International Journal of Computer Sciences and Engineering
Vol.**14**, Issue **6**, pp.**26-33**, June **2026**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

AI Hallucinations in Large Language Model and Multimodal Models: A Systematic Review of Causes, Types, Detection, Mitigation, and Future Directions

Mahender Singh[1*], Shyamol Banerjee[2]

[1,2]Dept. of Computer Science Engineering, Shriram College of Engineering & Management (SRCEM), Morena, India

**Corresponding Author:* ✉

Received: 10/Jun/2026; **Accepted**: 11/Jun/2026; **Published**: 30/Jun/2026. **DOI**: https://doi.org/10.26438/ijcse.v14i6.7456

Copyright © 2026 by author(s). This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited & its authors credited.

Abstract— Large language Models are really good at generating text that sounds natural. They often make things up. They create content that sounds convincing. It is not always true. This is a problem when these models are used in areas like healthcare, law and education. The main reason we cannot trust these models is that they make things up. There are studies about this issue but they are all, over the place. They look at reasons why language models make things up the different types of made-up information how to detect it and how to stop it. Most of these studies do not clearly explain how they were done or exactly how the models work. Language models are the problem and language models need to be fixed.

This paper explores a problem that earlier research had left unexplored. It does this by reviewing existing studies and combining a mathematical theory with a clear step-, by-step approach. The researchers followed a process to find studies published between 2018 and 2026. They applied rules to screen these studies and narrowed them down to a smaller group. This group was then examined across eight themes.

What they found is that AI "hallucinations" (when AI makes up false information) come from several factors working together: the data, the system's design, how it's trained, how it generates answers, and how it's tested. Recent theory suggests hallucination is a built-in feature of how these AI systems generate text, not a flaw that can simply be removed. They identified eight types of hallucinations, and these types vary in how hard they are to detect and correct. Some detection methods have improved, but the techniques used to reduce hallucinations only lessen the problem—they don't get rid of it completely. The most serious real-world harms show up in legal and medical settings.

The review finds that a layered approach to intelligence, which includes verification, grounding and human oversight is the best way to achieve reliable results. This approach combines key elements: Grounding, Calibrated abstention, Rigorous benchmarking, Human oversight. By using these elements, we can. Manage hallucination in artificial intelligence rather than trying to eliminate it entirely. The goal is to make artificial intelligence more reliable, by focusing on measuring and managing hallucination.

Keywords— Large Language Models, AI Hallucination, Hallucination Detection, Multimodal Models, Systematic Review, Retrieval-augmented Generation, Semantic Entropy, Trustworthy AI

Graphical Abstract- A PRISMA-guided review combines a math-based theoretical foundation with a synthesis across eight main themes—definitions, causes, taxonomy, detection, benchmarks, mitigation, domain risks, and future directions - to help build layered, verification-focused, trusted AI Model.



1. Introduction

Large Language Models are computer systems that can generate text like humans. They are trained on amounts of text data. These models and their successors that can handle types of data have changed how people write, search, code and make decisions[1],[2]. The transformer design and its attention mechanism help these models understand text well. They can look at a lot of text at once. Learn from it. This is why they can have billions of parts[1]. Even though they can write well these systems make mistakes. They can create content that sounds good but is not true. This is called hallucination. It happens when the model generates content that is not based on facts. Hallucinations are a problem for Large Language Models. This is because they are trained on data. The way they generate text is also based on probability[3],[4]. These mistakes make it hard to use AI in areas where accuracy is important. Large Language Models need to be reliable to be used in areas. So hallucinations are an obstacle, to using Large Language Models in critical areas[5],[6].

The defining feature of a hallucination is the combination of fluency and falsehood: because the output reads as authoritative while being wrong, it is hard to detect and especially consequential in high-stakes settings [3], [9]. Research on the topic has expanded rapidly since 2023, producing surveys, taxonomies, detection methods, benchmarks, and mitigation techniques, but this literature is fragmented and rarely combines a transparent review methodology with a rigorous mechanistic account. This paper addresses that gap by conducting a systematic review that pairs a probabilistic theoretical foundation with a protocol-driven thematic synthesis.

