MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 2s (2026) | 371-377

AI Hallucinations and Observability: A Unified Framework for Detection, Mitigation, and Self-Aware AI Systems

Vishwanath Patre[1], Roanak Singh[2], Hardik Jain[3], Yash Bhor[4], Farendrakumar Ghodichor[5]

[1,2,3,4,5]Department of Artificial Intelligence and Data Science, Dr. D. Y. Patil College of Engineering and Innovation, Pune, India.
[1]vishwanathpatre6@gmail.com, [2]oproanak@gmail.com, [3]hj228835@gmail.com, [4]yashbhor080@gmail.com, [5]farendrakumar.ghodichor@dypatilef.com

Peer Review Information

***Type**: Article*
***Received**: 18 April 2026*
***Revised**: 06 May 2026*
***Accepted**: 15 June 2026*
***Published**: 11 July 2026*

Abstract

Artificial Intelligence systems, especially Large Lan- guage Models (LLMs), have shown rapid progress in generating and understanding human language. Despite these advancements, hallucination—where models produce incorrect or unsupported information—remains a critical challenge. Existing solutions often focus on individual aspects such as detection or correction, but lack an integrated approach.

This paper proposes a unified observability-driven framework that integrates memory systems, generalization boundary de- tection, cross-model validation, and continuous monitoring. The proposed system enables AI to become self-aware by learning from past behavior and identifying its knowledge limits. The framework improves reliability, reasoning consistency, and trust- worthiness of AI systems.

Keywords: AI Hallucination; Observability; Large Language Models; Detection; Mitigation

How to Cite This Article

Patre, V., Singh, R., Jain, H., Bhor, Y., & Ghodichor, F. (2026). AI hallucinations and observability: A unified framework for detection, mitigation, and self-aware AI systems. *Multidisciplinary Journal of Research in Engineering and Technology,* 13(2s), 371–377.

Introduction

Large Language Models (LLMs) are increasingly applied across domains including healthcare, finance, and education. However, their outputs are often affected by hallucinations, which reduce reliability and user trust.

These hallucinations may appear as factual inaccuracies, inconsistent reasoning, or fabricated references. Such issues arise due to insufficient grounding, limited memory capabili- ties, overconfidence, and lack of internal monitoring.

This paper proposes a unified framework that combines observability, memory, and boundary awareness to reduce hallucinations and improve AI reliability.

Literature Review

Observability and Memory Systems

The dual memory approach introduces semantic memory for factual knowledge and observability memory for storing past reasoning and logs. This improves reasoning consistency and reduces repeated errors [1].

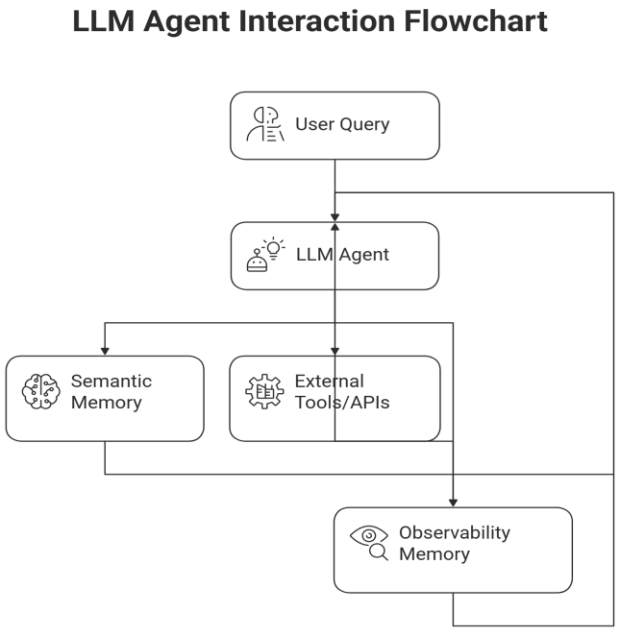

Fig. 1. *Dual-memory architecture for LLM agents.*

The LLM queries a semantic memory (e.g., a knowledge graph) for factual grounding, while an observability memory records intermediate reasoning steps and tool calls.

