Fulawade Harsh Nitin, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Hallucination In Large Language Models: A Review Of Causes, Taxonomy, Detection Methods And Mitigation Strategies

Fulawade Harsh Nitin[1] Prof. N. G. Rokade[2] Prof. M. C. Kolse[3] Prof. S. K. Said[4]
Department of Artificial Intelligence and Data Science
Jaihind College of Engineering
Pune, Maharashtra, India

Abstract—Large Language Models (LLMs) generate human-like text, answer questions and summarize information, but they sometimes produce incorrect, misleading or fabricated content while presenting it as fact. This phenomenon, known as hallucination, reduces the accuracy, trustworthiness and safety of generative AI systems, particularly in education, healthcare, law and finance. This paper reviews LLM hallucination in a structured way. It explains causes arising from training data, model architecture and training objectives, decoding and missing context, and presents a taxonomy that distinguishes factuality from faithfulness errors. Benchmarks that measure truthfulness and hallucination rates are discussed. Detection methods are reviewed, including self-consistency checks across sampled answers, uncertainty measures such as semantic entropy and fine-grained atomic fact verification. Mitigation strategies are examined across the model life cycle: data curation, alignment with human feedback, retrieval-augmented generation, prompting for verification and abstention when uncertain. A layered hallucination-aware LLM application architecture is proposed, along with research gaps, objectives and a comparison of methods. The review concludes that hallucination cannot be eliminated entirely but can be detected, reduced and communicated so that users can rely on LLM outputs appropriately.

Index Terms—Large language models, hallucination, factuality, faithfulness, uncertainty estimation, semantic entropy, fact verification, generative AI

I. INTRODUCTION

LLMs such as those behind modern chat assistants can write essays, explain concepts and draft code. Yet a student asking for references may receive citations to papers that do not exist, and a user asking about a medicine may receive a confident but wrong dosage. Such errors are called hallucinations.

Hallucinations are particularly dangerous because they are fluent and plausible. Ji *et al.* [1] surveyed hallucination in natural language generation, and Huang *et al.* [2] and Zhang *et al.* [3] extended the analysis to LLMs, documenting causes, benchmarks and mitigation methods. Kalai and Vempala [4] argued theoretically that calibrated language models must hallucinate at some rate on rare facts, implying that the goal is management rather than complete elimination.

This paper reviews why LLMs hallucinate, how hallucinations can be detected and how they can be reduced, and proposes an application architecture.

II. THEORETICAL BACKGROUND

1. Taxonomy

Fig. 1 summarizes the taxonomy. Factuality hallucinations contradict real-world facts or invent them; faithfulness hallucinations deviate from the user's instructions or the supplied context. Maynez *et al.* [5] showed that abstractive summarizers frequently add unsupported content, an early demonstration of faithfulness errors.

© 2026Fulawade Harsh Nitin, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

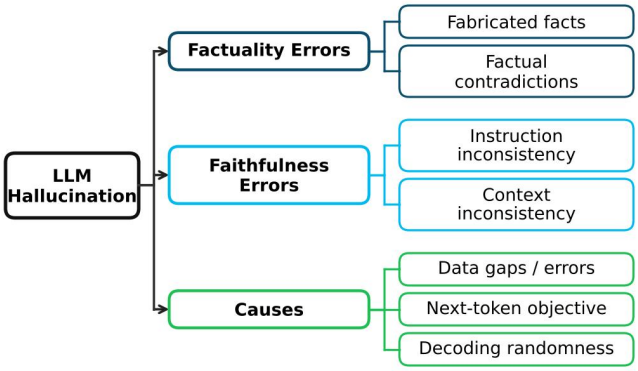


Fig. 1. Taxonomy of hallucination and its main causes.

2. Causes

LLMs are trained to predict the next token, which rewards plausible text rather than verified truth. Training data contains errors and gaps, rare facts are poorly memorized, decoding with sampling introduces randomness, and models may answer even when the question lies outside their knowledge.

III. SURVEY OF EXISTING APPROACHES

Lin *et al.* [6] built TruthfulQA, questions designed to elicit common misconceptions, and found that larger models were not necessarily more truthful because they imitated human falsehoods. Li *et al.* [7] introduced HaluEval, a large benchmark for recognizing hallucinated content in question answering, dialogue and summarization.

