IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 3, Ser. IV (May. – June. 2026), PP 18-30
www.iosrjournals.org

Deep Learning Approaches for Hallucination Detection in Large Language Models: A Comprehensive Survey

Redwana Farbin[1,3], Sakurah Ismail[1,3], Subrata Kumer Paul[2,3], Md. Ekramul Hamid[3]

[1]*North Bengal International University (NBIU), Chowddopai, Natore Road, Rajshahi-6206, Bangladesh.*
[2]*Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Natore-6431, Rajshahi, Bangladesh.*
[3]*Department of Computer Science and Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.*

Abstract:
The rapid advancement of Large Language Models (LLMs) has facilitated remarkable progress in natural language generation. Nevertheless, a major challenge remains, and this is the problem of hallucinations, where the models generate outputs that are linguistically valid but factually incorrect, logically unjustified, and even completely fabricated. The consequences of undetected hallucinations are severe, spanning patient safety risks in healthcare, fabricated legal precedents in judicial contexts, and distorted scholarly discourse in scientific research. Although significant research has focused on understanding and mitigating hallucinations in large language models, comparatively fewer studies have systematically examined deep learning-based detection methods and their associated benchmark datasets. This survey addresses that gap by presenting a structured taxonomy of detection methods organized into five paradigms: confidence and uncertainty-based detection, consistency-based detection, knowledge-grounded and graph-based detection, auxiliary model-based detection, and internal representation-based detection. For each paradigm, we analyze representative methods, datasets and outcomes. The insights synthesized in this survey provide guidance for designing reliable hallucination-aware LLM systems and support the development of adaptive detection–mitigation frameworks for trustworthy large-scale deployment.

Keywords: *Hallucination Detection, Deep Learning, Large Language Model, Model Reliability, AI Safety.*

Date of Submission: 15-06-2026 Date of Acceptance: 28-06-2026

I. Introduction

Large Language Models have quietly changed how we think about language and technology. What once required years of specialized training like writing code, drafting legal arguments, supporting clinical decisions, or simply holding a convincing conversation can now be approached by a machine with surprising fluency [1]. Models such as GPT-3, GPT-4, LLaMA 2, and Mistral have showcased remarkable capabilities in understanding and generating human-like language [2, 3]. Despite these advances, often these models hallucinate which is a serious problem. They can produce responses that sound completely convincing yet are factually wrong, misleading, or simply made up [4].

Hallucination in Large Language Models (LLMs) is a fundamental limitation arising from their probabilistic nature. Because these models are optimized for linguistic plausibility rather than factual correctness, they may generate fluent but inaccurate information based on learned statistical patterns [5]. Prior studies have shown that hallucinations occur across various NLG tasks, including summarization, translation, dialogue, and question answering [4].

The consequences of undetected hallucinations are severe and domain-specific. In healthcare, an LLM hallucinating drug dosages or misattributing clinical outcomes can directly endanger patient safety. In legal contexts, fabricated case citations the so-called "phantom precedent" problem can undermine judicial proceedings [6]. In scientific research, hallucinated references distort the integrity of scholarly discourse. As these systems are increasingly embedded into high-stakes pipelines, the inability to reliably distinguish factual from hallucinated content represents a fundamental barrier to their responsible adoption [5].

Motivated by the growing risks of hallucination, extensive research has focused on detecting unreliable LLM outputs using deep learning techniques. Early methods relied on heuristic metrics such as BLEU and ROUGE or reference-based comparisons requiring gold-standard answers [4], which proved limited in open-ended generation settings. As a result, research has shifted toward more robust deep learning-based strategies, broadly categorized into retrieval-based detection, uncertainty estimation, embedding-based analysis, supervised transformer

classifiers (e.g., BERT and DeBERTa), and self-consistency approaches that identify hallucinations through response variability across multiple generations [5, 7].

The rapid development of deep learning-based hallucination detection research reveals several important trends. Transformer models, originally developed for language understanding, are now commonly fine-tuned as dedicated hallucination detectors using benchmarks such as HaluEval [8] and TruthfulQA [9]. Retrieval-augmented methods further enhance reliability by grounding generated responses in external knowledge sources [5]. Similarly, graph-based frameworks such as GraphEval enable fine-grained claim verification through knowledge graph construction and NLI-based reasoning [10]. In addition, ensemble and multi-model approaches that combine semantic embeddings, NLI classifiers, and synthetic training data have shown strong detection performance even in low-resource settings [11].

Despite recent progress, hallucination detection in LLMs remains an ongoing challenge. Most existing surveys focus on hallucination types and mitigation strategies, while less attention has been given to analyzing detection methods from a deep learning architectural perspective. This survey addresses this gap by providing a structured taxonomy of deep learning-based hallucination detection methods, organized into five major methodological paradigms: confidence and uncertainty based, consistency based, knowledge-grounded and graph based, auxiliary model based, and internal representation-based approaches.

II. Research Methodology

The methodology adopted for systematically reviewing and analyzing prior literature on hallucination detection in Large Language Models (LLMs). The information synthesized in this survey was retrieved from several established academic databases and repositories, including Google Scholar, IEEE Xplore, Scopus, Springer, ACM Digital Library, arXiv. The search scope was restricted to deep learning-based methods specifically targeting hallucination detection in large language models. A predefined set of keywords was used in quotation form, including “Large Language Model,” “LLM hallucination,” “deep learning,” “hallucination detection techniques.” Only peer-reviewed articles and publicly available preprints written in English between 2022 and 2026 were considered. Studies were included if they (i) proposed a concrete hallucination detection framework or algorithm, (ii) employed deep learning techniques, internal representations, or LLM-based evaluation strategies, and (iii) reported quantitative experimental evaluation using standard benchmarks such as HaluEval, TruthfulQA, FEVER, GSM8K, MMLU, or other recognized datasets. During the screening process, duplicate entries were removed, followed by title and abstract filtering for relevance. Full-text screening was then performed to verify methodological novelty, experimental validation, and reproducibility of results. For each selected paper, structured attributes were extracted, including detection paradigm, access setting (black-box vs. white-box), supervision level (zero-shot vs. supervised), detection granularity (token, span, claim, or response level), reliance on multi-sampling, evaluation metrics (e.g. AUROC, AUPRC, F1-score), and computational overhead. Finally, the reviewed works were categorized into five methodological paradigms based on their primary detection signal and architectural principle: confidence and uncertainty-based detection, consistency-based detection, knowledge grounded and graph-based detection, auxiliary model-based detection, and internal representation-based detection. This taxonomy was developed iteratively to ensure conceptual coherence and to resolve overlaps by assigning each method to the category corresponding to its dominant detection mechanism.

