A Systematic Literature Review on Hallucination Detection Methods in LLMs

Elisa Bestetti[1,*], Zheying Zhang[1] and Kostas Stefanidis[2]

[2]Data Science Research Centre, Tampere University, Finland

Abstract

Large Language Models (LLMs) are increasingly used across diverse applications, from healthcare and education to legal services and journalism. As their use expands into domains where factual accuracy is critical, the challenge of detecting and mitigating hallucinations has become essential. This research conducts a systematic literature review on hallucination detection methods in LLMs, restricting the scope to detection methods that do not rely on task-specific grounding documents provided at the time of generation (in RAG, summarization, or translation). Instead, the review focuses on open-domain detection, which includes methods that may use public external knowledge (for example Wikipedia or Wikidata) for post-hoc verification of the model's inherent outputs. The review synthesizes 50 peer-reviewed studies published between 2023 and 2025, classifying detection strategies according to hallucination type, i.e. factuality or faithfulness, and technical approach of white-box or black-box. The findings reveal a predominance of factuality-focused methods and black-box techniques, reflecting practical constraints in accessing proprietary model internals. Prominent approaches include LLM-as-a-judge, knowledge graph techniques and fact-checking with external knowledge. This study contributes a systematic synthesis of open-domain LLM hallucination detection methods that do not rely on task-specific grounding documents, providing a structured taxonomy across hallucination type and technical access model, and distilling dominant approaches and evaluation gaps to guide future research.

Keywords

Large language models, hallucination detection methods, systematic literature review

1. Introduction

Large Language Models are capable of generating sophisticated, coherent, and contextually relevant text across a wide range of applications. However, a significant weakness of these models is hallucination, a phenomenon in which the model generates factually incorrect, unfaithful, or nonsensical information that is not grounded in the provided source content or established world knowledge [1]. These fabrications can range from minor inaccuracies to completely invented facts, posing a substantial risk to user trust and model reliability, especially in domains where factual reliability is fundamental (such as medicine, law and journalism).

As the adoption of LLMs grows, the need to ensure their reliability and trustworthiness has become a concern for both researchers and practitioners. This research addresses this challenge by conducting a Systematic Literature Review (SLR) of methods designed to detect hallucinations in LLMs. The scope is specifically focused on textual hallucinations, and concentrates exclusively on detection methods. By mapping the existing landscape of detection techniques, this research aims to provide a clear and structured overview of the current state of research.

To guide this systematic review and ensure a comprehensive analysis, the following research questions has been formulated: **What hallucination detection methods have been proposed for Large Language Models in the existing literature?**

The taxonomy proposed by Huang et al. [2] guides the analysis of this research. In the study, hallucinations are categorized into two main types: factuality hallucination and faithfulness hallucination. Factuality hallucination occurs when the model generates content that either contradicts established

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland

*Corresponding author.

elisa.bestetti@sanoma.com (E. Bestetti); zheying.zhang@tuni.fi (Z. Zhang); konstantinos.stefanidis@tuni.fi (K. Stefanidis)

0000-0002-6205-4210 (Z. Zhang); 0000-0003-1317-8062 (K. Stefanidis)

real-world knowledge or cannot be confirmed as accurate. Conversely, faithfulness hallucination refers to the extent to which an LLM's output remains aligned with the user's instructions, the given context, and maintains internal coherence.

Prior research has examined hallucinations in large language models, though with different scopes. Luo et al. [3] review detection and mitigation methods, including token- and sentence-level approaches. Their work is a general survey rather than a formal SLR and includes tasks beyond open-domain settings, such as summarization and translation. It also predates most studies analyzed here. Huang et al. [2] propose a taxonomy distinguishing factuality and faithfulness and discuss causes across data, training, and inference stages. While comprehensive, their survey spans mitigation and benchmarks, whereas this research focuses solely on detection.

The paper is divided as follows: Section 2 describes the research method, Section 3 presents the key findings of the research, Section 4 discusses the findings and Section 5 concludes the work.

2. Research Method

This research employs the systematic literature review (SLR) methodology proposed by Kitchenham and Charters [4], which define three main phases: planning, conducting, and reporting the review.

2.1. Search Strategy

The search was conducted with a metadata-focused search (*title*, *abstract*, *keywords*; or *abstract* where needed) across four subscription-accessible digital libraries in computer science and information technology: ACM Digital Library, Computer Science Database (ProQuest), IEEE Xplore (IEL), and Scopus (Elsevier). Searches were executed on October 7, 2025. The search string used was:

```
("language model*" OR LLM* OR "generative AI" OR "foundation model*" OR "
transformer model*" OR "generative model*") AND ("hallucination* detection"
OR "hallucination* identification" OR "hallucination* evaluation" OR
"fact-checking" OR "fact checking")
```

Inclusion and exclusion criteria were defined broad enough to capture open-domain detection research while excluding domain-bound or grounding-dependent approaches. The complete list is in Table 1.