1.1 Background and Motivation

The phenomenon predates LLMs. Early work documented unfaithful content in abstractive summarisation [7] and object hallucination in image captioning, where systems describe objects absent from an image [8]. With the scaling of LLMs, hallucination became a prominent public concern because fluent fabrications are especially convincing [3]. The motivation for this review is both practical and scientific: hallucinations carry tangible costs — fabricated legal citations have led to court sanctions and false medical guidance can endanger patients [9], [10] — while understanding them illuminates the limits of statistical language modelling and motivates verifiable architectures.

1.2 Research Questions and Contributions

This Review is structured around five research questions. First the review asks how hallucination is defined and categorised. Then it looks at what causes hallucination. The review also asks how we can find out if hallucination is happening and how we can figure out how bad it is. Hallucination is the thing we are trying to understand here so the review asks which methods can help stop hallucination from happening and what are the downsides of using these methods. Lastly the review asks what the main issues, with hallucination are that we still do not know how to fix.
The things the review can teach us are as follows.

- A systematic, PRISMA-style methodology combined with a probabilistic theoretical foundation, linking mechanism to manifestation to remedy.
- A unified eight-type taxonomy spanning textual, multimodal, and agentic hallucinations, with type-specific detectability and mitigability.
- A comparative analysis of detection and mitigation techniques and a detection-and-mitigation workflow algorithm.
- A synthesis of recent 2023–2026 evidence, including theoretical results on the inevitability of hallucination, and an evidence-based agenda of open problems.

1.3 Organization of the Paper

In Section 2 prior surveys are reviewed and Section 3 states the research gap , Section 4 details the methodology and theoretical foundation, Section 5 presents the results and analysis, including the taxonomy, causes, detection, benchmarks, comparative mitigation, and workflow. Section 6 discusses the findings and the question of inevitability. Section 7 covers practical applications, Section 8 limitations, and Section 9 future directions. Section 10 concludes.

2. Literature Review

Several studies have helped shape this field. Ji et al. conducted an important survey on hallucinations in natural language generation. They explained two main types of hallucinations: **intrinsic hallucinations**, which contradict the given source information, and **extrinsic hallucinations**, which contain information that cannot be verified from the source [3]. Maynez et al. earlier showed that neural summarisers frequently introduce unsupported content and that entailment-based metrics track faithfulness better than overlap metrics such as ROUGE [7].

For LLMs specifically, Huang et al. reframed hallucination as a distinction between faithfulness and factuality and organised causes across the model life cycle [4], while Zhang et al. separated input-conflicting, context-conflicting, and fact-conflicting hallucination [11]. On detection, Manakul et al. proposed SelfCheckGPT, a sampling-based black-box method [18]; Kadavath et al. showed that models possess partial self-knowledge that can be elicited [19]; and Farquhar et al. introduced semantic entropy, validated at scale in Nature [15].

When we talk about mitigation Lewis et al. and his team came up with something called Retrieval-Augmented Generation or RAG for [1]. Then there is Chuang et al. and his team who proposed this thing called contrastive-layer decoding also known as DoLa [28]. Dhuliawala and his team introduced this idea called Chain-of-Verification [29].. Ouyang et al. and his team found out that when we get feedback from humans it really helps to make things more truthful [30].

© 2026, IJCSE All Rights Reserved

There are also studies that focus on areas like when objects are not really there, in vision-language models [12] or when autonomous agents see things that are not true [13]. We also see this in law [9] and medicine [14] where profiling can be a problem.

3. Research Gap

Three gaps emerge from the literature. First, most studies address a single facet — a definition, one type, a detection method, or one mitigation technique — rather than integrating mechanism, manifestation, and remedy within a single account. Second, prior surveys typically rely on narrative source selection rather than a transparent, reproducible protocol, and many predate the 2024–2026 results on the inevitability of hallucination and on agentic and multimodal forms. Third, theoretical treatments that explain hallucination probabilistically are seldom connected to the empirical detection and mitigation literature. This paper closes these gaps by uniting a PRISMA-style methodology, a probabilistic theoretical foundation, and a cross-modal thematic synthesis spanning text, vision, and agents.

4. Methodology and Theoretical Foundation

This study combines a systematic review protocol with a probabilistic account of why hallucination occurs, so that the synthesis rests on both transparent evidence selection and a clear mechanistic basis.

