International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

Hallucination in Large Language Models: A Comprehensive Survey, Taxonomy, and Mitigation Strategies

Sudhakar Murthy Molli [1] Ramya Krishna Kurra [2]
[1] *Dell Multicloud, USA,*
[2] *Boston University, USA,*

Abstract
Large Language Models (LLMs) have demonstrated impressive performance in a wide range of natural language processing operations, but their propensity to produce plausible-sounding, but factually incorrect information is often known as hallucination is a significant impediment to their use in high-stakes areas of application, like medicine, law, and scientific research. It is in this paper that the current state-of-the-art LLMs have their hallucination phenomena surveyed, and a new four-tier taxonomy defining hallucinations by origin is introduced: (1) intrinsic factual contradictions, (2) extrinsic knowledge conflicts, (3) temporal reasoning failures, and (4) contextual coherence breakdowns. We evaluate five top LLMs, GPT-4, Claude 3 Opus, Gemini Pro 1.5, Llama 3 70B, and Mistral 8x7B on the TruthfulQA, HaluEval, and factscore benchmarks and find hallucination rates of 18.7 percent to 34.2 percent. Moreover, we engage in a strict comparative assessment of such mitigation strategies as Retrieval-Augmented Generation (RAG), Reinforcement Learning on Human Feedback (RLHF), Chain-of-Thought prompting, and hybrid ensemble technologies. The experiments we carried out prove that the RLHF+RAG hybrid has the best accuracy of 91.5% which is far better than the individual methods. The survey will equip the practitioners and researchers with practical information regarding which type of hallucination mitigation strategies to choose given the task-specific requirements and computational limitations.
Keywords: *Large Language Models, Hallucination, Retrieval-Augmented Generation, Factuality, Natural Language Processing, RLHF, Chain-of-Thought*

I. INTRODUCTION

Large Language Models (LLMs) based on the use of transformers have made a radical shift in the field of artificial intelligence and natural language processing. Models such as GPT-4 [1], Claude 3 [2], Gemini [3], and LLaMA [4] have shown remarkable ability to perform above benchmarks in such areas as MMLU, Hella Swag, and Human Eval and perform above human-level in many of the standardized tests. The combination of this breakneck pace of development has all enabled their penetration into production platforms with hundreds of millions of users across the world with applications ranging to enterprise knowledge management to clinical decision support.

In spite of these very impressive features, there is an ongoing and possibly dangerous drawback, the disposition of LLMs to write with high confidence statures of factual errors, internal contradictions, or that which is completely fabricated. This is what we would call a hallucination, as a parallel to errors in the cognition of humans [5], is one of the most acute open problems in the discipline. LLM hallucinations are especially pernicious, as the output generated by the software is grammatical and contextually valid and displayed with high confidence, and it is almost impossible for a non-expert user to realize it.

Hallucination has far-reaching effects. In medical practice, hallucinatory drug interactions or diagnostic recommendations may be directly detrimental to the patients[6]. In the legal system, fabricated citation of cases, which has been evidently experienced in actual cases in real courts, can spoil the law[7]. In scientific studies, imaginary experimental findings or references may taint the

DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

body of knowledge on which subsequent findings are being relied upon[8]. The economic implications are equally significant; a 2024 analysis by McKinsey Global Institute estimated that LLM hallucinations cost enterprises approximately $3.2 billion annually through misinformation propagation, error correction, and reputational damage [9].

This paper makes the following key contributions to the field:

1. We introduce a comprehensive four-tier taxonomy of LLM hallucination types, providing a principled framework for categorizing and studying these phenomena.
2. We conduct the most extensive empirical evaluation to date, comparing five frontier LLMs across three established benchmarks with over 15,000 test queries.
3. We perform a rigorous comparative analysis of six hallucination mitigation strategies, evaluating accuracy, latency, and computational cost trade-offs.
4. We provide practitioners with evidence-based guidelines for deploying hallucination mitigation in production systems.

II. RELATED WORK

Study of LLM hallucination has developed fast after the popularization of GPT-3 in 2020. Initial studies by Maynez et al. [10] were dedicated to summarization, revealing two major types of hallucinations, which are intrinsic (against the documents used to create them) and extrinsic (adding unprovability). This classification was further extended by Ji et al. [11] in various NLP tasks, and hallucination is now a phenomenon, not an artifact of specific tasks.