Figure 1 illustrates the workflow of the review methodology used in this survey on hallucination detection in Large Language Models. The process begins with defining the research scope, followed by a keyword-based literature search across major academic databases such as Google Scholar, IEEE Xplore, Scopus, Springer, ACM Digital Library, and arXiv. The retrieved studies then undergo systematic screening through duplicate removal, title and abstract filtering, and full-text evaluation. Relevant information regarding detection methods, datasets, and evaluation metrics is extracted from the selected studies. Finally, the reviewed works are organized into a taxonomy of hallucination detection techniques, enabling comparative analysis and synthesis of the findings.

III. Hallucination detection methods

We categorize existing techniques into five major groups based on the source of detection signals and methodological principles. Where retrieval-based methods are categorized under knowledge-grounded & graph-based detection, uncertainty estimation under confidence and uncertainty-based detection, embedding-level analysis under internal representation-based detection, supervised classifiers under auxiliary model-based detection, and self-consistency under consistency-based detection.

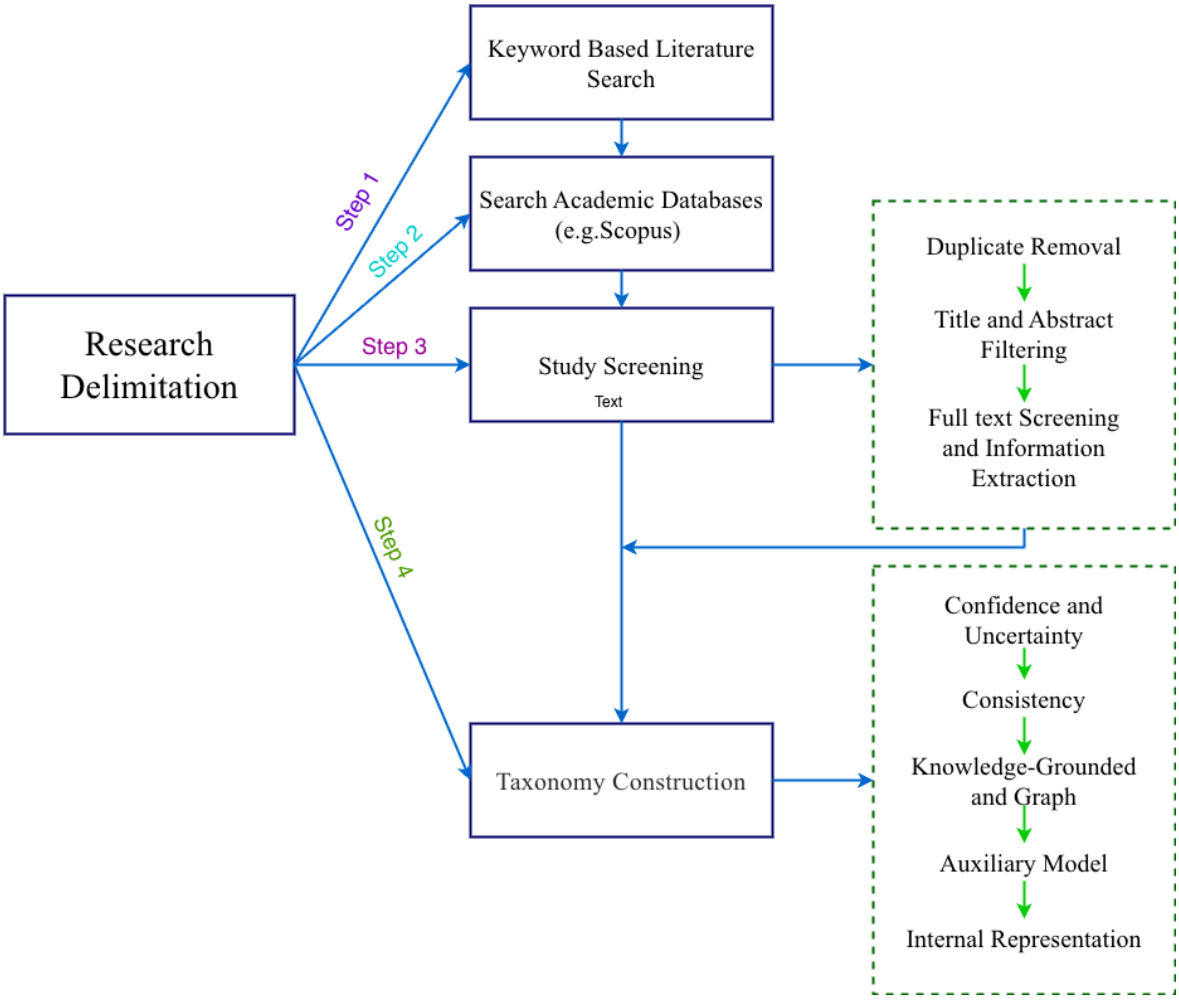

Figure 1: Survey Methodology Workflow.

3.1 Confidence and Uncertainty Based Detection

Confidence and uncertainty-based hallucination detection methods assess the reliability of model outputs using probabilistic signals such as confidence scores and entropy derived from the model itself. Since these approaches rely only on internal generation statistics, they are computationally efficient and well suited for black-box settings.

