



AI-Based Hallucination Detection in Large Language Models Learning: A Review

Navale Sanika Shashikant¹, Bangar Abhishek Popat²,
Bhosale Sachin Bajirao³, Khatri Anand Ashok⁴

Department of Computer Engineering / Jaihind College of Engineering / SPPU, Pune, India

Abstract. Large language models (LLMs) can produce fluent but incorrect or fabricated statements, called hallucinations, which limits their use in education, healthcare and other high-stakes areas. Most existing safeguards act only after generation, and few correct errors automatically. This paper presents a structured narrative review of [N] verified sources published between [2022] and 2026, organized into seven themes: taxonomies, sampling-consistency detection, semantic-clustering methods, retrieval and fact-verification, internal-state detection, correction and integrated architectures, and evaluation practices. Black-box methods can detect hallucinations without external knowledge or model access, but only 8.8 % of reviewed systems combine detection and mitigation. Evidence relies heavily on question-answering benchmarks, small datasets, automatically generated labels and non-comparable metrics, and correction layers are rarely validated. The paper contributes a thematic synthesis, a comparison of [N] studies, a conceptual seven-layer self-correcting architecture and a prioritized research agenda.

Keywords: Large Language Models, Hallucination Detection, Semantic Consistency, Retrieval-Augmented Generation, Self-Correction, Trustworthy AI

I. Introduction

Large language models (LLMs) are widely used for question answering, summarization, dialogue and code generation. Their usefulness depends on how reliable their answers are. LLMs generate text by predicting the next word from patterns learned during training, so they can produce content that reads fluently but is factually wrong, unverifiable or inconsistent with a source document. This is known as hallucination [1]. Such errors can come from flawed training data, outdated or incomplete knowledge, and the way text is decoded during generation [1]. If hallucinations are not identified before users rely on them, they can spread misinformation and cause harm, especially in fields such as law and medicine [1], [5].

Traditional safeguards address parts of the problem. Methods that check an answer against external knowledge bases or stronger models need reliable resources and can be costly, and methods that inspect a model's internal states need full white-box access [2]. Sampling-consistency methods ask the same question several times and treat disagreement between answers as a warning sign, without needing either resource [2]. Semantic-clustering methods group answers by meaning and measure how scattered they are, using inconsistency as a signal of uncertainty [3]. However, most protections act only after the text is



generated, and only 11 of 125 reviewed studies (8.8 %) combine detection and mitigation in a single architecture [1]. Related work outside text, such as multimodal change detection, shows that iterative correction and multi-path self-consistency can also reduce hallucination-type errors [4].

This review discusses selected research papers based on their objectives, data sources, techniques, results and limitations. It covers hallucination taxonomies, sampling-consistency detection, semantic-clustering methods, correction and integrated architectures, and evaluation practices. It also explains the difference between detecting whether a response contains a hallucination, identifying its type correcting the error, and re-verifying the corrected answer.

II. Review Methodology

A focused search of IEEE Xplore, was carried out in [using the terms hallucination, large language model, hallucination detection, semantic entropy, consistency checking, retrieval-augmented generation and self-correction, prioritising [2022]–2026 publications. Studies were included when they proposed, evaluated or reviewed a hallucination detection, mitigation or evaluation method for large language models and reported a stated method with a quantitative outcome or a clear analytical contribution. Studies that mentioned hallucination only briefly, were not written in English or were not peer-reviewed were excluded, except for [x] labelled preprints. Existing surveys were used for context only. Every reference was verified against a publisher, DOI or repository record. The review is narrative rather than systematic, and study counts are not reported.