Table 1
Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Literature published between 2018 and the search date	Literature not written in English
Literature written in English	Literature researching text generated in other languages
Literature researching text generated in English	Blogs or other gray literature
Articles, conference proceedings, or journals	Not peer-reviewed
Peer-reviewed literature	Not available in the selected libraries/databases
Accessible via the selected digital libraries/databases	Not relevant to the first research questions
Relevant to the first research questions	Does not have content to answer both research questions
Allows answering both research questions	Researching *multimodal* large language models
Researching large *textual* language models	Hallucination detection not the main topic
Hallucination detection is the main topic	Method is domain-restricted
Method designed for use in an open knowledge domain	Methods where the truth source is provided as part of the input
Method can use public, general-purpose knowledge bases for post-hoc	(summarization, translation, or RAG)
verification but the truth source is not part of the input	Method needs grounding documents
Method does not need grounding documents	Addresses deliberately generated misinformation or intentionally
Focuses on unintentional hallucinations arising from model limitations	deceptive content

Following the protocol, screening proceeded in three stages: (i) *title screening* removed clearly irrelevant entries; (ii) *abstract screening* excluded works outside the open-domain detection scope; (iii) *quality assessment* excluded studies failing the predefined threshold. Four otherwise relevant papers were inaccessible via institutional subscriptions and could not be obtained from authors. In total, 748 studies were retrieved, of which 50 studies were included in the final corpus. The selected studies are

listed in the references as items [5]-[55] and are explicitly marked with paper IDs in the form L* to enable traceability to the raw data recorded in the spreadsheet.

2.2. Data Extraction

To support transparency and consistent synthesis, a predefined data-extraction form was used. For each included study, the following information was recorded: bibliographic identifiers, hallucination type addressed, detection granularity (token/sentence/passage), access to model internals, dependence on external references, method classification and main technique, a brief method summary, detector output, evaluation setup and evaluation metrics. Data were extracted in chronological order and consolidated into a common spreadsheet for analysis, which is publicly available on GitHub[1].

3. Results

This section answers the research question: what hallucination detection methods have been proposed for Large Language Models in the existing literature?

Following the taxonomy proposed by Huang et al. [2], hallucination detection strategies are first divided into two fundamental categories: Factuality hallucination detection, which identifies factual errors in outputs, and faithfulness hallucination detection, which identifies the faithfulness of outputs to the provided context.

The majority of the studies, 37 in total, proposed methods for factuality hallucination detection, while 12 focused on faithfulness hallucination detection. In some cases, the distinction between the two categories was not clear-cut, especially with methods that had application in different domains and different phases. One study [5] clearly addressed both strategies.

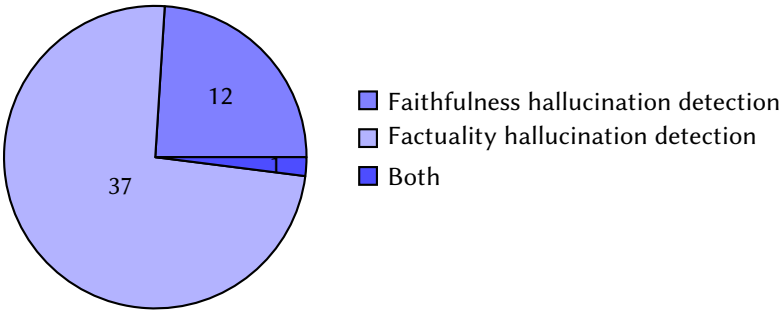

Figure 1: Hallucination detection method types.

Despite this primary classification, another important technical distinction between methods was found: 16 methods needs to access LLM internals or token probabilities (white-box), while 34 methods

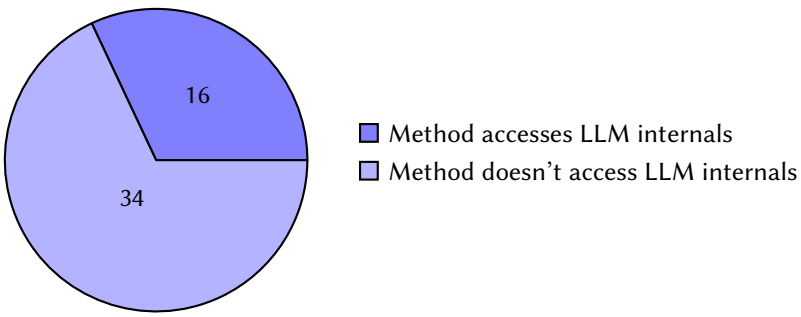

Figure 2: Hallucination detection method access to LLM internals.

relies solely on the LLM's outputs (black-box).