Early unsupervised hallucination detection methods relied on global uncertainty measures such as predictive entropy, negative log-likelihood, and perplexity. However, these sequence-level metrics often suffer from a dilution effect, where small hallucinated segments are hidden within longer high-confidence text. To overcome this limitation, Qiao et al. (2026) proposed Lowest Span Confidence (LSC) [12], a zero-shot metric that evaluates local token spans to identify low-confidence regions associated with factual errors using only a single forward pass, making it suitable for black-box APIs. Similarly, Moslonka et al. (2026) introduced Entropy Production Rate (EPR) [13], which estimates token-level uncertainty during generation, along with a supervised variant (WEPR) that improves detection under API-constrained settings without requiring multiple samples. Since entropy-based methods may fail for confidently generated hallucinations, Shelmanov et al. (2025) proposed Pre-trained Uncertainty Quantification (UQ) Heads [14], auxiliary modules trained on hallucination datasets to predict claim-level uncertainty using attention and hidden-state features, achieving strong cross-domain and cross-lingual performance with minimal overhead. To further improve efficiency, Kossen et al. (2024) introduced Semantic Entropy Probes (SEPs) [15], which approximate semantic uncertainty from a single generation using linear probes, reducing the computational cost of sampling-based approaches. Finally, Niu et al. (2025) proposed HaMI [16], which formulates hallucination detection as a multiple instance learning problem, adaptively selecting informative tokens and improving robustness against sparse hallucinated spans and varying sequence lengths.

Collectively, confidence and uncertainty-based methods are efficient and suitable for black-box settings but often struggle with confident hallucinations and domain shifts. Recent approaches, such as UQ heads and semantic entropy probes, improve robustness by integrating uncertainty with learned representations.

3.2 Consistency Based Detection

Consistency based hallucination detection methods assume that hallucinated responses show unstable semantic behavior under perturbations or repeated reasoning. Rather than relying on confidence or entropy signals, these approaches evaluate whether model outputs remain semantically consistent across different views or generations of the same query. Applicable in both black-box and white-box settings, they have recently emerged as effective training-free solutions for hallucination detection.

A representative zero-shot method in this category is Attention-Guided Self-Reflection (AGSER) [17], which uses attention information to create attentive and non-attentive query variants and measures consistency between their generated responses. By requiring only a few forward passes, AGSER achieves strong detection performance while reducing computational overhead compared to sampling-based approaches. A different strategy is proposed by MetaQA [18], which applies semantic perturbations using metamorphic relations such as synonym and antonym substitutions. Hallucinations are detected when these transformations violate factual consistency, leading to significant F1-score improvements over SelfCheckGPT across multiple datasets. Extending consistency analysis to internal representations, D²HScore [19] introduces a training-free white-box framework that evaluates semantic breadth and depth through intra-layer dispersion and inter-layer representation drift. This approach achieves competitive performance, with dispersion scores reaching an AUC of 0.74 on GSM8K and consistently outperforming other white-box baselines. In black-box settings, SINdex [20] measures semantic inconsistency through embedding-based clustering of multiple responses, achieving up to 9.3% AUROC improvement while providing substantial computational speedups over NLI-based methods. Consistency analysis has also been extended to multi-model scenarios through Consortium Consistency [21], which aggregates outputs from multiple LLMs and estimates disagreement using entropy over semantic clusters. Evaluated across 11 tasks and 15 models, this approach improves detection performance while reducing correlated hallucination errors among individual models.

3.3 Knowledge-Grounded & Graph Based Detection

Knowledge-grounded and graph-based hallucination detection methods identify factual inconsistencies by structuring generated content into interpretable relational forms, such as subject–relation–object triples. Unlike probabilistic or sampling-based approaches, these methods verify alignment between model outputs and source knowledge, enabling more fine-grained and interpretable hallucination detection.

A representative graph-based black-box method, FactSelfCheck [22], converts model outputs into knowledge graphs and evaluates factual consistency at the triple level by comparing multiple generated responses. This fact-level analysis improves hallucination correction, increasing factual content by 35.5%, and highlights that hallucinations often occur at individual claim levels rather than entire sentences. Similarly, Lie to Me [23] applies knowledge graph decomposition for self-detection, prompting models to verify individual facts instead of full passages. This structured reasoning improves interpretability and achieves up to 20% F1-score improvement over SelfCheckGPT without requiring external knowledge. Extending this idea, GraphEval+ [24] aligns generated and source content through entity–relation graphs combined with semantic similarity filtering and targeted NLI, enabling efficient and interpretable detection with visualization support for human auditing. Another related approach, the Q-S-E framework [25], detects hallucinations through structured question–answer alignment between generated summaries and source documents. Evaluations on CNN/DailyMail, PubMed, and ArXiv show improved factual consistency while maintaining summary quality.

Overall, knowledge-grounded and graph-based methods enable precise hallucination detection by operating at the level of atomic facts and aligning entities and relations in structured representations. These approaches improve interpretability and verification but may introduce additional computational overhead due to graph construction and semantic evaluation.

3.4 Auxiliary Model Based Detection

Auxiliary model-based hallucination detection methods employ additional models to evaluate the reliability of LLM outputs. Unlike intrinsic approaches that rely on confidence or consistency signals, these methods treat hallucination detection as a supervised or semi-supervised task using uncertainty features, hidden states, or reconstructed responses, enabling stronger discrimination and better generalization across tasks.

A representative auxiliary approach by Arteaga et al. [26] introduces a memory-efficient ensemble framework combining BatchEnsemble and LoRA to estimate epistemic and aleatoric uncertainty for hallucination detection. Using uncertainty features with a downstream classifier, the method achieves 97.8% accuracy on SQuAD v2 and 68% on MMLU, while significantly reducing computational cost. Moving beyond uncertainty estimation, HD-NDEs [27] model hallucination detection as a dynamic process by analyzing the evolution of hidden states across token sequences using neural differential equations. By capturing temporal representation changes, the framework achieves over 14% AUC-ROC improvement over prior methods. Another auxiliary strategy, InterrogateLLM [28], detects hallucinations through reconstruction consistency by asking the model to regenerate the original

query from its own answer. Deviations between reconstructed and original queries indicate hallucinations, achieving approximately 81% balanced accuracy without requiring labeled data or white-box access.