III. Literature Review

1. Systematic Review of Hallucinations in Large Language Models

Woesle, Fischer-Brandies, and Buettner [1] presented a systematic literature review of hallucinations in large language models (LLMs) following the PRISMA guidelines. They searched IEEE Xplore, ACM Digital Library, and ScienceDirect for studies published between 2015 and January 2025. Of the 132 publications identified, 125 were analyzed across nine application domains, and seven literature reviews were excluded. The authors classified the studies into four system architectures: unaugmented generation, post-hoc reactive validation, proactive detection-and-mitigation, and integrated detection-and-mitigation. They also defined seven outcome scenarios, ranging from correct generation and unchecked hallucination to unrecognized hallucination and errors introduced by the correction layer itself. According to the authors, question answering and multimodal tasks account for 48% of the studies, and only 11 studies (8.8%) use an integrated architecture.

Critical Analysis: This study gives a useful map of the field and identifies a clear gap in integrated detection-and-correction systems. However, the authors acknowledge that assigning papers to architecture categories involves some subjectivity, and that simplified categories can hide the details of complex methods. The review does not compare the quantitative results of the studies, so it cannot show which method performs best. Its claim that integrated architectures can be transferred to underrepresented domains is conceptual and has not been tested experimentally. The search was also limited to three databases and keyword filters, which may have excluded relevant work.

2. Multi-Perspective Consistency Checking for Black-Box Detection

Kong and co-authors [2] proposed a black-box, zero-resource method for detecting LLM hallucinations that needs neither external knowledge nor access to model internals. The method asks the same query several times to collect multiple responses (forward sampling) and also asks the model to regenerate the query from the response (backward sampling). Sentence-BERT embeddings and cosine similarity are used to score consistency, and the two scores are combined as a weighted sum with a

fusion coefficient of 0.8, using five samples in each direction. The authors built two datasets, a Books question-answering set of 1,000 items and a Square Num arithmetic set of 500 items, and tested Llama2-7B, Qwen2-7B, and Mistral-7B. On the Books dataset with Llama2-7B, the method reached 77.8% balanced accuracy and an AUC of 0.832, which is 2.3 percentage points above the single-perspective SelfCheckGPT baseline. Critical Analysis: The study shows that checking consistency from two directions can reduce the overconfidence of single-perspective methods without any external resources. However, the datasets are small, self-built, and limited to closed-domain tasks, and the hallucination rates are very uneven (for example 94.7% and 11.4%). The gains are not uniform: on the Square Num dataset, other baselines scored higher in some settings, and a variable temperature performed worse than a constant one. The method also requires multiple generations for every query, which adds computing cost, and it only detects hallucinations without correcting them. Only two perspectives, one similarity measure, and a simple linear fusion were examined, and only 7B models were tested.

3. SINdex: Semantic Inconsistency Index

Abdaljalil and co-authors [3] introduced a black-box framework that clusters sampled responses by meaning and scores them with a new measure called SINdex. For each question, 10 responses are generated, encoded together with the question using a sentence-embedding model (all-MiniLM-L6-v2), and grouped by hierarchical agglomerative clustering with a distance threshold of 0.05. SINdex is an entropy over the clusters in which each cluster's proportion is weighted by its average internal cosine similarity, so that lower values indicate more consistent answers. The method was tested on TriviaQA, Natural Questions, SQuAD, and BioASQ (400 questions each) using Llama-2, Mistral, and Falcon models. The authors report AUROC improvements of up to 7.6%, 9.3%, 9.1%, and 4.8% over semantic entropy on the four datasets, and a 60-fold speed-up at 200 generations.

Critical Analysis: The study addresses the cost of entailment-based semantic entropy by working directly in embedding space, and it supports its results with bootstrap confidence intervals and ablation studies. However, the ground-truth labels are created automatically by comparing embeddings with a reference answer, which the authors acknowledge may overestimate detection performance. The clustering threshold is fixed at 0.95 and may not suit other settings, and only 400 questions per dataset were used. All experiments are question-answering tasks, so the method may not generalise to code, mathematics, or open-ended generation where varied answers can be valid. Like the previous study, it detects hallucinations but does not correct them.