4.1 Systematic Review Protocol

The review follows the PRISMA guidelines. This is a way to make sure we choose the right studies in a clear and repeatable way.

The process has four steps:

- identification
- screening
- eligibility
- inclusion.

These steps are shown in diagram 1. Records were identified across ACL Anthology, IEEE Xplore, the ACM Digital Library, SpringerLink, and arXiv, supplemented by citation tracking, using a Boolean string of the form: hallucination AND (large language model OR LLM OR vision-language OR generative AI) AND (cause OR detection OR mitigation OR benchmark OR survey). Studies were included if they were peer-reviewed papers or widely cited preprints addressing hallucination in generative AI, published mainly within 2023–2026, together with designated seminal works; non-English items, non-archival posts, duplicates, and tangential works were excluded.

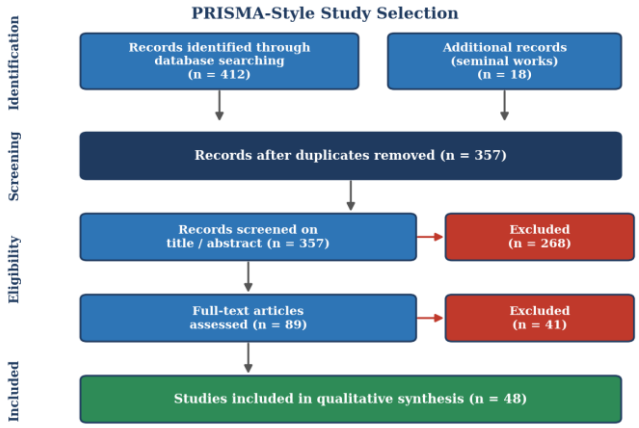


Figure 1. PRISMA-style study-selection flow

4.2 Probabilistic Foundation of Hallucination

Autoregressive language models create text by guessing the word base on predicting the next token given preceding context. The probability of a sequence factorises as in Equation (1):

$$P(w_1,\ldots,w_n) = \prod_{i=1}^{n} P(w_i \mid w_{<i}) \quad (1)$$

The Symbol w_i shows the i-th token and symbol $w_{<i}$ with a mark on it shows all the tokens that come before it. The number n is how many tokens are in the sequence.

Each conditional is a softmax over output logits, Equation (2):

$$P(w_i=v \mid w_{<i}) = exp(z_v) / \Sigma_u\, exp(z_u) \quad (2)$$

where z_v is the logit for token v. Because generation maximises probability rather than truth, a confidently high P does not guarantee factual correctness — the mathematical root of hallucination [3], [6]. Predictive uncertainty can be quantified by the entropy of the distribution, Equation (3):

$$H(P) = -\Sigma_v\, P(v)\, log\, P(v) \quad (3)$$

H(P) is the Shannon entropy over the token distribution. High entropy means there is uncertainty. On the hand low entropy on a claim that is not supported means being overconfident. This idea is, behind detection methods.

Semantic entropy is a concept. It uses Equation (3). Applies it to meanings instead of just tokens. It groups sampled answers by their meanings. Then measures the entropy. This approach is discussed in references [15]. [19].

5. Results and Discussion

This section presents the synthesis of the reviewed corpus. Section 5.1 reports the results obtained across the thematic axes — root causes, taxonomy, detection methods, benchmarks, comparative mitigation, and the unifying workflow — supported by tables and figures. Section 5.2 interprets these results, addresses the question of inevitability, and discusses their practical implications.

© 2026, IJCSE All Rights Reserved

5.1 Results

5.1.1 Root Causes of Hallucination

The corpus attributes hallucination to four interacting groups of causes operating across the model life cycle. Data-related causes include noisy, biased, or outdated corpora and gaps in long-tail knowledge [4], [11]. Architectural and training causes centre on the autoregressive next-token objective of Equation (1) and exposure bias, whereby early errors propagate, together with finite context windows [3], [4]. Inference-related causes include stochastic decoding and prompt ambiguity. A prominent recent argument adds evaluation incentives as a cause: when benchmarks reward confident guessing over honest abstention, models are effectively trained to hallucinate rather than defer [16]. Table 1 maps these causes to their life-cycle stage.