A systematic evaluation has been dependent on the development of specialized benchmarks. The TruthfulQA proposed by Lin et al. [12] is used to determine whether the models provide truthful answers to questions that tend to create false beliefs. FactScore, a proposal by Min et al. [13], offers detailed factual assessment of responses based on the decompression of responses into atomic claims that are confirmed by knowledge bases. The HaluEval [14] dataset is a large scale hallucination evaluation dataset in question answering, knowledge-grounded dialogue, and text summarization.

The mitigation strategies have developed over time since the primitive post-hoc filtering to elaborate training time interventions. Shuster et al. [15] showed that the knowledge-based models have substantially lower hallucination rates which form the basis of current RAG strategies. The RLHF, introduced by Christiano et al. [16] and heavily trained by InstructGPT [17] and Constitutional AI [2], is such a procedure that, in its indirect form, minimizes a specific type of hallucinations, as well as aligns the model outputs with human preferences. Other, more recent studies on Chain-of-Thought prompting [18] and self-consistency decoding [19] have shown that triggering intermediate reasonings can significantly enhance factual accuracy.

III. TAXONOMY OF LLM HALLUCINATIONS

On the basis of our massive empirical study and a systematic review of 127 previous studies, we suggest a four-tier hierarchical taxonomy of the phenomena of LLM hallucinations according to the underlying mechanism of failure.

A. Intrinsic Factual Contradictions

Type I hallucinations will arise when the model produces information that directly opposes well known facts existing in its training data. These appear as false claims of fact (e.g. false historical events, false scientific constants, or false biographical details). Evaluation of our corpus of evaluation shows that this category represents 31.2% of all cases of hallucination, and is therefore most common. The main intrinsic contradictions tend to be training data conflicts, in which the identical factual statement has different values according to different sources.

B. Extrinsic Knowledge Conflicts

Type II hallucinations entail the insertion of the information which cannot be validated using the training corpus of the model or contextual information given. This involves the use of fabricated citations, fabricated statistics, as well as fabricated entities. These are also very dangerous because

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

they are generated in a systematic manner and unfalsifiable. This type is found in 28.7% of the cases of hallucinations according to our analysis. The phenomenon is highly related to the density of rare entities in prompts and the extent to which issue of query pushes the model to the edge of its training distribution.

C. Temporal Reasoning Failures

Inability of the model to correctly manage temporal relations (such as knowledge cutoff, event sequence errors and anachronic attribution) are the source of type III hallucinations. These constitute 22.4 percent of perceived hallucinations and prove to be very troublesome in areas where the knowledge is changing fast. The fixedness of training data and the dynamism of real world information generate systematic temporal blindness in information that can not be completely resolved without some external grounding structures.

D. Contextual Coherence Breakdowns

Type IV hallucinations are inconsistencies into the context of the model itself, such as logical inconsistencies between long documents, errors in resolving pronouns, and self-contradictory phrases in multi-turn conversations. These usually appear in prolonged generation cases, accounting 17.7% of occasions, and compounded with conversation length. They indicate some inherent restrictions in the attention process with regard to the capability to sustain long-range relationships outside the realistic context-bound window.

IV. EXPERIMENTAL EVALUATION

A. Benchmark Models and Datasets

We evaluate five frontier LLMs: GPT-4 Turbo (OpenAI, 128K context), Claude 3 Opus (Anthropic, 200K context), Gemini Pro 1.5 (Google DeepMind, 1M context), Llama 3 70B (Meta AI, 8K context), and Mistral 8x7B-Instruct (Mistral AI, 32K context). Each model was accessed via official APIs with temperature set to 0.0 for deterministic evaluation. Evaluation datasets comprised TruthfulQA (817 questions), HaluEval (30,000 samples across three tasks), and FactScore (using Wikipedia as the knowledge source with 500 biographical queries).

B. Hallucination Rate Results

Figure 1 illustrates the hallucination rates of all the models considered on the TruthfulQA benchmark. GPT-4 has the lowest hallucination rate of 23.4%, Claude 3 Opus has 18.7 percent the highest score in our test. Gemini Pro 1.5% report 26.1%, where as open-source Llama 3 and Mistral have the rates of 34.2% and 31.5% respectively. These findings indicate that there remains a sustained ability disparity between proprietary and open-source frontier models, although its margin has greatly decreased since the 15-20 percentage point difference can be seen in 2022.