4. Hallucination-Resistant Change Detection with Multimodal LLMs

Zuo and co-authors [4] proposed CVAHR-CD-LLM, a multimodal framework for change detection in remote-sensing images. The system extracts temporal differences from paired images with a SigLip encoder, passes them with text prompts to an LLM, and refines the predicted change locations through an iterative calibration loop based on change vector analysis, which typically converges in three to five iterations. A self-consistency mechanism aggregates several chain-of-thought reasoning paths to reduce hallucinated detections. The model was tested on the CDD, GVLM, and WHU-CD datasets against FC-SC, SNU-Net, DT-SCN, and AMTNet. In the ablation study on GVLM, the full model reached 99.39% overall accuracy, 95.36% F1-score, and 90.02% IoU, compared with 97.16%, 89.11%, and 84.56% for the baseline without the two components.

Critical Analysis: This study is useful because it shows an integrated design in which detection, correction, and repeated verification form a loop, which is the type of architecture that Woesle et al. [1] found to be rare. However, "hallucination" here means false or irrelevant detected changes in images, not factually wrong text, so the results cannot be compared directly with text-based detectors. The comparison is made only with vision-only change-detection models rather than with other hallucination-mitigation methods, and the ablation is reported on a single dataset. The authors note that the calibration loop adds

latency, that semi-supervised data augmentation may spread domain bias, and that hallucinations can remain under severe radar-optical misalignment.

5. Survey of Fake Text Detection

Kwon and Jang [5] surveyed techniques for detecting fake text, which they divide into misinformation and LM-generated text. For misinformation, they review traditional feature-based and machine-learning methods, deep learning (neural networks, attention, graph neural networks, and multimodal models), information-based approaches that use content and context, and human-aided approaches. For LM-generated text, they review statistical methods, generative approaches such as DetectGPT and Grover, adversarial training, watermarking, and human-aided detection. Hallucination is discussed as one of five challenges of LM-generated text, together with copyright infringement, spam and phishing, bias, and academic disruption. The authors identify limited and biased datasets, the constant evolution of generated text, and false positives as the main limitations. Critical Analysis: This survey is a helpful background source because it places hallucination within the wider problem of unreliable text and highlights the importance of datasets and false-positive control. However, its scope is broad, and hallucination detection receives only limited coverage. Detecting whether a text is machine-generated or misleading is also a different task from checking whether the content of a particular LLM response is factually correct. The survey does not provide a quantitative comparison of methods, so its findings should be used as context rather than as a direct comparison with the hallucination-specific methods reviewed above.

IV. Comparative Analysis

Table 1. Comparative Analysis of Existing Research on Hallucination Detection and Mitigation in Large Language Models

Study	Input / approach	Main output	Key limitation
Woesle et al. [1]	PRISMA review of 125 studies, four system architectures, seven outcome scenarios	Taxonomy and gap analysis	Subjective categories; no quantitative comparison
Kong et al. [2]	Forward and backward sampling, SBERT consistency scores, linear fusion	Fact or hallucination label	Small self-built datasets; detection only
Abdaljalil et al. [3]	Sampled responses, embedding-based clustering, SIndex	Hallucination score	Fixed threshold; embedding-based labels; question answering only
Zuo et al. [4]	SigLip encoder, LLM, calibration loop, self-consistency	Refined change map with fewer false detections	Image-based task; added latency
Kwon and Jang [5]	Survey of misinformation and LM-generated text detection	Taxonomy and challenges	Hallucination covered only briefly

V. Research Gaps

- The analysis of the selected research papers on AI-Based Hallucination Detection in Large Language Models (LLMs) reveals several areas that require further investigation and improvement.
- Some studies examine different hallucination-related problems together. Detecting factual errors, identifying unsupported claims, measuring uncertainty, and correcting hallucinated responses should be evaluated as separate tasks.
- High detection accuracy alone does not guarantee that a system can reliably identify rare or subtle hallucinations. Additional evaluation measures, such as precision, recall, F1-score, AUROC, and AUPRC, should also be considered.