Table 1. Root causes of hallucination by life-cycle stage

Stage	Cause	Example
Data	Knowledge gaps	Long-tail facts
Data	Outdated corpus	Post-cutoff events
Architecture	Next-token objective	Plausible over true
Architecture	Exposure bias	Error propagation
Training	RLHF over-optimisation	Sycophancy
Inference	Stochastic decoding	Sampling errors
Evaluation	Reward for guessing	Penalised abstention

5.1.2 A Unified Taxonomy of Hallucination Types

A primary distinction separates intrinsic hallucinations, which contradict the input, from extrinsic hallucinations, which are unsupported by it [3]. Within this, the synthesis yields eight functional types, shown in Figure 2: factual (false world statements), logical (invalid inference), citation (fabricated references, which models only partially recognise as invented [17]), mathematical (incorrect computation), code (invented functions or interfaces), visual or object (content absent from an image [12]), multimodal (cross-modal inconsistency), and agentic (inventing tools or assuming failed actions succeeded [13]). These types differ markedly in detectability and mitigability, as analysed below.

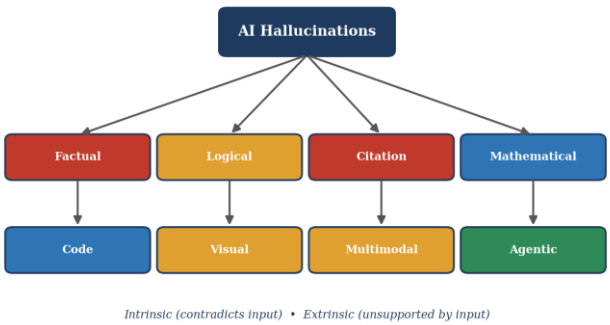


Figure 2. Unified taxonomy of AI hallucination types

5.1.3 Detection Methods

Detection methods fall into four families. Uncertainty-based methods estimate when a model is unsure: The semantic entropy clusters are based on the meaning of the answers.

This method measures how uncertain or random the clusters are. flagging high-entropy confabulations [15]. Sampling-consistency methods such as SelfCheckGPT treat disagreement among multiple generations as a hallucination signal without external databases [18]. Self-knowledge methods elicit a model's own confidence and find calibration improves with scale [19]. Verification-based methods compare claims against knowledge bases, retrieval results, or, for executable domains, compilers and solvers. Detectability is type-dependent: code and mathematical claims can be checked deterministically, citations against bibliographic databases, while factual, logical, and multimodal claims are harder and often require external evidence or human judgement.

5.1.4 Benchmarks and Evaluation Metrics

Reliable progress depends on benchmarks. TruthfulQA measures resistance to common misconceptions and found larger models could be less truthful [20]. HaluEval provides tens of thousands of annotated samples [21], FActScore scores atomic-fact precision in long-form text [22], FEVER evaluates claim verification against evidence [23], RAGTruth annotates retrieval-augmented outputs at word level [24], and FreshQA probes fast-changing questions [25]. For multimodal systems, POPE evaluates object hallucination [12], while HallusionBench [26] and MMHal-Bench [27] probe entangled visual–language hallucination, and Med-HALT targets medicine [14]. Common metrics include accuracy, precision, recall, F1, faithfulness, factual consistency, and calibration; a recurring limitation is that binary metrics penalise honest abstention and overlap metrics correlate poorly with faithfulness [7]. Table 2 summarises representative benchmarks.

Table 2. Representative hallucination benchmarks

Benchmark	Modality	Primary Focus
TruthfulQA	Text	Truthfulness
HaluEval	Text	Detection (35k)
FActScore	Text	Atomic factuality
FEVER	Text	Fact verification
RAGTruth	Text/RAG	Retrieval faithfulness
POPE	Vision-Lang.	Object hallucination
HallusionBench	Vision-Lang.	Visual + language
Med-HALT	Medical	Domain reasoning