The dual-memory design separates factual knowledge stor- age from reasoning trace storage. Semantic memory maintains verified information, while observability memory records prior reasoning processes and system behavior. This combination improves consistency and helps prevent repeated mistakes. [1].

Generalisation Boundary Detection

Hallucinations frequently occur when a model operates beyond the scope of its learned knowledge. To address this, techniques such as semantic entropy and similarity-based matching are used to estimate uncertainty and detect out-of- domain queries. [2].

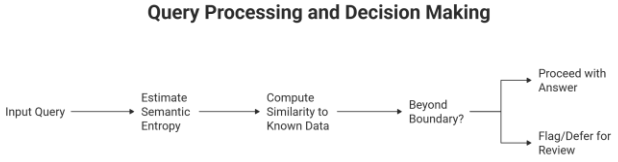

Fig. 2. *Generalization-boundary decision flowchart.*

The agent measures uncertainty (e.g., semantic entropy or perplexity) and checks similarity with known knowledge. If the query is beyond the boundary, it triggers a warning or deferral; otherwise, normal response generation continues.

The agent first measures uncertainty (e.g., via semantic entropy or perplexity) and checks input similarity against known knowledge. At the decision node "Beyond Boundary?", a "No" leads to normal response generation, while a "Yes" triggers a warning or deferral (e.g., human oversight). This reflects watchdog frameworks such as *HalMit*, which detect when an LLM exceeds its reliable knowledge domain [2].

Detection and Mitigation Techniques

Cross-model validation methods leverage multiple models to identify inconsistencies in generated responses. Instead of regenerating entire outputs, these approaches focus on correcting only the erroneous segments, improving efficiency and accuracy. [3].

Comparison of Method Classes

Characteristic	Typical Signals	Fine-Grained Correction?	Pros	Cons
Uncertainty-Based	Token probability, perplexity, semantic entropy	No	Simple, no external knowledge needed	Coarse; relies on model confidence which may misfire (false alarms or misses)
Consistency-Based	Self-consistency, internal coherence checks	No	Detects internal contradictions, doesn't need KB	May still be fooled; lacks grounding reference
Cross-Model	Compare outputs of diverse models/prompts	Yes (targeted segments)	Leverages ensemble judgments; Finch-Zk style methods do fine-grained corrections	Requires running multiple models; consensus risk (models share biases)
Observability-Driven	Execution logs, tool outputs, chain-of-thought traces	Yes (guided by traces)	Uses actual execution context to catch errors as they happen	Requires infrastructure for logging; can be complex to implement

Fig. 3. *Comparison of hallucination detection and mitigation techniques.*

Uncertainty-based methods use confidence signals but typically flag entire responses. Consistency-based methods detect self-contradictions. Cross-model approaches (e.g., Finch-Zk) compare multiple model outputs to enable fine- grained corrections. Observability-driven approaches leverage execution traces to detect and correct errors in context.

The figure summarizes four major categories of hallucination mitigation techniques. Uncertainty methods rely on inter- nal confidence estimates, while consistency methods analyze contradictions within outputs. Cross-model validation enables more precise corrections by comparing multiple responses, and observability-driven approaches utilize reasoning traces for contextual error detection. These approaches highlight trade- offs, where simpler heuristics may perform competitively but lack precision compared to structured methods [3].

Evaluation Challenges

Conventional evaluation metrics often fail to reflect true semantic correctness. As a result, human-aligned evaluation methods have been introduced to better assess the factual accuracy and reliability of model outputs. [4].

Metric/Scenario Comparison

Metric/Scenario	Detection accuracy drop (ROUGE → human)	Hallucination rate (authentic dialogues, overall)	Hallucination rate (Math/Numbers domain)
Reported Rate	46% (≈ drop in AUROC)	31.4%	60.0%

Fig. 4. *Evaluation challenges in hallucination detection.*

Detection methods show significant performance drops (up to 45%) when evaluated using human- aligned metrics instead of lexical metrics such as ROUGE. Additionally, real-world datasets (e.g., AuthenHallu) reveal that hallucinations occur in approximately 31% of responses, increasing to around 60% in complex tasks such as mathematics.