- Evaluating a detection method on similar or overlapping data may lead to misleading results. Using independent test sets and carefully designed data splits can provide a more reliable assessment of whether a method generalizes to unseen questions and responses.
- A detection approach that performs well on one dataset or language model may not achieve similar results with other models, languages, or application domains. Differences in model architecture, response style, question complexity, and domain knowledge can influence detection performance.
- Methods based on repeated response generation and semantic consistency may be affected by computational cost, response variability, and the quality of semantic representations. These factors should be considered when developing and evaluating hallucination detection systems.
- Techniques such as semantic clustering and uncertainty estimation can identify inconsistencies among generated responses. However, consistent responses are not necessarily factually correct, so additional verification against reliable evidence may be required.
- The practical usefulness of a hallucination detection system depends on more than its detection accuracy. False alarms, undetected hallucinations, processing time, computational resources, and the ability to provide understandable explanations should also be evaluated.

VI. Proposed Improvements

The following improvements are suggested for future development and testing.

- Define the detection objective: Select a clear objective, such as identifying factual errors in LLM-generated responses, detecting unsupported claims, measuring uncertainty, or classifying responses as hallucinated or non-hallucinated.
- Understand the dataset: Record details such as the data source, number of questions and responses, input features, types of hallucinations, class distribution, missing values, and the method used to label responses as correct or hallucinated.
- Test different detection techniques: Start with a basic detection method and compare its performance with suitable approaches, such as semantic consistency checking, uncertainty estimation, semantic clustering, and fact verification using external knowledge sources.
- Use a proper testing method: Divide the dataset into training, validation, and testing sets wherever applicable. Ensure that similar or duplicate questions and responses do not cause data leakage. Evaluate the system on independent data to measure its ability to detect hallucinations in unseen responses.
- Evaluate system performance: Use a confusion matrix, precision, recall, F1-score, and accuracy to understand the detection system's performance. PR-AUC can also be considered when the dataset contains an unequal number of hallucinated and non-hallucinated responses.
- Explain the detection results: Provide understandable explanations for why a response is identified as potentially hallucinated. Highlight unsupported claims, inconsistencies, or differences between the generated response and reliable evidence. The system should clearly distinguish between a potential hallucination and a confirmed factual error.
- Test the system under practical conditions: Evaluate the detection system using different language models, question types, and application domains. Measure processing time, computational requirements, false alarms, and missed hallucinations to determine whether the system is suitable for real-time use.

VII. Real Time Workflow

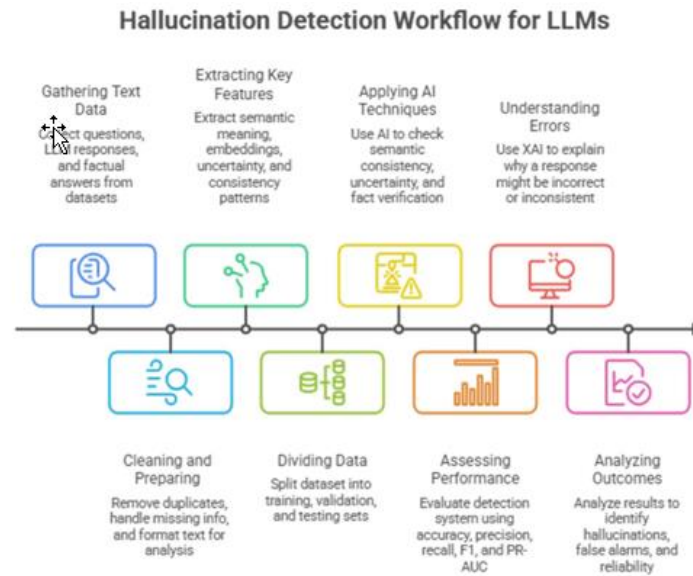


Figure 1: Proposed Workflow for AI-Based Hallucination Detection in Large Language Models (LLMs)