3.5 Internal Representation Based Detection

Internal representation-based hallucination detection analyzes signals available inside the model, such as hidden states and layer activations, instead of relying on output probabilities or external verification. Since intermediate representations retain rich semantic information, these methods can identify unreliable generations more effectively using lightweight probes or analytical metrics, often without requiring additional models or repeated sampling.

A representative approach in this category is INSIDE [29], which leverages internal model representations rather than output confidence for hallucination detection. Its EigenScore evaluates semantic consistency in embedding space and improves detection by analyzing activation patterns during inference. Building on this idea, Hallucination Detection with Internal Layers of LLMs [30] employs probing models that combine information from multiple transformer layers, showing that hallucination-related signals are typically concentrated in mid-to-late layers, although cross-model generalization remains challenging. Another direction focuses on spectral analysis of internal activations. EigenTrack [31] models hallucinations as temporal shifts in hidden-state geometry and detects early representation drift using spectral statistics, achieving strong AUROC improvements with a single forward pass. Similarly, Attention Head Embeddings with Deep Kernels [32] measure distributional differences between prompt and response embeddings using learned probabilistic distances, enabling efficient detection without external knowledge. Finally, LapEigvals [33] analyzes attention maps through spectral graph features, showing that disruptions in internal information flow provide reliable hallucination signals across datasets and models.

Overall, internal representation-based methods improve hallucination detection by directly analyzing hidden model dynamics rather than relying on output-level signals.

IV. Comparative studies and discussion

The reviewed studies reveal several important trends in hallucination detection research. Different detection paradigms demonstrate distinct trade-offs in terms of interpretability, benchmark datasets, and detection evaluation metrics. Table 1 presents a comprehensive comparison of representative methods across five major paradigms. The comparison summarizes detection methods, evaluation datasets, and reported performance as presented in the respective studies. Confidence and uncertainty-based approaches emphasize computational efficiency and compatibility with black-box settings, whereas consistency-based methods achieve strong detection performance by analyzing semantic stability across multiple responses. Knowledge-grounded approaches focus on fact-level verification and improve interpretability through structured reasoning. In contrast, auxiliary model-based and internal representation-based approaches often achieve higher detection accuracy by leveraging additional learning models or hidden-state representations. However, these methods may introduce higher computational complexity or require white-box access to model internals. These observations suggest that different paradigms offer complementary strengths, indicating the potential of hybrid hallucination detection frameworks that combine multiple detection signals for more robust performance.

Table 1: Comparative Analysis of Hallucination Detection Methods Across Different Paradigms

Paradigm	Detection Method	Dataset	Performance	Research Gap
Confidence and Uncertainty	Lowest Span Confidence (LSC) [12]	Natural Questions, TriviaQA, SQuAD 2.0, CoQA	Achieves highest AUROC across QA benchmarks and consistently outperforms existing zero-shot baselines.	LSC is evaluated only on QA-style short-answer tasks, leaving its effectiveness on long-form generation tasks unexplored. It can be extended with an adaptive window size mechanism to examine and improve its hallucination detection capability on long-form generation tasks.
	Entropy Production Rate (EPR/ WEPR) [13]	TriviaQA, WebQuestions, ArGiMi-Ardian Finance (RAG)	WEPR achieves ROC-AUC up to 93.6, consistently outperforming SelfCheckGPT and HalluDetect across multiple LLMs.	WEPR is trained only on short-answer QA tasks and explicitly acknowledges failure on "high-certainty hallucinations" where the model generates wrong facts with low entropy (high confidence), making the method blind to the most dangerous hallucination type. Combining WEPR's entropy-based signal with a lightweight semantic consistency check can help detect high-certainty hallucinations that entropy alone cannot capture.
	Pre-trained UQ Heads [14]	Multi-domain factual generation and multilingual biography datasets (e.g., Biographies, Cities, Movies, Books; Russian, Chinese, German)	Achieves state-of-the-art claim-level hallucination detection, outperforming unsupervised methods by up to 23 PR-AUC points and demonstrating strong cross-domain and cross-lingual generalization.	UQ Head requires white-box access to internal attention maps and token probabilities, making it completely inapplicable to black-box API-based LLMs (e.g., GPT-4, Claude) where such internal states are unavailable. Distilling the attention-based uncertainty signals from a white-box UQ Head into a lightweight black-box-compatible detector could extend hallucination detection capability to closed-source API settings.
	Semantic Entropy Probes (SEPs) [15]	TriviaQA, SQuAD, BioASQ, Natural Questions (NQ Open)	Achieves AUROC up to ~0.95 and outperforms accuracy probes in cross-task generalization while reducing computational cost by 5–10×	SEPs rely on a simple linear probe trained on single-token hidden states, which cannot capture the sequential and contextual uncertainty patterns that span across multiple tokens in a generated response, leaving span-level hallucination signals unexploited. Replacing the single linear probe with a lightweight sequential model (e.g., small Transformer or LSTM) trained on hidden states across multiple token positions could capture richer contextual uncertainty patterns and improve hallucination detection precision.
	HaMI [16]	TriviaQA, SQuAD, Natural Questions (NQ), BioASQ	Achieves AUROC up to 0.90+, significantly outperforming existing state-of-the-art hallucination detection methods across multiple QA benchmarks.	HaMI's uncertainty augmentation fuses uncertainty into token representations via a simple scalar multiplication applied uniformly across all hidden dimensions, which cannot selectively amplify the specific representation dimensions most informative for hallucination detection. Replacing the scalar fusion with a learned feature-wise attention gate would allow the model to selectively enhance hallucination-relevant dimensions, enabling more precise and discriminative token representations for detection
Consistency	Attention-Guided Self-Reflection (AGSER) [17]	Books, Movies, Global Country Information (GCI) datasets	Achieves best hallucination detection performance with AUC up to 0.988, consistently outperforming SelfCheckGPT, INSIDE,	AGSER uses Rouge-L as its sole consistency scoring metric to compare generated answers against the original, which is a surface-level lexical overlap measure that fails when semantically equivalent answers differ in wording, causing false hallucination detections