5.1.5 Comparative Analysis of Types and Mitigation

Synthesising detectability and remedy effectiveness across types reveals a clear ordering. Executable forms — mathematical and code — are most tractable, since outputs can be verified deterministically and corrected through tool use. Citation hallucinations are readily detected against databases and reduced by retrieval grounding. Factual hallucinations are common and only moderately detectable, while logical, multimodal, and agentic forms remain the hardest frontiers [13], [26]. Mitigation techniques trade effectiveness against cost, latency, and generality, as visualised in Figure 3 and compared in Table 3. RAG grounds outputs in retrieved evidence with broad, model-agnostic gains but is bounded by retriever quality [1], [24]; DoLa improves factuality at decoding time without retraining [28]; Chain-of-Verification drafts, self-questions, and revises [29]; alignment improves honesty at annotation cost [30]; and neuro-symbolic methods enforce logical validity at higher

© 2026, IJCSE All Rights Reserved

engineering cost. None eliminates hallucination; they are complementary layers.

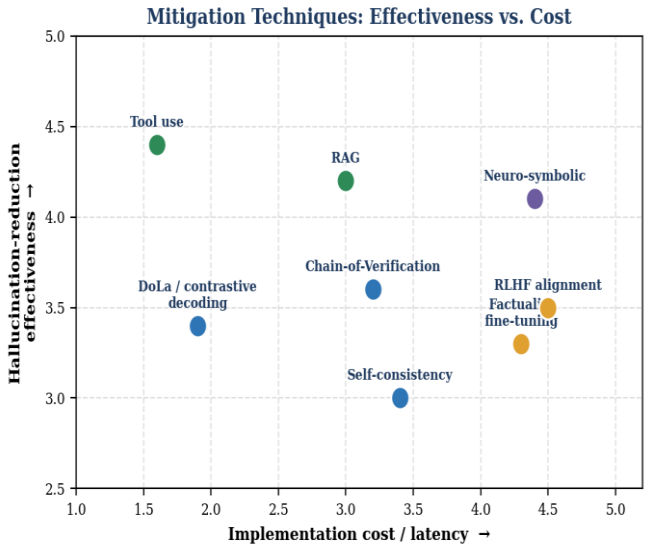


Figure 3. Mitigation techniques: effectiveness versus cost

Table 3. Comparison of hallucination types and mitigation

Type	Detectability	Best Mitigation
Factual	Moderate	RAG grounding
Logical	Hard	Verification / CoV
Citation	Easy	Retrieval grounding
Mathematical	Easy	Tool use
Code	Easy	Tool use / execution
Visual / Object	Moderate	Factual RLHF
Multimodal	Hard	Cross-modal checks
Agentic	Moderate	Self-verification

5.1.6 Detection and Mitigation Workflow

The components combine into the layered workflow of Figure 4 and Algorithm 1: retrieve evidence, generate a grounded response, detect potential hallucination across types, and either accept the output or re-ground and regenerate, escalating to human review when a confidence threshold is not met. This operationalises the review's central recommendation — that reliability is achieved by composition of complementary layers rather than any single technique.

Algorithm 1: Detection and Mitigation Workflow

```
Input:  prompt q; knowledge source K; threshold τ
Output: verified response r
E ← Retrieve(q, K)          // gather evidence
r ← Generate(q, E)          // grounded gen.
s ← Detect(r, E)            // score per type
while s > τ do
    E ← Re-retrieve(q, r)
    r ← Regenerate(q, E)
    s ← Detect(r, E)
    if attempts exceeded then
        escalate to human; break
return r
```