The proposed system aims to detect hallucinations in Large Language Models (LLMs) by analysing generated responses using AI-based techniques. The workflow consists of the following steps:

- **Data Collection:** Questions, LLM-generated responses, and factual reference answers are collected from available datasets or reliable knowledge sources.
- **Data Preprocessing:** The collected text data is cleaned by removing duplicate responses, handling missing information, and preparing the data for analysis.
- **Feature Extraction:** Important features, such as semantic meaning, text embeddings, uncertainty, and consistency patterns, are extracted from the generated responses.
- **Dataset Splitting:** The prepared dataset is divided into training, validation, and testing sets to support model development and reliable performance evaluation.
- **Hallucination Detection:** AI-based techniques, such as semantic consistency checking, uncertainty estimation, and fact verification, are applied to identify potentially hallucinated responses.
- **Model Evaluation:** The detection system is evaluated using performance measures such as accuracy, precision, recall, F1-score, and PR-AUC to assess its effectiveness.
- **Explainable AI:** Explainable AI techniques are used to understand why a response may be incorrect, unsupported, or inconsistent with reliable evidence.
- **Result Assessment:** The final results are analysed to identify detected hallucinations, false alarms, missed errors, processing time, and overall system reliability.

The proposed workflow combines data preprocessing, feature extraction, AI-based detection, performance evaluation, and result interpretation to support the reliable identification of hallucinations in Large Language Models.



VIII. Conclusion

The research papers discussed in this review focus on different approaches to detecting hallucinations in Large Language Models (LLMs). These approaches include semantic consistency checking, uncertainty estimation, fact verification using external knowledge, semantic clustering, and explainable AI techniques. Each method has its own purpose, data requirements, advantages, limitations, and evaluation methods. Therefore, the results reported in these studies cannot always be compared directly because they may use different datasets, language models, detection techniques, and performance measures.

For future student projects, it is important to select a clear hallucination detection objective, use suitable datasets, and evaluate the system using appropriate performance measures. Understanding why a response is identified as potentially hallucinated and testing how the system performs on different language models and unseen data are also important. Overall, the effectiveness of a hallucination detection system should be judged by its actual experimental results, reliability, and ability to identify incorrect or unsupported responses rather than by expected improvements alone.

References

1. C. Woesle, L. Fischer-Brandies, and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," *IEEE Access*, vol. 13, pp. 148231–148253, 2025. DOI: 10.1109/ACCESS.2025.3601206.
2. M. Rahman and M. Chowdhury, "Hallucination Detection in Large Language Model Generated Texts: A Survey of Internal and External Approaches," in *2026 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, 2026. DOI: 10.1109/ICECTE69292.2026.11429437.
3. "Roberta with Low-Rank Adaptation and Hierarchical Attention for Hallucination Detection in LLMs," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2024. DOI: 10.1109/ICICML63543.2024.10957858.
4. S. Dasgupta, S. Nath, A. Basu, P. Shamsolmoali, and S. Das, "HalluShift: Measuring Distribution Shifts towards Hallucination Detection in LLMs," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8. DOI: 10.1109/IJCNN64981.2025.11228484.
5. "Assessing Hallucination in Large Language Models Under Adversarial Attacks," in *2024 Ninth International Conference on Mobile and Secure Services (MobiSecServ)*, 2024. DOI: 10.1109/MobiSecServ63327.2024.10759969.
6. "Semantic Reformulation Entropy for Robust Hallucination Detection in QA Tasks," in *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. DOI: 10.1109/ICASSP55912.2026.11460452.
7. "AI Hallucinations: A Misnomer Worth Clarifying," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024. DOI: 10.1109/CAI59869.2024.00033.
8. "SHIFT: Smoothing Hallucinations by Information Flow Tuning for Multimodal Large Language Models," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. DOI: 10.1109/ICCV51701.2025.00347.