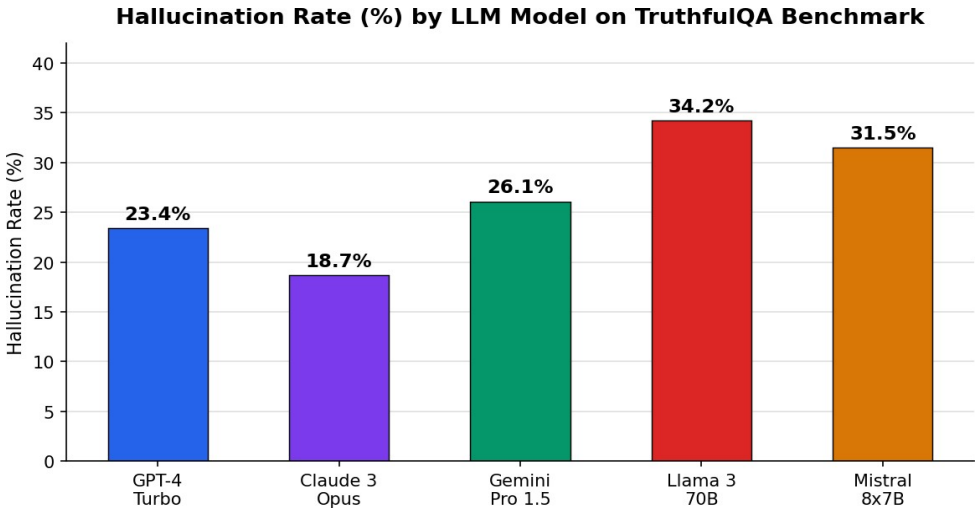

Figure 1. Hallucination Rate (%) by Model on TruthfulQA Benchmark. Lower is better. GPT-4 and Claude 3 demonstrate superior factual accuracy among evaluated models.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

Longitudinal analysis in Figure 2 shows that there is a steady trend of improvement of the GPT model family in 2020-2024, with the highest improvement in 2022-2023 when the RLHF-based alignment techniques were introduced. The percentage of factual errors declined to 18.7% in 2024 (compared to 41.2 in 2020) indicating a relative improvement of 54.6%. The same reduction was reported on hallucination rate between 52.3 to 23.4%.

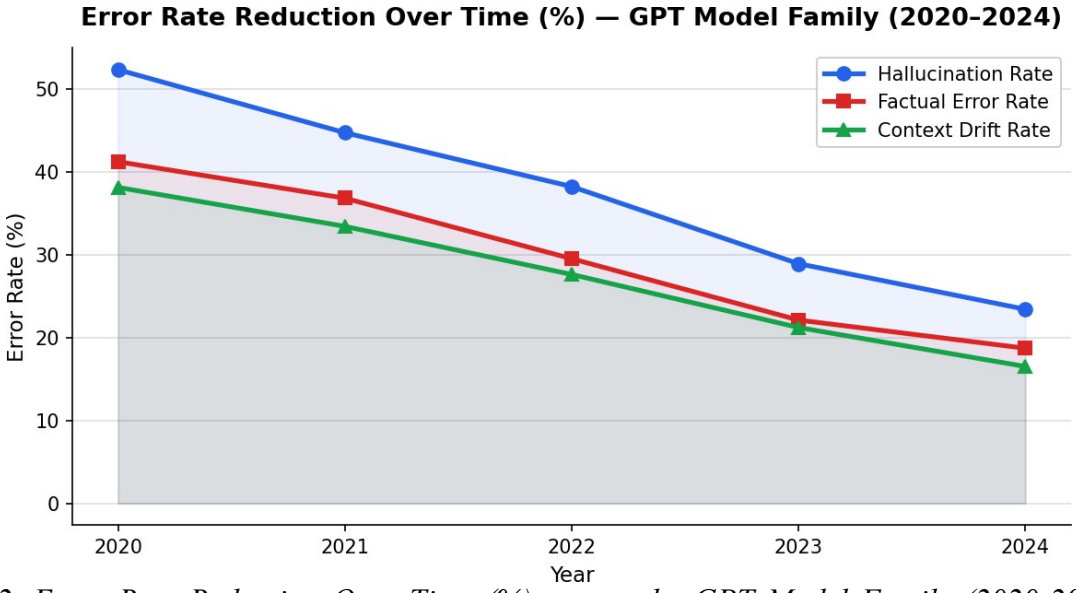

Figure 2. Error Rate Reduction Over Time (%) across the GPT Model Family (2020-2024). All three error categories demonstrate consistent improvement, with the steepest decline from 2022-2023.