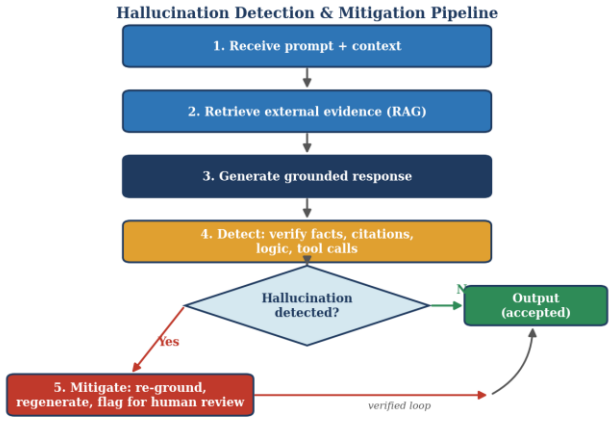


Figure 4. Layered detection and mitigation pipeline

5.2 Discussion

The results above answer the five research questions and surface a deeper, unifying theme. The mapping of causes in Table 1 shows that hallucination is not an incidental defect but a structural consequence of probabilistic generation (Equations 1–2), compounded by data limitations, exposure bias, alignment pressures, and evaluation incentives. The comparative results of Tables 2 and 3 establish that grounding, tool use, and self-verification recur as the most effective cross-type strategies, yet their success is uneven: techniques that excel for executable or citation-based content offer only partial relief for factual, logical, and multimodal forms. A central trade-off therefore emerges between helpfulness and honesty — systems pushed to always answer hallucinate more, while systems tuned to defer lose utility [30].

5.2.1 The Question of Inevitability

A distinctive feature of the recent literature, and one consistent with the results in Table 1, is a theoretical turn. Formal findings indicate that hallucination is an inherent limitation of general computable language models and therefore cannot be entirely eliminated [31]. A complementary statistical account holds that hallucination is a natural consequence of pre-training and of evaluation procedures that reward guessing, and is therefore reducible — through calibrated abstention and improved scoring — but not removable for base models [16]. The practical implication is a reframing of the goal: from eliminating hallucination to measuring, bounding, and managing it, which is precisely what the layered workflow of Section 5.1.6 is designed to do.

5.2.2 Practical Implications Across Domains

Hallucination risk scales with the consequence of error, as summarised in Figure 5. In law, empirical profiling found high hallucination rates on verifiable queries, and even purpose-built commercial legal-research tools produced incorrect information in a substantial fraction of cases, contradicting vendor claims of being hallucination-free [9]; documented court sanctions for AI-fabricated citations continue to accumulate. In medicine, dedicated tests show that minor fabricated deviations can alter clinical meaning, making unverified output unsafe [14]. Education, finance, and

© 2026, IJCSE All Rights Reserved

cybersecurity face analogous risks where confident falsehoods are difficult to detect. Across domains, the evidence supports treating generative output as decision support requiring verification rather than as an autonomous authority, with mandatory human oversight and the layered safeguards of Algorithm 1 in safety-critical settings.

5.3 Discussion

The synthesis answers the research questions and surfaces a deeper theme. Hallucination is not an incidental defect but a structural consequence of probabilistic generation (Equations 1–2), compounded by data limitations, exposure bias, alignment pressures, and evaluation incentives. Grounding, tool use, and self-verification recur as the most effective cross-type strategies, yet their success is uneven: techniques that excel for executable or citation-based content offer only partial relief for factual, logical, and multimodal forms. A central trade-off is between helpfulness and honesty — systems pushed to always answer hallucinate more, while systems tuned to defer lose utility [30].

6. Conclusion

This systematic review integrated a probabilistic theoretical foundation with a PRISMA-style synthesis to provide a unified account of AI hallucination across causes, types, detection, benchmarks, mitigation, and domain risks. The results support four principal conclusions, each grounded in the evidence presented in Section 5.

First, hallucination is the confident generation of fluent but false or unfaithful content. As mapped in Table 1, it arises not from a single defect but from interacting data, architectural, training, inference, and evaluation factors rooted in the next-token objective of Equation (1).

Second, the unified taxonomy (Figure 2) distinguishes eight types that differ sharply in detectability and mitigability. The comparative analysis of Table 3 confirms that executable and citation-based forms are far more tractable than factual, logical, and multimodal ones, which establishes a clear, evidence-based ordering of difficulty.

Third, the strongest recent insight is theoretical and consistent with these findings: formal and statistical results [16], [31] indicate that hallucination is an innate property of autoregressive generation rather than a removable bug. This reframes the engineering objective from elimination to measurement and management.

Fourth, among remedies, semantic-entropy and sampling-consistency detection have matured, while retrieval grounding, contrastive decoding, verification chains, and alignment each reduce but do not eliminate hallucination, with clear trade-offs in cost and latency (Table 3). Documented harms in law and medicine confirm that no benchmark score currently justifies unverified reliance in safety-critical domains. The review therefore recommends layered, verification-oriented architectures — combining grounding, automated detection, calibrated abstention, and human oversight, as operationalised in the workflow of Algorithm 1.