			and InterrogateLLM across multiple LLMs.	for correct paraphrases and false non-detections for hallucinations that happen to share surface tokens with the original answer. Replacing or augmenting Rouge-L with a semantic similarity measure for consistency computation would make the hallucination estimator robust to lexical variation while remaining sensitive to genuine semantic inconsistencies.
	MetaQA [18]	TruthfulQA-Enhanced, HotpotQA, FreshQA	Outperforms SelfCheckGPT with improvements of 0.041–0.113 (Precision), 0.143–0.430 (Recall) and 0.154–0.368 (F1-score) across multiple LLMs.	The MetaQA approach relies on synonym and antonym-based metamorphic mutations, which may not capture deeper semantic or reasoning inconsistencies in complex LLM responses. Incorporating reasoning-aware or semantic-structure mutations could enhance the detection of deeper factual inconsistencies beyond lexical mutations.
	D^2HScore [19]	GSM8K, TheoremQA, MMLU, Belebele (English), MGSM	Achieves highest AUROC and AUPR across benchmarks, outperforming existing training-free white-box baselines.	The proposed D^2HScore method requires access to internal hidden states of LLMs, which limits its applicability to white-box models and cannot be applied to closed-source LLMs such as GPT-4. Extending the framework by designing output-level approximations of semantic breadth and depth signals could enable similar hallucination detection capabilities while supporting black-box LLM environments.
	SINdex (Semantic INconsistency Index) [20]	TriviaQA, Natural Questions (NQ), SQuAD, BioASQ	Improves AUROC by up to 9.3%, achieving 0.87 (TriviaQA), 0.83 (NQ), 0.84 (SQuAD), and 0.94 (BioASQ) across multiple LLMs.	The SINdex framework relies on multiple sampled responses and semantic clustering, which increases inference cost and limits its practicality for real-time hallucination detection. Integrating single-pass semantic inconsistency estimation could reduce the need for multiple generations while enabling faster hallucination detection.
	Consortium Consistency [21]	GSM8K, GPQA-Diamond, TruthfulQA and 8 MMLU subsets (11 evaluation tasks)	Improves hallucination detection with +5.63% AUROC, +3.70% accuracy, and +5.39% AURAC, outperforming single-model consistency in over 92% of evaluated model teams	The proposed consortium consistency method relies on simple majority voting and entropy-based agreement across multiple LLMs, which may fail when several models share similar training biases and produce the same hallucinated answer. Incorporating model reliability weighting or confidence-aware aggregation during voting could improve hallucination detection by reducing the influence of biased or less reliable models while emphasizing more trustworthy predictions.
Knowledge-Grounded & Graph	FactSelfCheck [22]	WikiBio GPT-3 Hallucination Dataset, FavaMultiSamples	Achieves AUC-PR up to 93.41, outperforming SelfCheckGPT in fact-level detection and improving factual content by 35.5% during hallucination correction	Although FactSelfCheck enables fine-grained fact-level hallucination detection, its pipeline relies on multiple LLM-based steps, making the detection process computationally expensive and less efficient for large-scale deployment. Reducing detection complexity by merging knowledge-graph extraction and fact-consistency scoring into a single step or using lightweight structured extraction models could improve the efficiency and scalability of fact-level hallucination detection.

	Lie to Me: Knowledge Graphs for Robust Hallucination Self-Detection [23]	SimpleQA, WikiBio GPT-4o	Improves hallucination detection with up to 14% accuracy, 20% F1, and 7.5% AUC-PR gain; achieves F1 up to 0.87 and AUC-PR up to 0.88	Although the proposed method improves hallucination detection using knowledge graphs, it relies on self-detection by the same LLM that generated the response, which may introduce confirmation bias and reduce reliability when the model confidently produces hallucinated facts. Integrating cross-model verification or external semantic validation mechanisms during fact-level consistency analysis could improve hallucination detection by reducing bias from self-evaluation and providing more reliable detection signals.
	GraphEval+ (Graphing the Truth) [24]	SummEval (CNN/DailyMail summarization benchmark)	GraphEval achieves 71.5% balanced accuracy, while GraphEval+ reaches 53%	Although the proposed GraphEval+ framework improves interpretability through graph-based visualization, its hallucination detection performance is highly dependent on accurate triple extraction and NLI evaluation, making the method sensitive to extraction errors and increasing computational cost. Integrating robust semantic representations such as sentence-level embeddings or hybrid graph–embedding similarity measures could reduce dependence on fragile triple extraction and improve the reliability of hallucination detection.
	Q-S-E Framework [25]	CNN/DailyMail, PubMed, ArXiv	Improves hallucination detection with FactCC scores up to 38.14, consistently outperforming baseline LLM summaries across all datasets	The proposed QA-based hallucination detection relies heavily on ROUGE-based overlap and FactCC scoring for answer comparison, which may fail to detect semantically equivalent but lexically different answers. Replacing or augmenting ROUGE-based matching with semantic similarity or entailment-based scoring could improve hallucination detection by capturing deeper semantic inconsistencies beyond lexical overlap.
Auxiliary Model	BatchEnsemble + LoRA Uncertainty Framework [26]	SQuAD, SQuAD v2.0 (Faithfulness Detection), MMLU (Factual Detection)	Achieves 97.8% accuracy for faithfulness hallucination detection and 68% accuracy for factual hallucination detection using uncertainty-based classification.	The proposed uncertainty-based detection method struggles with factual hallucination detection and out-of-distribution (OOD) scenarios, showing significantly lower accuracy compared to faithfulness hallucinations. Incorporating external evidence verification or semantic consistency checks alongside uncertainty estimation could improve detection of factual hallucinations and enhance robustness in OOD settings.
	HD-NDEs (Hidden Dynamics Neural Differential Equations) [27]	True-False Dataset (Company*, Fact*, City*, Invention*), TruthfulQA, TriviaQA, HaluEval, Natural Questions	Improves hallucination detection with over 14% AUC-ROC gain on the True-False dataset and achieves strong detection performance (up to ~97% AUC-ROC) across multiple QA benchmarks.	The HD-NDEs method relies on internal hidden-state representations of LLMs, making it applicable mainly to open-source models and limiting its usability for black-box APIs. Developing a hybrid detection framework that approximates latent dynamics using observable signals could extend HD-NDE–style detection to black-box LLMs while preserving dynamic hallucination detection capability.
	InterrogateLLM [28]	Movies, Books, Global Country Information (GCI)	Achieves balanced accuracy ≈81% and AUC up to 0.87–1.00, outperforming semantic	InterrogateLLM assumes that hallucinated answers will produce inconsistent reconstructed queries, but hallucinated responses can sometimes remain internally

