction of correct predictions. Precision measures how many of the responses flagged as hallucinated actually were hallucinated. Recall measures how many of the actual hallucinations the system caught. F1-score balances precision and recall. Which metric matters most depends on the application: safety-critical systems need high recall to avoid missing hallucinations, while user-facing tools may prioritize precision to avoid false alarms [12].

RESULTS AND ANALYSIS

Random Forest performed best across all metrics: 87% accuracy, 0.85 precision, 0.84 recall, and 0.84 F1-score. This is likely because the relationship between the features and hallucination involves nonlinear interactions that tree-based ensembles handle well. SVM achieved 82% accuracy and 0.79 F1, which is respectable but below the ensemble method. Logistic Regression was the weakest at 78%

© STM Journals 2026. All Rights Reserved

accuracy and 0.75 F1, which is expected given its limitation to linear decision boundaries. The fact that even the linear model performs well above chance confirms that the features carry real discriminative signal (Table 1).

Table 1. Performance comparison of classifiers for hallucination detection.

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic regression	0.78	0.76	0.74	0.75
SVM	0.82	0.80	0.79	0.79
Random forest	0.87	0.85	0.84	0.84

Across all three classifiers, precision was slightly higher than recall. This means the system is somewhat conservative: when it flags something as hallucinated it is usually correct, but it lets some actual hallucinations pass through. Whether this tradeoff is acceptable depends on the deployment context.

Feature Importance

In the Random Forest model, semantic similarity score was the most important feature by a clear margin. This confirms the basic premise of the framework: hallucinated responses diverge semantically from verified references in ways that are measurable through embedding-based comparison, even when the text reads perfectly well on the surface.

Named entity mismatch was the second most important feature. During labeling we noticed a recurring pattern where responses would reference researchers, journals, or institutions that sounded plausible but did not actually exist. The entity mismatch feature picks up on this directly. Numerical inconsistency ranked third. We saw many cases where hallucinated responses had numbers that were close to correct but slightly off, like a year being wrong by one or two, or a percentage being inflated. These near-miss errors are easy for humans to overlook and hard to catch through similarity alone.

Keyword overlap and perplexity had the lowest individual importance, though both contributed positively to the overall classification. They capture different aspects of the text that the other features miss, which justifies including them even if their standalone performance is modest.

Types of Hallucinations Observed

Labeling the dataset gave us a good understanding of how hallucination actually looks in practice. The most common type was entity fabrication. Responses would contain citations to papers with realistic-sounding titles by real authors in real journals, except the specific paper did not exist. Sometimes two out of three citations were real and the third was fabricated, which made the fake one harder to spot.

The second type was what we call marginal factual distortion. The response is mostly correct but one or two specific facts are wrong: a date is off by a year, a statistic is inflated, or a discovery is attributed to the wrong person. These cases are difficult for our framework because the overall semantic similarity with the reference remains high. The error is localized to a small portion of the text.

The third type involved responses that blended accurate information from different but related topics. For example, when asked about one scientific concept, the response might incorporate details that actually belong to a related but distinct concept. The individual facts might each be correct in their original context but combining them produces a response that is misleading. We found this to be the hardest type for our system to detect reliably [13].

LIMITATIONS

Our work has several limitations that should be noted.

© STM Journals 2026. All Rights Reserved

The main one is the need for verified reference text. The framework requires a factually accurate reference for each query, which is feasible in structured information retrieval settings but not for open-ended questions or novel topics where no authoritative reference exists. Extending the approach to work without references is an open problem.

We evaluated on a single manually labeled dataset of moderate size. The results are promising but we cannot be sure how well the classifier would perform on a larger or differently distributed dataset. The small scale also limits the precision of our feature importance estimates.

We only tested on outputs from one LLM. Different models may hallucinate in different ways depending on their training data and architecture. A system tuned to one model's hallucination patterns may not transfer well to another without retraining.

The feature set, while effective for the patterns we observed, does not cover all types of hallucination. Logical inconsistencies and faulty reasoning chains are not well captured by our five features. A richer feature representation, possibly including structural or pragmatic linguistic features, could improve detection of these harder cases.

CONCLUSION

We presented a post-processing framework for detecting hallucinations in LLM outputs. The system compares generated responses against verified reference text using sentence embeddings, extracts four additional features targeting specific hallucination patterns, and trains a classifier on the combined feature set. Random Forest achieved the best performance at 87% accuracy and 0.84 F1-score.

The main practical advantage of the approach is that it requires no access to model internals. It works on the output text alone, so it can be deployed alongside any LLM including closed-source API services. In our view, this makes it a useful tool for organizations that want to add a verification layer to their existing LLM deployments without modifying the underlying system.

For future work, we see several directions worth exploring. Replacing the traditional classifiers with a fine-tuned transformer might improve detection of the subtler hallucination types, particularly the marginal distortions and context conflation patterns that our current system struggles with. Adding a retrieval component that automatically finds reference material at inference time would reduce the reference dependency, which is currently the biggest practical limitation. Extending the framework to non-English languages would require adapting both the embedding model and the entity recognition pipeline, but would substantially expand the system's applicability. Finally, incorporating features that capture logical coherence between sentences, rather than just semantic similarity and entity matching, could help with the reasoning-based hallucination patterns that our current features do not address well.

As LLMs continue to be deployed in settings where accuracy matters, the need for automated reliability checks will only grow. We hope this framework provides a useful starting point for that work.

REFERENCES

1. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM computing surveys. 2023 Mar 3;55(12):1-38.
2. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics 2020 Jul (pp. 1906-1919).
3. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S. Retrieval-augmented generation for knowledge-intensive nlp tasks.

© STM Journals 2026. All Rights Reserved

Advances in neural information processing systems. 2020; 33:9459-74.
4. Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, Xu J, Ott M, Smith EM, Boureau YL, Weston J. Recipes for building an open-domain chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume 2021 Apr (pp. 300-325).
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30 5998–6008
6. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).
7. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020; 33:1877-901.
8. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM computing surveys. 2023 Mar 3;55(12):1-38.
9. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024 Jun 20;630(8017):625-30.
10. Manakul P, Liusie A, Gales M. 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 2023 Dec (pp. 9004-9017).
11. Yang B, Al Mamun MA, Zhang JM, Uddin G. Hallucination detection in large language models with metamorphic relations. Proceedings of the ACM on Software Engineering. 2025 Jun 19;2(FSE):425-45.
12. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. InProceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) 2019 Nov (pp. 3982-3992).
13. Sun Z, Zang X, Zheng K, Xu J, Zhang X, Yu W, Song Y, Li H. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In International Conference on Learning Representations 2025 May 1 (Vol. 2025, pp. 50250-50279).

© STM Journals 2026. All Rights Reserved