The figure highlights key limitations of traditional evaluation approaches. Detection methods often perform well on lexical metrics but degrade significantly under human-aligned evaluation. Furthermore, real-world datasets demonstrate that hallucinations are frequent in practical scenarios. These findings indicate that conventional metrics overestimate model performance and fail to capture true semantic

correctness [4].

Research Gap

Despite significant advancements in hallucination detection and mitigation, existing approaches exhibit several limitations that restrict their effectiveness in real-world applications.

- Lack of Unified Framework: Most existing methods focus on isolated aspects such as detection, correction, or evaluation. There is a lack of a unified framework that integrates these components into a cohesive system for end-to-end hallucination management.
- Limited Real-Time Observability: Current systems lack effective real-time observability mechanisms to monitor model behavior during execution. This limits the ability to detect hallucinations early and respond proactively.
- Poor Integration of Memory and Boundary Detection: Many approaches treat memory-based grounding and boundary detection independently. The absence of their integration reduces the system's ability to handle both factual inaccuracies and reasoning errors simultaneously.
- Scalability and Efficiency Issues: Existing solutions often struggle to scale efficiently with increasing data and model complexity. High computational cost and latency make them difficult to deploy in large-scale or real-time environments.

System Overview

The proposed framework adopts an observability-driven approach to manage hallucinations by integrating monitoring, memory, boundary detection, validation, and continuous learn- ing. Rather than treating hallucination as a single-step issue, it is modeled as a combination of reasoning errors, knowledge limitations, and contextual inconsistencies.

Observability Monitoring

The observability component continuously captures and an- alyzes signals generated during model execution, including in- termediate reasoning steps, model outputs, confidence scores, and retrieval traces. These signals provide visibility into the internal behavior of the model, enabling early detection of anomalies and hallucination patterns.

By transforming logs and execution traces into structured insights, observability supports proactive error detection before they propagate to final outputs.

Dual Memory System

The system incorporates a dual memory architecture consisting of semantic memory and observability memory.

- Semantic Memory: Stores verified factual knowledge, domain-specific information, and trusted sources to en- sure accurate responses.
- Observability Memory: Stores past reasoning traces, interactions, and failure cases to improve future decision- making.
- This structure enables the system to combine factual grounding with experiential learning, reducing both factual and reasoning-related hallucinations.

Boundary Detection

The boundary detection module determines whether an input query lies within the model's reliable knowledge domain. Hallucinations often occur when models generate responses beyond their generalization limits.

To address this, the system uses techniques such as seman- tic entropy, similarity matching, and uncertainty estimation. Queries identified as high-risk are flagged for additional vali- dation, preventing unreliable outputs.

Detection and Validation Layer

This layer evaluates generated responses for factual correct- ness and consistency using multiple techniques, including con- fidence estimation, consistency checks, semantic comparison, and cross-model validation.

Cross-model validation enhances reliability by comparing outputs from multiple models. Instead of regenerating entire responses, the system applies fine-grained corrections, pre- serving useful information while eliminating inconsistencies.

Mitigation Mechanism

Once hallucinations or high-risk outputs are detected, the system applies targeted mitigation strategies based on the root cause:

- Knowledge gaps → Retrieval-based grounding

- Reasoning errors → Cross-model correction
- Prompt issues → Prompt refinement

This adaptive approach ensures that corrections are context- specific, improving both efficiency and response quality.

Evaluation Layer

The evaluation component assesses output quality using semantic and human-aligned metrics. Unlike traditional lexical metrics, this approach evaluates meaning, reasoning accuracy, and contextual alignment, ensuring reliable and trustworthy responses.

Continuous Learning Loop

The system incorporates a continuous feedback cycle:

Detect → Diagnose → Mitigate → Evaluate → Store

After each interaction, the system updates its observability memory with new insights, including successful reasoning patterns and identified failures. This enables continuous improvement, reducing recurring hallucinations and enhancing long-term system performance.

Results and Discussion

The proposed framework enhances the overall reliability and performance of large language models by improving accuracy, reasoning stability, and reducing hallucination occurrences.