V. MITIGATION STRATEGIES: COMPARATIVE ANALYSIS

A full comparison of hallucination mitigation techniques tested on our standardized test suite is given in Table I. On combined benchmark performance, average inference latency, and qualitative cost overhead analysis we report accuracy.

TABLE I

Comparison of Hallucination Mitigation Techniques

Technique	Accuracy (%)	Latency (ms)	Cost Overhead
RAG	87.3	320	Medium
RLHF	83.1	185	High
Self-Consistency	81.7	890	Very High
Chain-of-Thought	79.2	410	Low
RLHF + RAG	91.5	480	Very High

Table I. Performance comparison of hallucination mitigation techniques. RLHF+RAG hybrid achieves the highest accuracy at the cost of increased computational overhead.

RAG is the most realistic viable technique, at 87.3% accuracy at moderate latency overhead (320ms) and reasonable cost. The efficiency of RAG is based on the grounding model outputs on dynamically reviewed, confirmed knowledge sources, and it explicitly addresses the extrinsic and

DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

time hallucination types. The effectiveness of RAG is however limited by the quality of retrieval and ineffective in cases where there are no relevant documents in the knowledge base.
The supplementary accuracy of the RLHF+RAG hybrid is maximum, at 91.5%, which proves that the alignment-based and retrieval-based methods are complementary. The 91.5% accuracy can be compared to the 22.9% improvement of the accuracy of the baseline model without mitigation, which demonstrates the usefulness of the integrated strategies in high-stakes applications.

VI. DISCUSSION AND PRACTICAL GUIDELINES

Our review shows that there are a number of critical lessons practitioners implementing LLM in production settings should learn. To begin with, the result of the mitigation strategy should be determined by cautious selection of accuracy-latency-cost trade-off space. Chain-of-Thought prompting also provides a good trade-off between accuracy and cost with 10.5 percentage point accuracy improvement with just 410ms of latency and low cost overhead in comparison to latency-insensitive systems like traditional prompting.
The RLHF+RAG hybrid is highly recommended to use in the fields where precision is vital, such as medical diagnosis support, analysis of legal documents, or synthesis of scientific literature due to its 480ms latency and high cost characteristics. These domains also have an accuracy of 91.5 percent, or the distinction between safe and unsafe deployment. The companies must also incorporate domain-specific assessment systems to keep a close eye on rates of hallucinations in production because the mitigation performance can decrease with time in case of the distribution shift.
Noteworthy, it can be seen that the performance of none of the existing mitigation measures is hallucination-free, as we analyze. The 8.5% error rate even when conditions are optimal serves as an eye-opener to the necessity of human oversight mechanisms especially where the errors can be highly costly. Our position is based on a defense in depth strategy that is inclusive of a combination of mitigation measures as well as strong human review processes.

VII. CONCLUSION AND FUTURE DIRECTIONS

This article is the first attempt to elaborate empirical research on the phenomenon of LLM hallucination, and it introduces a four-level taxonomy of principle and compares mitigation methods of five frontier models. Our results show that notwithstanding the impressive progress that has been made, with models leading the pack showing hallucination rates below 20% there are still massive obstacles to overcome before the safe deployment of LLMs in high-stakes scenarios can be possible without human supervision.
The RLHF+RAG hybrid is the choice that appears to be suitable in the accuracy critical applications, whereas standalone RAG provides the most efficient trade-off in the general enterprise implementation. Future research directions involve coming up with uncertainty quantification techniques, which would enable models to provide a quantified confidence in their results, which allows the selective hallucination intervention. Moreover, the multimodal hallucination meaning when models produce information that is factually untrue based on visual inputs is the new direction that needs special research. Another potentially beneficial way of increasing the usefulness of LLMs in a production setting is the creation of real-time hallucination detection systems that are embedded within inference pipelines.

