International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

AI Hallucination Detection in Large Language Models

Vennila.M[1], Priya.P[2]

[1]Assistant Professor, Department of Computer Science with AI, Mahalashmi Women's College of Arts and Science
[2]Assistant Professor, Department of Computer Application, Mahalashmi Women's College of Arts and Science

***Abstract*— Artificial Intelligence (AI) has advanced through Large Language Models (LLMs) that generate human-like text. However, these models may produce incorrect or fabricated information, known as AI hallucinations, affecting their reliability in healthcare, finance, education, and legal applications. Detecting hallucinations has become an important research area. This article presents the causes, types, detection techniques, applications, and challenges of AI hallucination detection. It also highlights recent approaches such as Retrieval-Augmented Generation (RAG), Explainable AI (XAI), Natural Language Inference (NLI), and fact verification.**

***Keywords*— Artificial Intelligence, Hallucination Detection, Large Language Models, Explainable AI, Retrieval-Augmented Generation, Natural Language Inference, Fact Verification, Trustworthy AI.**

I. Introduction

Artificial Intelligence has rapidly advanced with Large Language Models (LLMs) such as ChatGPT, Gemini, Claude, and Llama, enabling applications in healthcare, education, finance, and software development. However, these models may generate incorrect or fabricated information, known as AI hallucinations, reducing their reliability and trustworthiness. Using methods like Retrieval-Augmented Generation (RAG), Explainable AI (XAI), Natural Language Inference (NLI), and fact verification, AI hallucination detection seeks to detect and validate such deceptive content, guaranteeing more precise and reliable AI-generated solutions.

II. Objectives Of AI Hallucination Detection

The primary objectives of AI hallucination detection include:

- Improve the reliability of AI-generated responses.
- Detects fabricated or unsupported information.
- Verify generated content using trusted knowledge sources.
- Increase transparency and explainability of AI systems.
- Enhance user trust in AI-based applications.

III. Types Of AI Hallucinations

AI hallucinations occur in different forms depending on the type of information generated by the model. Understanding these categories helps researchers develop effective detection techniques.

A. Factual Hallucination

Factual hallucination occurs when an AI model generates information that is incorrect or unsupported by reliable sources. Examples include incorrect historical events, wrong scientific facts, fabricated statistics, or inaccurate medical advice. This is the most common type of hallucination observed in LLMs.

B. Citation Hallucination

Citation hallucination refers to the generation of fake research papers, incorrect author names, fabricated Digital Object Identifiers (DOIs), or non-existent journal references. This problem frequently appears when AI is asked to generate academic content.

C. Logical Hallucination

Logical hallucination occurs when the reasoning process becomes inconsistent or contradictory. Although the generated sentences appear grammatically correct, the logical relationship between statements is invalid.

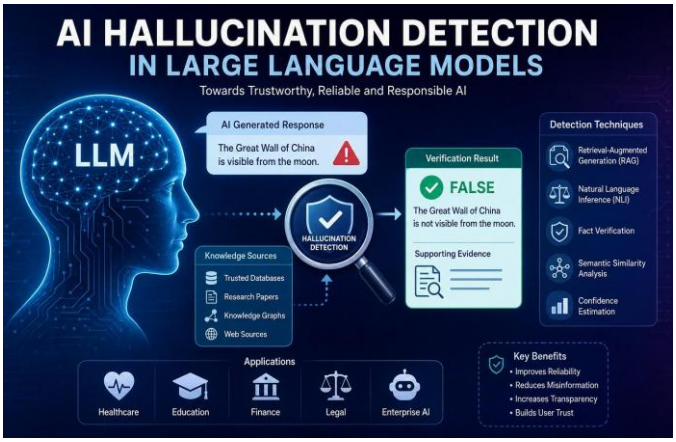
D. Context Hallucination

Context hallucination occurs when the generated response contradicts or ignores the information provided in the user's prompt or reference document. The AI introduces information that is unrelated or unsupported by the available context.

E. Instruction Hallucination

Instruction hallucination occurs when the model misunderstands or ignores user instructions, producing responses that fail to satisfy the intended task or constraints.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)



IV. Hallucination Detection Techniques

Several computational techniques have been developed to identify hallucinated information in AI-generated content. Each technique focuses on verifying the factual correctness and consistency of model outputs.

A. Retrieval-Based Detection

Retrieval-based detection compares AI-generated responses with trusted external knowledge sources such as scientific databases, digital libraries, official websites, and enterprise knowledge bases. If the retrieved evidence contradicts the generated content, the response is flagged as a potential hallucination.

B. Confidence Estimation

Modern language models can estimate confidence scores for generated responses. Low-confidence predictions often indicate uncertainty and a higher probability of hallucination. Confidence estimation helps prioritize responses requiring further verification.

C. Semantic Similarity Analysis

Semantic similarity techniques compare AI-generated text with reference documents using sentence embeddings and transformer-based language models. Low semantic similarity may indicate fabricated or irrelevant information.

D. Self-Consistency Verification

Self-consistency methods generate multiple responses to the same query and compare them. If significant differences exist among the generated answers, the response is considered less reliable.

V. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) reduces AI hallucinations by retrieving relevant information from trusted external sources before generating responses.

The retrieved evidence improves factual accuracy and reliability. RAG combines user queries with verified documents, making it widely used in healthcare, education, finance, legal services, and enterprise knowledge management for trustworthy AI applications.