			similarity and self-consistency baselines in zero-resource settings.	consistent, reducing detection reliability. Integrating semantic fact verification alongside consistency checks could improve detection by identifying hallucinations that remain internally consistent but factually incorrect.
Internal Representation	INSIDE (EigenScore + Feature Clipping) [29]	CoQA, SQuAD v2.0, TriviaQA, Natural Questions (NQ), TruthfulQA	Achieves state-of-the-art hallucination detection, reaching AUROC up to 83.8% on SQuAD with 5–9% improvement over Perplexity, LN-Entropy, Lexical Similarity, and consistency baselines.	The method relies on multiple generated responses (K generations) to compute EigenScore, which increases inference cost and latency during detection. Developing a single-pass detection mechanism that estimates semantic divergence directly from internal states without requiring multiple generations could significantly reduce computational overhead while maintaining detection accuracy.
	Internal Layer Probing Framework [30]	TruthfulQA, HaluEval, ReFact	Achieves AUROC up to 0.86 on HaluEval, highlighting the effectiveness of mid-to-late transformer layers for hallucination detection.	The method relies on supervised training using labeled hallucination datasets, which restricts scalability to new domains with limited annotations. Incorporating self-supervised or contrastive learning objectives on internal representations could enable hallucination detection with minimal labeled data.
	EigenTrack [31]	HaluEval(Detection); WebQuestions & Eurlex (OOD evaluation)	Achieves AUROC of 0.82–0.94 across LLaMA, Qwen, Mistral, and LLaVA models, reaching 0.89 on LLaMA-7B and outperforming INSIDE, LapEigvals, and SelfCheckGPT.	EigenTrack requires white-box access to hidden activations of LLMs, limiting applicability to proprietary or API-based models where internal states are unavailable. Developing a hybrid detection framework that approximates spectral activation dynamics using observable signals could extend hallucination detection to black-box LLM settings.
	Attention Head Embeddings with Deep Kernels [32]	RAGTruth (QA & Summarization), CoQA, SQuAD	Achieves state-of-the-art hallucination detection, reaching ROC-AUC up to 0.988 (SQuAD) and 0.770 / 0.707 ROC-AUC on RAGTruth QA & Summ, consistently outperforming SelfCheckGPT, INSIDE, and uncertainty baselines	The method relies on internal hidden states and attention head embeddings of the LLM, restricting its applicability to white-box models and limiting use in API-based black-box systems. Developing proxy features that approximate attention-head or hidden-state signals using observable outputs could extend this detection approach to black-box LLM environments.
	LapEigvals (Spectral Attention Graph Detector) [33]	NQ-Open, TriviaQA, CoQA, SQuADv2, HaluEvalQA, TruthfulQA, GSM8K	Achieves state-of-the-art hallucination detection, outperforming AttentionScore, AttnLogDet, and AttnEigvals on 6 out of 7 datasets, reaching AUROC up to 0.925 across multiple LLMs.	The proposed method relies on internal attention maps and spectral features, which limits its applicability to white-box LLMs and models where internal attention weights are accessible, restricting deployment in closed or API-based systems. Developing approximate spectral indicators derived from observable outputs could extend spectral-based hallucination detection to black-box LLM environments while preserving detection capability.

The comparison also reveals two dominant methodological directions: training-free detection methods (LSC, AGSER, SINdex etc.) and supervised auxiliary approaches (UQ Heads, HD-NDEs, BatchEnsemble etc.). Training-free methods are attractive due to their simplicity and ease of deployment, whereas supervised approaches often achieve stronger detection performance by learning discriminative features from labeled hallucination datasets.

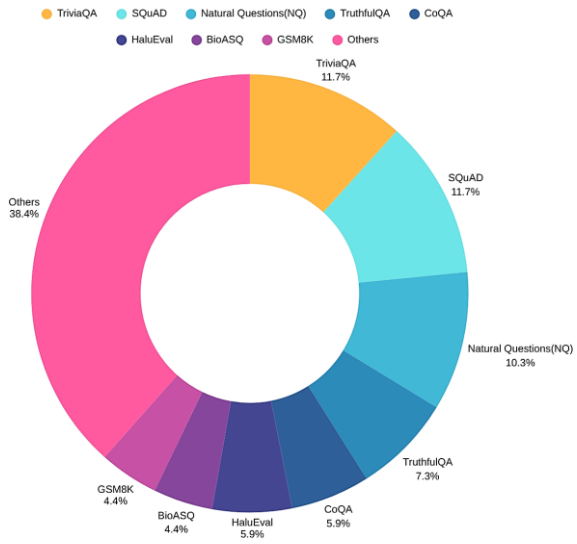

Figure 2: Dataset Distribution Across Hallucination Detection Methods (2022-2026).

The distribution of benchmark datasets used across the surveyed hallucination detection studies showed in Figure 2. The figure highlights the relative frequency of datasets employed in the evaluation of different detection methods. Among the datasets, TriviaQA and SQuAD appear most frequently, each accounting for approximately 11.7% of the total dataset usage, followed by Natural Questions (10.3%). These datasets are widely adopted due to their large-scale question answering benchmarks that enable evaluation of factual correctness and response reliability in large language models. Additionally, datasets specifically designed for evaluating factuality and hallucination behavior, such as TruthfulQA (7.3%) and HaluEval (5.9%), are commonly used in recent studies. Conversational and domain-oriented benchmarks including CoQA (5.9%), BioASQ (4.4%), and GSM8K (4.4%) also contribute to the evaluation of reasoning and domain-specific hallucination detection performance.