For detection, Manakul *et al.* [8] proposed SelfCheckGPT, which samples several answers and flags statements that are inconsistent across samples, requiring no external database. Farquhar *et al.* [9] introduced semantic entropy, measuring uncertainty over meanings rather than exact wordings, and showed it detects confabulations across tasks. Min *et al.* [10] proposed FActScore, which splits long text into atomic facts and checks each against a knowledge source. Kadavath *et al.* [11] showed that models can be trained to estimate whether they know the answer, supporting abstention.

For mitigation, Ouyang *et al.* [12] used reinforcement learning from human feedback to make models more truthful and helpful, Lewis *et al.* [13] grounded generation in retrieved documents, and Dhuliawala *et al.* [14] proposed chain-of-verification, in which the model drafts an answer, plans verification questions, answers them independently and revises. Table I summarizes these works.

Ref.	Authors	Year	Technique Used	Merits	Demerits
[5]	Maynez et al.	2020	Analysis of summarization faithfulness	Found frequent unsupported content	Summarization only
[6]	Lin et al.	2022	TruthfulQA benchmark	Larger models not always truthful	English, short answers
[12]	Ouyang et al.	2022	RLHF alignment	More truthful, helpful outputs	Does not remove hallucination
[7]	Li et al.	2023	HaluEval benchmark	Large hallucination recognition set	Generated test data
[8]	Manakul et al.	2023	SelfCheckGPT sampling consistency	Zero-resource detection	Multiple samples needed
[10]	Min et al.	2023	FActScore atomic facts	Fine-grained factual precision	Needs knowledge source
[9]	Farquhar et al.	2024	Semantic entropy	Detects confabulations across tasks	Sampling cost

Ref.	Authors	Year	Technique Used	Merits	Demerits
[14]	Dhuliawala et al.	2024	Chain-of-verification	Self-verification reduces errors	Extra generation steps

TABLE I Summary of Literature Survey

IV. COMPARATIVE STUDY

Table II compares detection and mitigation methods. Retrieval grounding reduces factual errors most directly, self-consistency and semantic entropy detect uncertainty without external data, fact-level verification gives fine-grained checks, and alignment improves general truthfulness.

Method	Type	External Data	Cost
RLHF alignment	Mitigation	No	High (training)
Retrieval grounding	Mitigation	Yes	Medium
Chain-of-verification	Mitigation	Optional	Medium
SelfCheckGPT	Detection	No	Medium–High
Semantic entropy	Detection	No	Medium–High
FActScore	Detection	Yes	Medium

TABLE II Comparison of Detection and Mitigation Methods

V. RESEARCH GAPS

- **G1:** No single benchmark captures hallucination across domains, languages and long-form tasks.
- **G2:** Detection methods based on sampling are computationally expensive.
- **G3:** Mitigation can reduce helpfulness by making models overly cautious.
- **G4:** Hallucination in Indian languages and code-mixed text is little studied.
- **G5:** Users are rarely shown calibrated confidence alongside answers.

VI. AIMS AND OBJECTIVES

- To explain the causes and taxonomy of LLM hallucination.
- To review benchmarks and detection methods.
- To review mitigation strategies across the LLM life cycle.
- To propose a hallucination-aware LLM application architecture.
- To identify open challenges and research directions.

VII. PROPOSED ARCHITECTURE

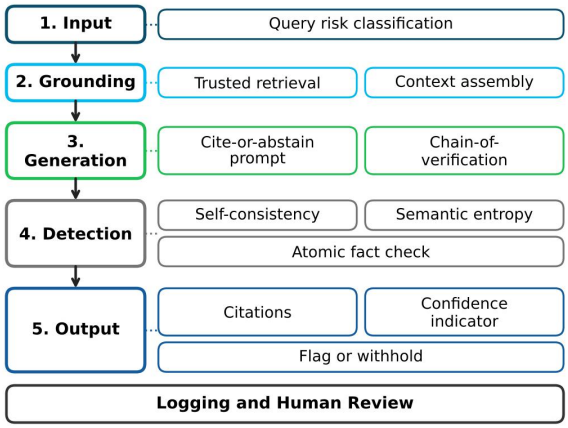


Fig. 2 presents a layered architecture for building hallucination-aware LLM applications.

Fig. 2. Proposed hallucination-aware LLM application architecture.