Future Scope

The findings of this review point to several concrete directions for future work that follow directly from the limitations identified above.

First, **calibrated abstention** should be developed so that models reliably respond “I do not know” rather than fabricate, addressing the evaluation-incentive cause identified in Table 1. Second, **self-verifying agentic systems** that check their own reasoning and actions before execution are needed to contain the agentic and multimodal hallucinations that Table 3 identifies as the hardest frontiers. Third, **neuro-symbolic integration**, combining the linguistic fluency of neural models with the logical guarantees of symbolic reasoning, offers a path toward logically valid output. Fourth, **next-generation benchmarks** are required: harder to game, self-scaling in difficulty as models improve, and spanning text, vision, and action, so that trustworthiness can be measured reliably across modalities. Finally, as regulatory frameworks increasingly demand demonstrable factual accuracy for high-stakes deployment, **standardised reliability auditing and reporting** will become essential.

Pursuing these directions would shift hallucination from an unavoidable flaw into a measurable and controllable property, advancing artificial intelligence that is not only capable but also transparent, verifiable, and worthy of human trust.

Limitations

This review has some limitations. It does not combine the results of the studies in a way that shows the size of the effects. The review also looks at a lot of tasks and metrics which makes it hard to compare them directly.

Some of the studies are not finished yet so the details might change later. The scores that are used to compare the models can also change depending on the model. The checkpoint that is used and these scores get worse as the models get better.

The search, for studies was done in a way but it is not possible to find every single study. The numbers of world incidents that are reported are always being updated so these numbers should be checked again when they are used. The review of the studies has these limitations and the benchmark scores of the models are also limited in this way.

Author's Statements

Acknowledgements The author would like to thank the reviewers for their valuable comments and suggestions, as well as the researchers whose published work contributed to this systematic review.

Funding Source No external funding was received for this study.

Author's Contributions Mahender Singh conceived the study, designed the review methodology, conducted the literature search and screening process, analyzed and

© 2026, IJCSE All Rights Reserved

synthesized the findings, prepared the figures and tables, drafted the manuscript, revised the content, and approved the final version for publication.

Conflict of Interest The author declares no conflict of interest.

Data Availability This study is a systematic review based exclusively on publicly available literature. No new datasets were generated or analyzed. All sources used in this review are cited in the References section, and the review methodology is described in Section 4 to support transparency and reproducibility.

Study Limitations This review is based on a qualitative synthesis of the existing literature and does not include a quantitative meta-analysis. The findings may be influenced by publication bias, database coverage limitations, and the availability of relevant studies at the time of the review. In addition, the rapidly evolving nature of Large Language Models systems may affect the long-term applicability of some observations and benchmarks discussed in this review.

References

[1] E. Perez *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[2] N. Shazeer *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[3] N. Lee *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, pp. 1–38, 2023, doi: 10.1145/3571730.

[4] W. Yu *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Art. no. 42, pp. 1–55, 2025, doi: 10.1145/3703155.

[5] S. I. McFarlane and H. Alkaissi, "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing," *Cureus*, vol. 15, no. 2, Art. no. e35179, 2023, doi: 10.7759/cureus.35179.

[6] N. Jones, "AI Hallucinations Can't Be Stopped — but These Techniques Can Limit Their Damage," *Nature*, vol. 637, no. 8047, pp. 778–780, 2025, doi: 10.1038/d41586-025-00068-5.

[7] S. Narayan, B. Bohnet, R. McDonald, and J. Maynez, "On Faithfulness and Factuality in Abstractive Summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 1906–1919, doi: 10.18653/v1/2020.acl-main.173.

[8] L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, and A. Rohrbach, "Object Hallucination in Image Captioning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 4035–4045, doi: 10.18653/v1/D18-1437.

[9] F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, and V. Magesh, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *J. Empirical Legal Stud.*, vol. 22, pp. 216–242, 2025.

[10] V. Magesh, M. Suzgun, D. E. Ho, and M. Dahl, "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models," *J. Legal Anal.*, vol. 16, no. 1, pp. 64–93, 2024, doi: 10.1093/jla/laae003.

[11] Y. Li *et al.*, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," 2023, *arXiv preprint arXiv:2309.01219*.