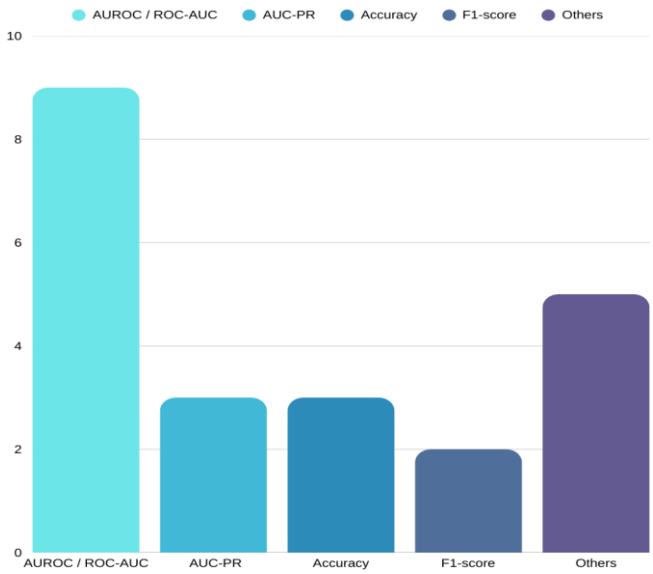

Figure 3: Frequency of evaluation metrics used across the hallucination detection methods summarized in Table 1.

A significant portion of the distribution (38.4%) falls under the "Others" category, which includes diverse datasets such as WikiBio, CNN/DailyMail, HotpotQA, RAGTruth, and several specialized evaluation benchmarks. This distribution indicates that hallucination detection research relies heavily on question-answering benchmarks while also incorporating a variety of task-specific datasets to evaluate robustness across different domains and generation settings.

As shown in Figure 3, AUROC/ROC-AUC is the most commonly used evaluation metric among the surveyed hallucination detection methods. This trend reflects a preference for threshold independent performance evaluation, which allows more reliable comparison across different detection models and datasets. Other metrics such as AUC-PR, accuracy, and F1-score appear less frequently and are generally used in task-specific evaluation settings.

V. Conclusion

Hallucination detection in large language models has matured into a diverse, multi-paradigm research area that includes uncertainty estimation, semantic consistency evaluation, knowledge-grounded verification, auxiliary supervision, and internal representation-based analysis. Each paradigm captures complementary aspects of hallucination behavior, from probabilistic confidence signals and semantic instability to structured fact alignment and latent activation dynamics. While uncertainty and consistency-based approaches are computationally efficient and suitable for black-box deployment, they may fail in cases of overconfident or reasoning-intensive hallucinations. Knowledge-grounded techniques enhance interpretability and factual precision but rely heavily on reliable source alignment. Auxiliary and internal representation-based methods often achieve stronger detection performance, yet they introduce additional architectural complexity or require white-box access.

Despite substantial progress, key challenges persist, including limited cross-model generalization, benchmark fragmentation, deployment constraints, and vulnerability to distribution shifts and adversarial prompting. This literature review is therefore essential to systematically organize existing methods, clarify their strengths and limitations, and identify research gaps across paradigms. By synthesizing current advancements, this survey provides a foundation for developing unified, adaptive hallucination-aware frameworks that integrate detection with mitigation strategies. Future research should prioritize hybrid and model-agnostic approaches, scalable black-box solutions, standardized evaluation protocols, and robust mechanisms capable of handling reasoning-based and real-world high-stakes applications, ultimately supporting the trustworthy and large-scale deployment of LLM systems.

References

[1]. L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, 2025. doi: https://doi.org/10.1145/3703155

[2]. H. Touvron *et al.*, "LLaMA 2: Open foundation and fine-tuned chat models," *arXiv preprint*, 2023. doi: https://doi.org/10.48550/arXiv.2307.09288

[3]. A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint*, 2023. doi: https://doi.org/10.48550/arXiv.2310.06825

[4]. Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023. doi: https://doi.org/10.1145/3571730

[5]. A. Alansari and H. Luqman, "Large language models hallucination: A comprehensive survey," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2510.06265

[6]. P. Sahoo *et al.*, "A comprehensive survey of hallucination in large language, image, video and audio foundation models," *Findings of EMNLP*, 2024. doi: https://doi.org/10.18653/v1/2024

[7]. P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," *Proc. EMNLP*, 2023. doi: https://doi.org/10.18653/v1/2023

[8]. Kong, L., Zhong, X., Chen, J. et al. Multi-perspective consistency checking for large language model hallucination detection: a black-box zero-resource approach. Front Inform Technol Electron Eng 26, 2298–2309 (2025). https://doi.org/10.1631/FITEE.2500180

[9]. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *Proc. ACL*, 2022. doi: https://doi.org/10.18653/v1/2022.acl-long.229

[10]. G. Hong *et al.*, "The hallucinations leaderboard: An open effort to measure hallucinations in large language models," *arXiv preprint*, 2024. doi: https://doi.org/10.48550/arXiv.2404.05904

[11]. C. Savelli *et al.*, "Leveraging synthetic data for LLM hallucination detection," *Proc. SemEval*, 2024. doi: https://doi.org/10.48550/arXiv.2403.00964

[12]. Y. Qiao *et al.*, "Lowest Span Confidence: A zero-shot metric for efficient and black-box hallucination detection in LLMs," 2026. Doi: https://doi.org/10.18653/v1/2024.emnlp-main.84

[13]. C. Moslonka *et al.*, "Learned hallucination detection in black-box LLMs using token-level entropy production rate," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2509.04492

[14]. A. Shelmanov *et al.*, "Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs," *Proc. EMNLP*, 2025. doi: https://doi.org/10.18653/v1/2025.emnlp-main.1809

[15]. J. Kossen *et al.*, "Semantic entropy probes: Robust and cheap hallucination detection in LLMs," *arXiv preprint*, 2024. doi: https://doi.org/10.48550/arXiv.2406.15927

[16]. M. Niu *et al.*, "Robust hallucination detection in LLMs via adaptive token selection," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2504.07863