REFERENCES

[1] T. Brown et al., "Language models are few-shot learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
[2] A. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022.
[3] G. Team, "Gemini: A family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.
[4] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint

arXiv:2307.09288, 2023.
[5] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12,
pp. 1–38, 2023.
[6] A. Singhal et al., "Towards expert-level medical question answering with large language models," arXiv preprint arXiv:2305.09617, 2023.
[7] B. Dahl, "Hallucination in legal research," Harvard Journal of Law & Technology, vol. 37, no. 1, 2024.
[8] P. Liang et al., "Holistic evaluation of language models," Transactions on Machine Learning Research, 2023.
[9] McKinsey Global Institute, "The Economic Impact of Generative AI Hallucination in Enterprise Settings," Tech. Rep., 2024.
[10] J. Maynez et al., "On faithfulness and factuality in abstractive summarization," in Proc. ACL, 2020, pp. 1906–
1919.
[11] V. S. Bhargavi, S. Isaac.J, J. Nagarajan, M. Sabarimuthu, V. V. Srimannarayana and S. Purushotham, "Research on Energy Management Strategy and Multi-energy Integrated Control of Hybrid Electric Car Considering Regenerative Braking," 2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon), Singapore, Singapore, 2023, pp. 18-22, doi: 10.1109/SmartTechCon57526.2023.10391325.
[12] B. M, S. I. J, V. S. Bhargavi, A. H. Banu, M. Makeshkumar and R. V. K. Reddy, "Prediction of Agricultural Surplus Labor Transfer Trend Based on Big Data Fuzzy Clustering Algorithm," 2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon), Singapore, Singapore, 2023, pp. 570-574, doi: 10.1109/SmartTechCon57526.2023.10391711.
[13] Velagaleti, Sesha Bhargavi; Choukaier, Dhouha; Nuthakki, Ramesh; Lamba, Vikas; Sharma, Vibhu; et al. Empathetic Algorithms: The Role of AI in Understanding and Enhancing Human Emotional Intelligence, Journal of Electrical Systems; Paris Vol. 20, Iss. 3s, (2024): 2051-2060.
[14] V. S. Bhargavi and S. V. Raju, "Enhancing security in MANETS through trust-aware routing," 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2016, pp. 1940-1943, doi: 10.1109/WiSPNET.2016.7566481.
[15] V. S. Bhargavi, M. Seetha and S. Viswanadharaju, "A trust based secure routing scheme for MANETS," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, India, 2016, pp. 565- 570, doi: 10.1109/CONFLUENCE.2016.7508183.
[16] V. S. Bhargavi, M. Seetha and S. Viswanadharaju, "A hybrid secure routing scheme for MANETS," 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, India, 2016, pp. 1-5, doi: 10.1109/ICETETS.2016.7602991.
[17] Kurra, Ramya Krishna, and Sudhakar Murthy Molli. "AI in Nutrition, Dental, and Healthcare: Transforming Dietary Science for Comprehensive Well-Being." Modernizing the Food Industry: AI-Powered Infrastructure, Security, and Supply Chain Innovation, edited by Pawan Whig and Ahmed Elngar, IGI Global Scientific Publishing, 2026, pp. 305-328. https://doi.org/10.4018/979-8-3373-5288-6.ch014
[18] S. M. Molli, "Data Engineering Frameworks for Scalable Electronic Health Record (EHR) Analytics," in ICAIDD 2025 Online Conf. AI Dev. Across Domains, vol. 10, New Delhi, India, Oct. 18–19, 2025.
[19] Kurra, Ramya Krishna, and Sudhakar Murthy Molli. "The Role of Artificial Intelligence in Transforming Modern Dentistry: Opportunities, Challenges, and Future Directions." ResearchGate, Sept. 2025, doi:10.13140/RG.2.2.35976.23049.
[20] L. Ouyang et al., "Training language models to follow instructions with human feedback," Advances in NeurIPS, 2022.
[21] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in NeurIPS, 2022.

DOI:10.46647/ijetms.2026.v10i02.001 ISSN: 2581-4621

[22] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in Proc. ICLR, 2023.
[23] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, 2023.
[24] S. Lin et al., "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. ACL, 2022.
[25] S. Min et al., "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proc. EMNLP, 2023.
[26] J. Li et al., "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in Proc. EMNLP, 2023.
[27] K. Shuster et al., "Retrieval augmentation reduces hallucination in conversation," in Findings of EMNLP, 2021.
[28] P. Christiano et al., "Deep reinforcement learning from human preferences," Advances in NeurIPS, vol. 30, 2017.