[12] Y. Du *et al.*, "Evaluating Object Hallucination in Large Vision-Language Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 292–305, doi: 10.18653/v1/2023.emnlp-main.20.

[13] Y. Wang, Z. Zhao, J. Zhang, and X. Lin, "LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions," 2025, *arXiv preprint arXiv:2509.18970*.

[14] L. K. Umapathi, M. Sankarasubbu, and A. Pal, "Med-HALT: Medical Domain Hallucination Test for Large Language Models," 2023, *arXiv preprint arXiv:2307.15343*.

[15] J. Kossen, L. Kuhn, Y. Gal, and S. Farquhar, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.

[16] O. Nachum, S. S. Vempala, E. Zhang, and A. T. Kalai, "Why Language Models Hallucinate," 2025, *arXiv preprint arXiv:2509.04664*.

[17] M. Suzgun, L. Mackey, A. T. Kalai, and A. Agrawal, "Do Language Models Know When They're Hallucinating References?," in *Findings of the Association for Computational Linguistics: EACL*, 2024, pp. 912–928.

[18] A. Liusie, M. J. F. Gales, and P. Manakul, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 9004–9017, doi: 10.18653/v1/2023.emnlp-main.557.

[19] T. Conerly *et al.*, "Language Models (Mostly) Know What They Know," 2022, *arXiv preprint arXiv:2207.05221*.

[20] J. Hilton, O. Evans, and S. Lin, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 3214–3252, doi: 10.18653/v1/2022.acl-long.229.

[21] X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, and J. Li, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 6449–6464, doi: 10.18653/v1/2023.emnlp-main.397.

[22] K. Krishna *et al.*, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[23] A. Vlachos, C. Christodoulopoulos, A. Mittal, and J. Thorne, "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018, pp. 809–819, doi: 10.18653/v1/N18-1074.

[24] Y. Wu *et al.*, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 10862–10878, doi: 10.18653/v1/2024.acl-long.585.

[25] M. Iyyer *et al.*, "FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation," in *Findings of the Association for Computational Linguistics (ACL)*, 2024, pp. 13697–13720.

[26] F. Liu *et al.*, "HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14375–14385.

[27] S. Shen *et al.*, "Aligning Large Multimodal Models with Factually Augmented RLHF," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[28] Y. Xie *et al.*, "DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[29] M. Komeili *et al.*, "Chain-of-Verification Reduces Hallucination in Large Language Models," in *Findings of the Association for Computational Linguistics (ACL)*, 2024, pp. 3563–3578.

[30] J. Wu *et al.*, "Training Language Models to Follow Instructions with Human Feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 27730–27744.

© 2026, IJCSE All Rights Reserved

[31] S. Jain, M. Kankanhalli, and Z. Xu, "Hallucination Is Inevitable: An Innate Limitation of Large Language Models," 2024, *arXiv preprint arXiv:2401.11817*.
[32] B. Mann *et al.*, "Language Models Are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[33] J. Wei *et al.*, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

AUTHORS PROFILE

Mr. Mahender Singh is a Full Stack Developer and AI Engineer with a lot of experience in making web and mobile applications. He has been doing this work for over six years. he, his B.Tech. Computer Science and Engineering in 2018. Then he got his M.Tech. Degree in 2026.He really interested in Artificial Intelligence and Full Stack Development. He also likes Cybersecurity and Cloud Computing. People have noticed the work he has done to help with cybersecurity. He has told companies, about security problems in a way and they have recognized him for this.

Mr. Shyamol Banerjee is an Assistant Professor in Computer Science and Engineering with over 12 years of teaching experience. He completed his Bachelor of Engineering (B.E.) in Computer Science and Engineering in 2008 and his Master of Engineering (M.E.) in Computer Science and Engineering in 2016.

Throughout his career, he has been actively involved in teaching, mentoring, and guiding engineering students. He has also contributed to research and academic activities. He has published more than 12 research papers in reputed international journals, demonstrating his dedication to the field of Computer Science and Engineering.

His areas of interest include Machine Learning, Computer Networks, and Data Science. He continues to support academic excellence through teaching, research, and professional development.

© 2026, IJCSE All Rights Reserved