- **A. Input Layer**
- The user query is classified by risk; high-stakes topics such as health or law trigger stricter handling.
- **2. Grounding Layer**
- Relevant documents are retrieved from trusted sources and supplied as context.
- **3. Generation Layer**
- The LLM is prompted to answer from context, cite sources and say "I don't know" when unsure; chain-of-verification may be applied for high-risk queries.

- **4. Detection Layer** Sampled-answer consistency and semantic entropy estimate uncertainty; atomic facts are checked against sources.

E. Output Layer

Answers are shown with citations and a confidence indicator; low-confidence answers are flagged or withheld.

VIII. PROCESS FLOW AND EVALUATION

Fig. 3 shows the processing steps. Semantic entropy over meaning clusters C of sampled answers is

$$SE = -\Sigma_C\, p(C \mid x) \cdot \log p(C \mid x) \quad (1)$$

where high entropy indicates that the model's answers disagree in meaning and are likely unreliable. FActScore measures the proportion of supported atomic facts:

$$FS = (1/|A|)\, \Sigma_{a \in A}\ \ [a \text{ is supported}] \quad (2)$$

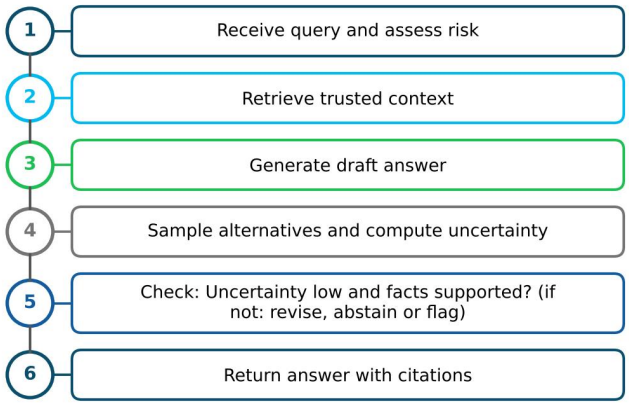


Fig. 3. Hallucination detection and mitigation workflow.

IX. LIMITATIONS AND FUTURE SCOPE

Table III summarizes challenges. Future work includes efficient single-pass uncertainty estimation, multilingual hallucination benchmarks, training objectives that reward abstention, and user interfaces that communicate confidence clearly.

Challenge	Mitigation
Cannot fully eliminate	Detect, cite and communicate uncertainty
Detection cost	Single-pass uncertainty probes
Over-cautious models	Calibrated abstention thresholds
Multilingual gaps	Benchmarks in Indian languages
User over-trust	Visible confidence and citations

TABLE III Challenges and Mitigation

X. CONCLUSION

Hallucination is an inherent risk of generative language models, arising from their training objective, data and decoding. The reviewed literature shows that hallucinations can be measured with benchmarks, detected through consistency, uncertainty and fact-level checking, and reduced through alignment, retrieval and verification prompting. Combining these methods in a layered application, and honestly communicating uncertainty, allows users to benefit from LLMs while limiting harm.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, 2023.
2. L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv:2311.05232, 2023.
3. Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," arXiv:2309.01219, 2023.
4. A. T. Kalai and S. S. Vempala, "Calibrated language models must hallucinate," in Proc. 56th Annu. ACM Symp. Theory Comput. (STOC), 2024, pp. 160–171.
5. J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proc. 58th Annu. Meeting ACL, 2020, pp. 1906–1919.

6. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. 60th Annu. Meeting ACL, 2022, pp. 3214–3252.
7. J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in Proc. EMNLP, 2023, pp. 6449–6464.
8. P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proc. EMNLP, 2023, pp. 9004–9017.
9. S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, pp. 625–630, 2024.
10. S. Min et al., "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proc. EMNLP, 2023, pp. 12076–12100.
11. S. Kadavath et al., "Language models (mostly) know what they know," arXiv:2207.05221, 2022.
12. L. Ouyang et al., "Training language models to follow instructions with human feedback," in Proc. NeurIPS, vol. 35, 2022, pp. 27730–27744.
13. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, vol. 33, 2020, pp. 9459–9474.
14. S. Dhuliawala et al., "Chain-of-verification reduces hallucination in large language models," in Findings ACL, 2024, pp. 3563–3578.