3.1. White-box hallucination detection methods

White-box approaches, represented by 16 studies, leverage internal signals such as hidden activations, attention patterns, gradients, or token-level probabilities , including logits and entropy. These signals support finer-grained detection at the token or claim level and often enable lightweight inference without external calls. Some methods combine multiple internal cues , such as embeddings and entropy, to improve robustness. Detection granularity ranges from token-level ([6], [7]) to sentence/passage-level ([8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]), with some methods supporting both ([20], [5]). All surveyed white-box approaches are reference-free and do not rely on external retrieval, making them applicable to tasks centered on internal model confidence and uncertainty.

Table 2
White-box hallucination detection methods

Detection category	Methods	Studies
LLM internal states	HaloScope, INSIDE, InternalInspector, Layer-wise classifiers, MIND, Lookback Lens	[9] [10] [11] [13] [6] [14]
Token probabilities	Claim-Conditioned Probability (CCP), Entropy-based Scores, Bayesian Semantic Entropy	[20] [15] [18]
Hybrid (internal + logit)	EGH, LLM-Check, HADEMIF, MixHD, Semantic Consistency Analysis, BatchEnsemble + LoRA, AMR Graphs	[8] [12] [5] [7] [16] [17] [19]

3.2. Black-box hallucination detection methods

Black-box methods (34 studies) do not rely on internals and instead analyze generated text and, optionally, external evidence. Two broad families emerge: (1) **Knowledge-based** methods that retrieve external information, such as web, Wikipedia/Wikidata, and Resource Description Framework (RDF) knowledge graphs to fact-check outputs. RDF is a standard data model for representing information as structured 'triples' (Subject-Predicate-Object), which allows for precise automated verification; and (2) **Reference-free** methods that judge reliability using the model's own behavior, auxiliary LLMs, structured representations, or consistency signals without retrieval.

3.2.1. Knowledge-based methods

The 11 knowledge-based methods target factuality hallucinations and generally follow a fact-checking paradigm with external retrieval. These methods are in studies [21], [22], [23], [24], [25], [26], [27], [28], [29], [30] and [31]. Knowledge is sourced primarily from Wikipedia/Wikidata or web search, while some approaches leverage RDF-based knowledge graphs, such as DBpedia, LODsyndesis, OpenDialKG.

Two main phases dominate these methods: fact extraction and verification. Extraction typically decomposes LLM outputs into atomic claims or constructs knowledge graphs, enabling finer-grained validation. Verification then matches these units against retrieved evidence using semantic entailment or graph-based techniques. NLI classifiers and web search are common for textual claims, whereas KG-based approaches employ structural matching and embedding-based similarity.

Although most methods adhere to this pipeline, some integrate alternative architectures combining multi-source retrieval, fusion, and decision-making without explicit decomposition. These systems often incorporate specialized modules for evidence aggregation and verdict generation, aiming to improve robustness in detecting and mitigating hallucinations.

[1] https://github.com/Eelsie/master_thesis_data/blob/main/extraction_forms.csv

3.2.2. Knowledge graph: knowledge-based and reference-free methods

As shown in Table 3, seven studies propose methods based on knowledge graphs as the primary technique: [21], [22], [24], [32], [33], [34], and [5]. Among these, [21], [22] and [24] rely on external knowledge in the form of knowledge bases that utilize the RDF data model, such as DBpedia, LODsyndesis, and OpenDialKG. These methods verify model-generated facts against established knowledge bases. They transform LLM outputs into triplets (typically subject–predicate–object) statements and cross-check these against curated or dynamically retrieved graphs.

ConFcheKG [21] links entities to multiple knowledge graphs and scores reliability by contrasting intersecting versus conflicting subgraphs: low coherence signals hallucination. GPT LODS [24] prompts the LLM to produce RDF triples for a question, then checks those triples in real time against DBpedia or LODsyndesis. [22] differs because it proposes a method based on Graph Neural Networks: given a dialogue history and its generated response, it extracts entities or relations from the response and retrieves KB facts about those entities to form a reference. It then encodes both graphs with RGAT, pools graph features, and performs binary classification.