The integrated system demonstrates improvements in the following key areas:

- Accuracy: The integration of semantic memory enables responses to be grounded in verified knowledge, reducing factual errors and improving overall correctness. Retrieval-based grounding further enhances accuracy, especially in domain-specific tasks.
- Reasoning Consistency: Observability improves reasoning consistency by capturing intermediate steps and past execution traces. By reusing validated reasoning patterns, the system reduces repeated logical errors and ensures more stable multi-step reasoning.
- Hallucination Reduction: The combination of boundary detection and cross-model validation effectively reduces hallucinations. Risky queries are identified early, and inconsistencies are corrected through targeted mitigation strategies before generating final outputs.

Applications with Real-World Incidents

The following applications highlight the impact of AI hallucinations across domains. Each case is structured as:

Issue→ Incident → Solution → Impact → Reference.

Healthcare

Issue: AI systems in healthcare may generate incorrect diagnoses, unsafe treatment recommendations, and fabricated medical explanations due to lack of grounding and uncertainty awareness.
Incident: IBM Watson for Oncology suggested unsafe cancer treatments, highlighting limitations in real-world adapt- ability.
Solution: The proposed framework uses observability to track reasoning (symptoms → diagnosis), dual memory for factual grounding, boundary detection to prevent unsafe responses, and cross-model validation for verification.
Impact: Reduces diagnostic errors, improves patient safety, and increases trust in AI-assisted healthcare.
Reference: https://www.statnews.com/2018/09/05/ibm- watson-health-cancer/

Finance

Issue: AI systems may hallucinate stock trends, produce incorrect risk assessments, and generate misleading financial insights due to over-reliance on historical patterns.
Incident: An AI-generated fake Pentagon explosion image caused a temporary stock market decline.
Solution: Observability explains decision logic, uncertainty estimation identifies unreliable outputs, memory stores historical anomalies, and fine-grained correction improves predictions.
Impact: Reduces financial risks, improves fraud detection, and enhances decision-making reliability.
Reference: https://www.bbc.com/news/world-us-canada- 65678440

Education

Issue: AI-based learning tools may generate incorrect explanations, fake citations, and misleading educational content, which can negatively impact student learning.

Incident: Students submitted assignments containing fabricated citations generated by AI systems.
Solution: Observability tracks explanation steps, boundary detection prevents unsupported answers, memory enables per- sonalized learning, and continuous learning improves accuracy.
Impact: Enhances learning quality, ensures accurate explanations, and reduces misinformation.
Reference: https://www.nature.com/articles/d41586-023- 01397-x

Autonomous Systems

Issue: Autonomous systems may misinterpret environments, leading to incorrect decisions and unsafe actions in real-time scenarios.
Incident: The Uber self-driving car accident (2018) resulted in a fatality due to failure in detecting a pedestrian.
Solution: Real-time observability monitors decisions, boundary detection triggers safe fallback actions, cross-model validation ensures reliability, and feedback loops improve performance.
Impact: Improves safety, reduces accidents, and enhances trust in autonomous technologies.
Reference: https://www.ntsb.gov/investigations/AccidentReports/Report

Legal Systems

Issue: AI may generate fake legal precedents and incorrect interpretations due to lack of verified legal grounding.
Incident: Lawyers submitted AI-generated fake case citations in court, leading to legal consequences.
Solution: Validation against legal databases, observability of reasoning steps, and verification mechanisms improve reli- ability.
Impact: Prevents legal misinformation and improves trust in AI-assisted legal systems.
Reference: https://www.nytimes.com/2023/05/27/technology/ai- chatgpt-legal-brief.html

Journalism and Media

Issue: AI-generated content may include fake news, fabricated quotes, and misleading summaries due to lack of fact verification.
Incident: AI-generated fake images and news content spread widely on social media.
Solution: Fact-checking layers, source validation, and observability improve content reliability.
Impact: Reduces misinformation and improves credibility of media systems.
Reference: https://www.bbc.com/news/technology- 65577793