[17]. Q. Liu *et al.*, "Attention-guided self-reflection for zero-shot hallucination detection in LLMs," in Proc. 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Suzhou, China, Nov. 2025, pp. 21005–21021, doi: https://doi.org/10.18653/v1/2025.emnlp-main.1063

[18]. B. Yang *et al.*, "Hallucination detection in large language models with metamorphic relations," *ACM Trans. Softw. Eng.*, 2025. https://dl.acm.org/doi/10.1145/3715735

[19]. Y. Ding *et al.*, "D2HScore: Reasoning-aware hallucination detection via semantic breadth and depth analysis in LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2509.11569

[20]. S. Abdaljalil *et al.*, "SINdex: Semantic inconsistency index for hallucination detection in LLMs," *arXiv preprint*, 2025. doi:

https://doi.org/10.48550/arXiv.2503.05980
[21]. D. Till *et al.*, "Teaming LLMs to detect and mitigate hallucinations," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2510.19507
[22]. A. Sawczyn *et al.*, "FactSelfCheck: Fact-level black-box hallucination detection for LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2503.17229
[23]. S. Kale and A. L. Alfeo, "Lie to Me: Knowledge graphs for robust hallucination self-detection in LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2512.23547
[24]. T. Agrawal *et al.*, "Graphing the truth: Structured visualizations for automated hallucination detection in LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2512.00663
[25]. Liu, S., Gao, Y., Li, S. et al. A hallucination detection and mitigation framework for faithful text summarization using LLMs. Sci Rep 16, 1374 (2026). https://doi.org/10.1038/s41598-025-31075-1
[26]. Arteaga *et al.*, (2025). *Hallucination detection in LLMs: Fast and memory-efficient finetuned models.* In *Proceedings of the 6th Northern Lights Deep Learning Conference (NLDL)* (pp. 1–15), PMLR. DOI**:** https://doi.org/10.48550/ARXIV.2409.02976
[27]. Kallem, P. (2026). Learning to Trust the Crowd: A Multi-Model Consensus Reasoning Engine for Large Language Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2601.07245
[28]. Li Zituo *et al, "*Survey of Hallucination Detection Methods for Large Language Models". Journal of Computer Research and Development, 2026, 63(1): 123-146. DOI: https://doi.org/10.7544/issn1000-1239.202550069
[29]. C. Chen *et al.*, "INSIDE: LLMs' internal states retain the power of hallucination detection," *arXiv preprint*, 2024. doi: https://doi.org/10.48550/arXiv.2402.03744
[30]. M. Preiss, "Hallucination detection with the internal layers of LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2509.14254
[31]. D. Ettori *et al.*, "EigenTrack: Spectral activation feature tracking for hallucination and OOD detection in LLMs and VLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2509.15735
[32]. R. Oblovatny *et al.*, "Attention head embeddings with trainable deep kernels for hallucination detection in LLMs," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2506.09886
[33]. J. Binkowski *et al.*, "Hallucination detection in LLMs using spectral features of attention maps," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2502.17598

BIOGRAPHIES OF AUTHORS (10 PT)

Redwana Farbin completed her Bachelor of Science (B.Sc.) degree in Computer Science and Engineering from the University of Rajshahi in 2025. She is currently working as a Lecturer at North Bengal International University (NBIU), Chowddopai, Rajshahi-6206, Bangladesh. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Large Language Models, Natural Language Processing, and Computer Vision. She can be contacted at email: redwanafarbi31@gmail.com.

Sakurah Ismail earned her Bachelor of Science degree in Computer Science and Engineering from the University of Rajshahi, Bangladesh. She is currently working as a Lecturer in the Department of Computer Science and Engineering at North Bengal International University. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Medical Image Segmentation, Computer Vision, and Natural Language Processing. She can be contacted at email: ismailsakurah@gmail.com.

Subrata Kumer Paul completed his B.Sc. and M.Sc. in Computer Science and Engineering from the University of Rajshahi in 2016 and 2017, respectively. Currently, he is serving as an Assistant Professor in the Department of Computer Science and Engineering at Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Natore-6431, Bangladesh. Recently, he has been awarded an MPhil (Master of Philosophy) degree in Computer Science and Engineering from the University of Rajshahi. Now, he is pursuing PhD from the same university. His research interests include Human Activity Recognition, Deep Learning, Artificial Intelligence, movement disorders (ASD, AD, PD), and Signal Processing. Currently, he received ICSETEP project as PhD fellowship. To date, he has published more than 30 research works in reputed international journals, conferences, and book chapters. He has actively participated in numerous symposiums, poster presentations, workshops, and seminars on research-related topics. Mr. Subrata has received a fellowship from the Information and Communication Technology (ICT) Division under the Ministry of Posts, Telecommunications, and Information Technology of Bangladesh. He has also been honored with Best Paper and Best Researcher awards for his outstanding research contributions. Mr. Subrata is a Graduate Member of IEEE. He can be contacted at email: sksubrata96@gmail.com

Dr. Md Ekramul Hamid received his B.Sc. and M.Sc. degrees in Applied Physics and Electronics from the University of Rajshahi. Later on, received an MCS degree from Pune University, India, and PhD degree from Shizuoka University, Japan. He is currently working as a professor at the Department of Computer Science & Engineering, University of Rajshahi, Bangladesh. He has published more than 70 international journal/conference papers. He is a recipient of the Monbukagakusho scholarship, JASSO Fellowship, and NIST fellowship for his contribution to Science and Technology. In 2026, he secured the ICSETEP project and is serving as the Principal Investigator (PI) of the project. He worked as a faculty member at King Khalid University, KSA in 2010-11 and visiting researcher at Shizouka University, Japan in 2012, 2014 and 2017 respectively. He worked as the Chairman of the CSE Department from September 2011 to June 2015 and as Dean of the Faculty of Engineering from April 2018 to November 2021 at the University of Rajshahi. His research interests include Audio signal processing, Speech enhancement, Machine Learning, and Image processing. He can be contacted at email: ekram_hamid@ru.ac.bd