In contrast, the methods reported in [32], [33], [34] and [5] are knowledge-free, relying only on the provided context or the model's internal knowledge. FactAlign [32] constructs graphs from the source text and the output, aligns their triples, and flags low-alignment facts. GraphEval [33] also turns outputs into triples but tests each against the given context, improving base NLI detectors while returning the offending triples. GCA [34] builds a graph for each response, models dependencies between facts with an RGCN, and combines multi-sample consistency with reverse verification to score hallucination. Finally a semantic-graph uncertainty approach [5] uses AMR graphs to propagate uncertainty across entities and sentences and calibrates it with NLI-based contradiction signals.

Table 3
Knowledge graph-based black-box methods

Core technique	Knowledge source	Study
Subgraph coherence	External	[21]
RDF verification	External	[24]
GNN (RGAT)	External	[22]
Graph alignment	Free	[32]
Triples + NLI	Free	[33]
GNN (RGCN)	Free	[34]
AMR Graphs	Free	[5]

3.2.3. LLM-as-a-judge: knowledge-based and reference-free methods

Seventeen methods use LLM-as-a-judge as their main technique or an important part of their method. LLM-as-a-judge is the practice of using a large language model as an automatic evaluator of other models' outputs, typically by prompting it to compare or grade responses on open-ended tasks [35]. The detection granularity of the methods is at the sentence or passage level. In some methods, claims are first extracted and verified in a second phase.

Among these studies, six studies use external references to judge the veracity of the outputs. These methods decompose text into claims, retrieve supporting evidence from the web or knowledge bases, and use an LLM to verify the consistency between each claim and the retrieved evidence.

The remaining eleven studies rely exclusively on the model's internal knowledge and adopt four distinct strategies:

- *Self-consistency* methods treat hallucinations as instability in the model's own behavior: if the model doesn't give consistent answers when asked the same question multiple times under slightly different conditions, it's likely hallucinated.
- *Metamorphic-testing* methods also treat hallucinations as instability in the model's own behavior

by applying systematic transformations (metamorphic relations) to the input and checking whether the output changes in a predictable way.

- *LLM-judge classifiers and ensembles* methods, where the LLM is prompted as a classifier with no external retrieval: reliability comes from prompt design, few-shot examples and ensemble voting.
- *LLM to train data* methods don't want to call a LLM at inference. Instead, they use a strong LLM as a teacher to label data (or generate reliability signals) and then train a smaller model or ensemble of models.

Table 4 summarizes the LLM-as-a-judge methods by strategy and granularity.

Table 4
Comparison of LLM-as-a-judge methods by strategy and granularity

Strategy	Granularity	Studies
External knowledge retrieval	Claim	[26] [28] [31]
	Sentence	[29]
	Passage	[25] [27]
Self-consistency	Sentence	[36] [37] [38]
Metamorphic-testing	Sentence	[39]
	Passage	[40]
LLM-judge classifiers and ensembles	Sentence	[41] [42] [43]
	Passage	[44]
LLM to train data	Sentence	[45] [46] [43]

3.2.4. Other techniques for knowledge-free black-box methods

Beyond LLM-as-a-judge and knowledge graphs, the selected studies present several additional methodologies for hallucination detection that operate without access to model internals or external knowledge sources. These approaches, nine in total, can be categorized into four groups.

- *Reverse validation* method in [47] and [48] assesses reliability by reconstructing the original query from the generated answer; hallucinations are indicated by reconstruction mismatches. For example, InterrogateLLM [47] draws on human interrogation techniques, using consistency across repeated questioning as an indicator of truthfulness.
- *Uncertainty estimation* method detects hallucinations by quantifying uncertainty in generated text. For example, [49] paraphrases query into multiple scenarios and applies factor analysis to separate shared semantic uncertainty from scenario-specific variation, while [50] computes token-level uncertainty using negative log-likelihood and entropy over informative keywords.
- *Natural Language Inference (NLI)* method casts detection as entailment or contradiction between model outputs and task-specific textual anchors.[51] performs zero-shot detection by checking entailment between source, hypothesis, and target using pretrained NLI models, while [52] evaluates internal consistency by decomposing responses into claims and measuring support or contradiction across multiple sampled answers.
- *Synthetic data* method is to generate faithful/hallucinated pairs [45] or weak labels [53] with LLMs, then fine-tune discriminative detectors such as DeBERTa/RoBERTa ([53], [54]) or LoRA adapters [54].