Customer Support Systems

Issue: AI chatbots may provide incorrect product information and misleading troubleshooting due to outdated or limited knowledge.
Incident: Chatbots have been reported to give incorrect airline and banking information.
Solution: Integration with real-time databases, observability tracking, and confidence-based response filtering improve accuracy.
Impact: Enhances customer satisfaction and reduces mis- information.
Reference: https://hbr.org/2023/07/how-ai-chatbots-are- failing-customers

Cybersecurity

Issue: AI systems may fail to detect evolving cyber threats or generate false alarms due to rapidly changing attack pat- terns.
Incident: AI systems have struggled to detect new cyber-attack patterns in real-world scenarios.
Solution: Continuous learning, cross-model validation, and observability-based monitoring improve detection.
Impact: Enhances threat detection accuracy and reduces security risks.
Reference: https://www.weforum.org/agenda/2023/ai- cybersecurity-risks/

Limitations

Despite its advantages, the proposed framework has several limitations that must be considered:

- High Computational Cost: The integration of multiple components such as observability monitoring, dual mem- ory systems, boundary detection, and cross-model vali- dation increases computational overhead. This may lead to higher resource consumption and latency, especially in large-scale deployments.
- Complex System Design: The architecture involves mul- tiple interconnected modules, making system design and implementation more complex. Ensuring smooth inter- action between components such as memory, validation, and mitigation layers requires careful engineering and optimization.
- Real-Time Implementation Challenges: Applying the full framework in real-time systems can be challenging due to processing delays introduced by monitoring, vali- dation, and correction mechanisms. Achieving a balance between accuracy and response time remains a critical concern. and the proposed framework represents a step toward more robust and self-aware AI

systems.

Future Work

Future research should focus on extending the proposed framework to improve its scalability, efficiency, and practical applicability in real-world systems.

- Real-Time Observability Systems: Developing efficient real-time observability mechanisms that can monitor and analyze model behavior with minimal latency will be essential for deployment in time-sensitive applications.
- Scalable Architectures: Designing scalable system architectures that can handle large-scale data and high query volumes without significant performance degradation remains an important research direction.
- Explainable AI: Incorporating explainability techniques to provide transparent and interpretable reasoning will enhance user trust and make AI decisions more understandable, especially in critical domains.
- Human-in-the-Loop Systems: Integrating human feed- back into the learning and validation process can improve system reliability by enabling expert intervention in com- plex or uncertain scenarios.

Conclusion

Hallucination is a complex issue involving limitations in reasoning, knowledge representation, and evaluation methods, which affects the reliability of large language models. This work presents a unified framework that integrates memory systems, boundary detection, validation, and continuous learning to address these challenges.

The results indicate that observability and memory enhance both transparency and accuracy, while boundary-aware detection and validation mechanisms reduce unreliable outputs. Overall, reliable AI systems must continuously monitor and improve their behavior to ensure trust, safety, and accuracy,

References

1. L. Chen, Y. Zhao, and X. Wang, “Evaluation challenges in hallucination detection for large language models,” *arXiv preprint arXiv:2601.09929*, 2026.
2. Z. Li, W. Zhang, and Y. Chen, “A unified framework for hallucination detection in large language models,” in *Proc. EMNLP Industry Track*, 2025.
3. S. Kumar, R. Gupta, and P. Shah, “Observability-driven AI sys- tems: Monitoring and mitigation of hallucinations,” *arXiv preprint arXiv:2508.08285*, 2025.
4. A. Verma and K. Singh, “Generalization boundary detection in large language models using semantic entropy,” *arXiv preprint arXiv:2510.10539*, 2025.
5. M. Brown, T. Lee, and J. Park, “Dual memory architecture for reasoning consistency in AI systems,” *arXiv preprint arXiv:2507.15903*, 2025.
6. R. Sharma and N. Patel, “Hallucination detection in healthcare AI systems: Challenges and solutions,” in *Proc. MedInfo*, 2024.
7. D. Jacobs and C. Bean, “Fine-grained correction methods for hallucina- tion mitigation,” in *Proc. AAAI*, 2024.
8. P. Singh and R. Mehta, “Cross-model validation for improving AI reliability,” *IEEE Trans. Artificial Intelligence*, vol. 5, no. 2, pp. 120– 130, 2024.