4. Discussion

This research systematically reviewed hallucination detection methods for LLMs in open-domain settings, where no grounding documents are available. The review analyzed 50 peer-reviewed studies

published between 2023 and 2025. The research landscape is recent and rapidly evolving: no studies were found prior to 2022, and 2024 emerged as the most prolific year, reflecting the surge in interest following widespread LLM adoption.

Majority of detection methods addressing factuality. Following the taxonomy of [2], the research revealed a prominence of factuality hallucination detection (37 studies) over faithfulness detection (12 studies). This imbalance suggests that the primary concern in open-domain generation is the fabrication of world knowledge (entity errors, factual fabrications) rather than internal logical consistency. This focus is also intuitive given the open-domain nature of the selected studies: without grounding documents (as used in RAG, translation or summarization), the *truth* must be derived from the model's internal parametric knowledge or from external verification. However, the distinction is becoming blurred in knowledge graph -based approaches. These methods transform text into knowledge graphs or semantic graphs, treating factuality and faithfulness as the same mathematical problem: graph alignment or consistency.

Predominance of black-box methods. A significant finding of this review, is the majority of black-box methods (34 studies) over white-box methods (16 studies). This trend is probably a consequence of the commercial reality of the LLM landscape: as the most capable proprietary models (GPT, Claude, Gemini) are accessible primarily via API with no access to internal weights, logits, attention maps, gradients or hidden activations, researchers have been forced to innovate outside the model architecture. White-box methods generally offer finer granularity (token-level detection), but their detection is limited to open-source models (for example, LLaMA and Mistral). By accessing the model's uncertainty directly, these methods also avoid the latency and computational expense of generating multiple external outputs. More capable proprietary models require also more computationally expensive (black-box) detection methods.

LLM-as-a-judge is the most used detection technique. Seventeen methods use LLM-as-a-judge approaches, where one language model evaluates the outputs of another. This evaluation strategy reflects both the capabilities and limitations of current AI systems. On one hand, LLMs possess the linguistic sophistication and reasoning ability to make nuanced judgments about factuality and consistency; on the other hand, LLM-as-a-judge methods inherit the very vulnerabilities they aim to detect. Evaluator models may themselves hallucinate, exhibit biases, or demonstrate inconsistent judgment, introducing the critical risk of recursive hallucinations. This phenomenon occurs when the evaluator model, in the process of assessing another model's output, generates its own hallucinations. If the judge model lacks the necessary world knowledge or shares the same inductive biases as the target model, it may incorrectly validate a hallucinated claim (a false negative) or flag a correct statement as an error (a false positive). The literature attempts to mitigate this in different ways: grounding the judge in external evidence, querying multiple judges and using a majority vote, using self-consistency (ask the judge the same question under paraphrases or perturbations), combining LLM judge with other non LLM judgments or using LLM judges only to create supervision for separate detectors.

External knowledge detection in black-box methods. Among black-box methods, 11 studies rely on external references. External-knowledge methods combines fact extraction into atomic claims or triples and verification against web search, Wikipedia/Wikidata, or RDF knowledge graphs. These methods are particularly suited to domains where the relevant information is well covered by public knowledge bases. However, they are vulnerable to the same knowledge boundaries that affect the underlying LLMs: long-tail, up-to-date, or copyright-restricted knowledge remains difficult to verify. They also introduce engineering complexity and runtime cost due to retrieval and reasoning over external sources.

Reference-free detection in black-box methods. Zero-knowledge detection is attractive because it avoids dependencies on external infrastructure and can be applied in settings where retrieval is unavailable, unreliable, or undesirable. LLM-as-a-judge in this category is the most used technique (11 methods), along with knowledge graphs (4 methods). Moreover, other interesting minority approaches emerged: reverse validation, uncertainty estimation, Natural Language Inference and generation of synthetic data.

Broad diversity in evaluation benchmarks. Studies rely on a variety of benchmarks, many of which differ substantially in task formulation, annotation protocol, size, and granularity. Some detectors are evaluated primarily on question answering datasets (TruthfulQA [40][16], FreshQA [28][40], TriviaQA [10][50] and SQuAD [10][17]), others on summarisation corpora (xSum [49][32], SummEval [33]), and others on dedicated detection datasets like HaluEval [8][55][26][11][27][29][30][31], SelfCheckGPT [49][23][8][55][12][40][30] or SHROOM [51][53][41][46][44]. This diversity complicates systematic comparison across methods and reported performance is often tightly coupled to the characteristics of a particular benchmark. Further research could benefit from standardized benchmarks and evaluation protocols.

Threats to validity. Several limitations should be considered when interpreting these findings. The review was conducted by the first author, with guidance from the other two authors. Although the literature search, study selection, and data extraction were discussed among the authors, the risk of subjectivity remains, particularly in study selection, quality assessment, and data extraction. Involving multiple reviewers would likely have improved reliability. The exclusion of gray literature ensures a focus on peer-reviewed quality but may omit relevant preprints. Additionally, four potentially suitable papers were inaccessible despite attempts to contact the authors. Determining whether studies addressed open-domain hallucination detection was occasionally ambiguous, potentially introducing selection bias. Finally, categorizing methods as targeting factuality or faithfulness required interpretive judgment, particularly for multi-technique and multi-granularity approaches such as knowledge-graph–based methods. While predefined forms and established taxonomies helped mitigate misclassification risks, some ambiguity remains.

5. Conclusion

This study provides a comprehensive overview of hallucination detection methods for LLMs in open-domain settings. The findings reveal a rapidly growing and diverse research field, with a strong emphasis on factuality detection, a predominance of black-box approaches, and widespread reliance on LLM-as-a-judge techniques. The findings reveal both progress and persistent challenges. Detection methods are diverse, intending to resolve the balance between scalability, accuracy, and independence from external resources. White-box approaches offer more precision but lack applicability to closed models. Black-box methods are broadly deployable but often computationally expensive or reliant on imperfect judges.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI ChatGPT (GPT-4 and GPT-5 models), Anthropic Claude (Sonnet 4.5) in order to: Improve the use of spelling and grammar throughout the text, synthesize or paraphrase complex concepts for comparison with own understanding.

References

[1] Z. Ji, N. Lee, R. Frieske, D. S. Tiezheng Yu, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Computing Surveys, 2022.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. (2025).

[3] J. Luo, T. Li, D. Wu, M. Jenkin, S. Liu, G. Dudek, Hallucination detection and hallucination mitigation: An investigation, arXiv preprint arXiv:2401.08358 (2024). URL: https://arxiv.org/abs/2401.08358.

[4] B. Kitchenham, S. M. Charters, Guidelines for performing systematic literature reviews in software engineering, Keele University and Durham University Joint Report (2007).

[5] K. Chen, Q. Chen, J. Zhou, X. Tao, B. Ding, J. Xie, M. Xie, P. Li, F. Zheng, Enhancing uncertainty modeling with semantic graph for hallucination detection (2025). doi:10.1609/aaai.v39i22.34528, **L52**.

[6] M. Suresh, R. Aljundi, I. Nkisi-Orji, N. Wiratunga, Towards improving open-box hallucination detection in large language models (llms) (2024). **L47**.

[7] X. Zhou, M. Zhang, Z. Lee, W. Ye, S. Zhang, Hademif: Hallucination detection and mitigation in large language models (2025). **L55**.

[8] X. Hu, Y. Zhang, R. Peng, H. Zhang, C. Wu, G. Chen, J. Zhao, Embedding and gradient say wrong: A white-box method for hallucination detection (2024). doi:10.18653/v1/2024.emnlp-main.116, **L16**.

[9] X. Du, C. Xiao, Y. Li, Haloscope: Harnessing unlabeled llm generations for hallucination detection (2024). **L25**.

[10] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, J. Ye, Inside: Llms' internal states retain the power of hallucination detection (2024). **L28**.

[11] M. Beigi, Y. Shen, R. Yang, Z. Lin, Q. Wang, A. Mohan, J. He, M. Jin, C.-T. Chang-Tien, L. Huang, Internalinspector i2: Robust confidence estimation in llms through internal states (2024). doi:10.18653/v1/2024.findings-emnlp.751, **L29**.

[12] G. Sriramanan, S. Bharti, V. Sadasivan, S. Saha, P. Kattakinda, S. Feizi, Llm-check: Investigating detection of hallucinations in large language models (2024). **L32**.

[13] Y.-S. Chuang, L. Qiu, C.-Y. Hsieh, R. Krishna, Y. Kim, J. Glass, Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps (2024). doi:10.18653/v1/2024.emnlp-main.84, **L33**.

[14] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, Y. Liu, Unsupervised real-time hallucination detection based on the internal states of large language models (2024). doi:10.18653/v1/2024.findings-acl.854, **L49**.

[15] E. Joo, Y.-J. Lee, H.-J. Choi, Entropy-based sentence-level hallucination score in large language models (2025). doi:10.1109/BigComp64353.2025.00022, **L53**.

[16] R. B. Beyene, F. Faghih, T. A., Hallucination detection in llms via beam search sampling and semantic consistency analysis, 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W) (2025). doi:10.1109/DSN-W65791.2025.00076, **L57**.

[17] G. Arteaga, T. Schön, N. Pielawski, Hallucination detection in llms: Fast and memory-efficient fine-tuned models (2025). **L58**.

[18] K. Ciosek, N. Felicioni, S. Ghiassian, Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy, Transactions on Machine Learning Research (2025). **L59**.

[19] C. Li, B. Xing, D. Huo, Q. Zhou, Z. Xu, Y. Wang, Mixhd: A method for detecting hallucinations based on the internal state and output probability of large language models, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025). doi:10.1109/ICASSP49660.2025.10889328, **L61**.

[20] E. Fadeeva, A. Rubashevskii, A. Shelmanov, S. Petrakov, H. Li, H. Mubarak, E. Tsymbalov, G. Kuzmin,

A. Panchenko, T. Baldwin, Fact-checking the output of large language models via token-level uncertainty quantification (2024). doi:10.18653/v1/2024.findings-acl.558, **L18**.

[21] H. Lee, M. Sohn, Context-based fact-checking using knowledge graph (2023). doi:10.1109/BigData59044.2023.10386121, **L3**.

[22] K. Furumai, Y. Wang, M. Shinohara, K. Ikeda, Y. Yu, T. Kato, Detecting dialogue hallucination using graph neural networks, 2023 International Conference on Machine Learning and Applications (ICMLA) (2023). doi:10.1109/ICMLA58977.2023.00128, **L4**.

[23] X. Wang, Y. Yan, L. Huang, X. Zheng, X. Huang, Hallucination detection for generative large language models by bayesian sequential estimation (2023). doi:10.18653/v1/2023.emnlp-main.949, **L6**.

[24] M. Mountantonakis, Y. Tzitzikas, Real-time validation of chatgpt facts using rdf knowledge graphs (2023). **L11**.

[25] X. Chen, D. Song, H. Gui, C. Wang, N. Zhang, Y. Jiang, F. Huang, C. Lyu, D. Zhang, H. Chen, Factchd: Benchmarking fact-conflicting hallucination detection (2024). **L20**.

[26] Y. Wang, R. Reddy, Z. Mujahid, A. Arora, A. Rubashevskii, J. Geng, O. Afzal, L. Pan, N. Borenstein, A. Pillai, Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers (2024). doi:10.18653/v1/2024.findings-emnlp.830, **L21**.

[27] X. Zhao, J. Yu, Z. Liu, J. Wang, D. Li, Y. Chen, B. Hu, M. Zhang, Medico: Towards hallucination detection and correction with multi-source evidence fusion (2024). doi:10.18653/v1/2024.emnlp-demo.4, **L37**.

[28] H. Iqbal, Y. Wang, M. Wang, G. Georgiev, J. Geng, I. Gurevych, P. Nakov, Openfactcheck: A unified framework for factuality evaluation of llms (2024). doi:10.18653/v1/2024.emnlp-demo.23, **L41**.

[29] X. Cheng, J. Li, W. Zhao, H. Zhang, F. Zhang, D. Zhang, K. Gai, J.-R. Wen, Small agent can also rock! empowering small language models as hallucination detector (2024). doi:10.18653/v1/2024.emnlp-main.809, **L45**.

[30] S. Heo, S. Son, H. Park, Halucheck: Explainable and verifiable automation for detecting hallucinations in llm responses, Expert Systems with Applications (2025). doi:10.1016/j.eswa.2025.126712, **L60**.

[31] X. Sun, J. Li, Y. Zhong, D. Zhao, R. Yan, Towards detecting llms hallucination via markov chain-based multi-agent debate framework, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025). doi:10.1109/ICASSP49660.2025.10889448, **L63**.

[32] M. Rashad, A. Zahran, A. Amin, A. Abdelaal, M. AlTantawy, Factalign: Fact-level hallucination detection and classification through knowledge graph alignment (2024). doi:10.18653/v1/2024.trustnlp-1.8, **L19**.

[33] H. Sansford, N. Richardson, H. Maretić, J. Saada, Grapheval: A knowledge-graph based llm hallucination evaluation framework (2024). **L22**.

[34] X. Fang, Z. Huang, Z. Tian, M. Fang, Z. Pan, Q. Fang, Z. Wen, H. Pan, D. Li, Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling (2025). doi:10.1609/aaai.v39i22.34559, **L50**.

[35] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track, 2023.

[36] J. Zhang, Z. Li, K. Das, B. Malin, S. Sricharan, Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency (2023). doi:10.18653/v1/2023.findings-emnlp.1032, **L12**.

[37] P. Manakul, A. Liusie, M. Gales, Shroomshroom: Zero-resource black-box hallucination detection for generative large language models (2023). **L13**.

[38] R. Mehta, A. Hoblitzell, J. O'Keefe, H. Jang, V. Varma, Halu-nlp at semeval-2024 task 6: Metacheckgpt - a multi-task hallucination detection using llm uncertainty and meta-models

(2024). doi:10.18653/v1/2024.semeval-1.52, **L26**.
[39] W. Wu, Y. Cao, N. Yi, R. Ou, Z. Zheng, Detecting and reducing the factual hallucinations of large language models with metamorphic testing, Proc. ACM Softw. Eng. (2025). doi:10.1145/3715784, **L51**.
[40] B. Yang, M. A. Al Mamun, J. M. Zhang, G. Uddin, Hallucination detection in large language models with metamorphic relations, Proc. ACM Softw. Eng. (2025). doi:10.1145/3715735, **L56**.
[41] R. Sanayei, A. Singh, M. Rezaei, S. Bethard, Maria at semeval 2024 task-6: Hallucination detection through llms, mnli, and cosine similarity (2024). doi:10.18653/v1/2024.semeval-1.225, **L36**.
[42] A. Bui, S. Brech, N. Hußfeldt, T. Jennert, M. Ullrich, T. Breuer, N. Khasmakhi, P. Schaer, The two sides of the coin: Hallucination generation and detection with llms as evaluators for llms (2024). **L46**.
[43] S. Das, R. Śrihari, Compos mentis at semeval2024 task6: A multi-faceted role-based large language model ensemble to detect hallucination (2024). doi:10.18653/v1/2024.semeval-1.208, **L10**.
[44] B. Allen, F. Polat, P. Groth, Shroom-indelab at semeval-2024 task 6: Zero- and few-shot llm-based classification for hallucination detection (2024). doi:10.18653/v1/2024.semeval-1.120, **L44**.
[45] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, Y. Xiao, Hallucination detection: Robustly discerning reliable answers in large language models (2023). doi:10.1145/3583780.3614905, **L7**.
[46] C. Wei, Z. Chen, S. Fang, J. He, M. Gao, Opdai at semeval-2024 task 6: Small llms can accelerate hallucination detection with weakly supervised data (2024). doi:10.18653/v1/2024.semeval-1.104, **L40**.
[47] Y. Yehuda, I. Malkiel, O. Barkan, J. Weill, R. Ronen, N. Koenigstein, Interrogatellm: Zero-resource hallucination detection in llm-generated answers (2024). doi:10.18653/v1/2024.acl-long.506, **L30**.
[48] S. Yang, R. Sun, X. Wan, A new benchmark and reverse validation method for passage-level hallucination detection (2023). doi:10.18653/v1/2023.findings-emnlp.256, **L2**.
[49] T. Zhang, L. Qiu, Q. Guo, C. Deng, Y. Zhang, Z. Zhang, C. Zhou, X. Wang, L. Fu, Enhancing uncertainty-based hallucination detection with stronger focus (2023). doi:10.18653/v1/2023.emnlp-main.58, **L5**.
[50] Z. Wen, Z. Liu, Z. Tian, S. Pan, Z. Huang, D. Li, M. Huang, Scenario-independent uncertainty estimation for llm-based question answering via factor analysis (2025). doi:10.1145/3696410.3714880, **L64**.
[51] P. Bhamidipati, A. Malladi, M. Shrivastava, R. Mamidi, Maha bhaashya at semeval-2024 task 6: Zero-shot multi-task hallucination detection (2024). doi:10.18653/v1/2024.semeval-1.241, **L34**.
[52] F. Cheng, V. Zouhar, S. Arora, M. Sachan, H. Strobelt, M. El-Assady, Relic: Investigating large language model responses using self-consistency (2024). doi:10.1145/3613904.3641904, **L42**.
[53] F. Borra, C. Savelli, G. Rosso, A. Koudounas, F. Giobergia, Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection (2024). doi:10.18653/v1/2024.semeval-1.240, **L35**.
[54] J. Lu, S. Li, Roberta with low-rank adaptation and hierarchical attention for hallucination detection in llms, 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (2024). doi:10.1109/ICICML63543.2024.10957858, **L43**.
[55] D. Zhang, V. Gangal, B. Lattimer, Y. Yang, Enhancing hallucination detection through perturbation-based synthetic data generation in system responses (2024). doi:10.18653/v1/2024.findings-acl.789, **L17**.