VI. Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) improves transparency by explaining how AI models make decisions. In hallucination detection, XAI identifies why incorrect information is generated and highlights unsupported claims. Techniques such as LIME, SHAP, attention visualization, and feature importance analysis enhance user trust and support responsible AI deployment.

VII. Fact Verification Models

Fact verification models determine whether AI-generated content is supported by trusted evidence. They compare responses with reliable sources such as research papers and databases. Transformer models like BERT, RoBERTa, DeBERTa, and T5 classify statements as Supported, Contradicted, or Insufficient Evidence, improving the reliability of AI systems.

Comparison of Hallucination Detection Techniques

Technique	Method	Accuracy	Computational Cost	Common Applications
Retrieval-Augmented Generation (RAG)	External Knowledge Retrieval	Very High	Moderate	Enterprise AI, Healthcare
Natural Language Inference (NLI)	Semantic Reasoning	High	Moderate	Fact Verification
Knowledge Graph Verification	Graph-Based Validation	High	High	Scientific Knowledge Systems
Semantic Similarity	Embedding Comparison	Moderate	Low	Document Verification
Confidence Estimation	Probability Analysis	Moderate	Very Low	Chatbots
Explainable AI (XAI)	Model Interpretation	High	Moderate	Responsible AI
Self-Consistency	Multiple Response Comparison	High	Moderate	Question Answering

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

Large Language Models (LLMs)

Large Language Models (LLMs) are deep learning systems based on the transformer architecture. They are trained on vast amounts of text data to produce responses that resemble human language. Examples of these models include GPT, Gemini, Claude, Llama, and Mistral. While these models can generate coherent and fluent text, they may sometimes provide incorrect information. Therefore, detecting hallucinations—instances where the model creates false or misleading content—is crucial for improving the reliability of AI systems.

Natural Language Inference (NLI)

Natural Language Inference (NLI) is a method used to assess the relationship between a statement and a given context.

It helps determine whether a statement is supported by the evidence, contradicts it, or is unrelated.In the context of hallucination detection, NLI is used to check if the information generated by an AI model aligns with reliable sources.This process helps in enhancing the accuracy and trustworthiness of AI responses in areas like question answering and conversation systems.

Knowledge Graph Verification

Knowledge Graph Verification is a technique that checks the factual correctness of information generated by AI models.

It uses structured knowledge bases such as Wikidata and DBpedia to cross-reference the content produced by the model with established facts. By identifying discrepancies between the generated information and verified data, this method improves the factual accuracy of AI outputs. It is commonly applied in fields such as healthcare, legal services, scientific research, and enterprise knowledge management.

Challenges in Hallucination Detection

Hallucination detection faces challenges such as identifying subtle factual errors, verifying domain-specific information, supporting multilingual and multimodal AI, maintaining computational efficiency, and providing interpretable explanations. Continuous research aims to improve the accuracy, reliability, and transparency of AI-generated content.

Hallucination Evaluation Metrics

Hallucination detection models are evaluated using Accuracy, Precision, Recall, F1-Score, BERTScore, and FactScore.

These metrics measure classification performance, semantic similarity, and factual correctness, helping researchers assess and improve the reliability of AI-generated responses.

Applications of AI Hallucination Detection

AI hallucination detection has become an essential component in many real-world applications where accurate and trustworthy information is required.

A. Healthcare

B. Education

C. Financial Services

D. Legal Services

E. Scientific Research

F. Customer Support

Future Trends in AI Hallucination Detection

Research in hallucination detection continues to evolve rapidly with the advancement of generative AI technologies. Future systems are expected to integrate multiple verification techniques to achieve higher accuracy and transparency.

Emerging trends include real-time hallucination detection during text generation, multimodal hallucination detection for text, images, audio, and video, privacy-preserving verification using federated learning, and explainable fact verification systems. The integration of Retrieval-Augmented Generation (RAG), Knowledge Graphs, and Explainable Artificial Intelligence (XAI) is expected to significantly improve the reliability of future Large Language Models.

In addition, domain-specific hallucination detection models for healthcare, law, finance, and scientific research will become increasingly important as AI systems are deployed in safety-critical environments.

Future Trends in AI Hallucination Detection

Future AI hallucination detection will integrate advanced verification techniques for greater accuracy and transparency. Emerging trends include real-time detection, multimodal verification, federated learning, and explainable AI. Combining RAG, Knowledge Graphs, and XAI will improve the reliability of Large Language Models, particularly in healthcare, law, finance, and scientific research.

VIII. Conclusion

AI hallucination detection is essential for improving the reliability and trustworthiness of Large Language Models.

Techniques such as RAG, NLI, Knowledge Graph Verification, XAI, and Fact Verification effectively reduce misinformation. Future research will focus on real-time, explainable, and domain-specific solutions to support responsible AI across various applications.

REFERENCES

[1] Y. Ji, J. Lee, T. Frieske, et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 58, no. 2, pp. 1–38, 2025.

[2] OpenAI, "GPT-4 Technical Report," OpenAI, 2024.

[3] S. Rawte, A. Sheth, and A. Das, "A Survey of Hallucination in Large Foundation Models," arXiv preprint, arXiv:2309.05922, updated 2024.

[4] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, updated editions and follow-up studies, 2024.

[5] T. Schuster, A. Fisch, J. Weston, and D. Chen, "Consistent Retrieval for Faithful Knowledge-Intensive Generation," EMNLP, 2024.

[6] J. Guo, X. Zhang, and Y. Liu, "Explainable Artificial Intelligence for Trustworthy Large Language Models: A Survey," IEEE Access, vol. 13, 2025.

[7] H. Li, M. Wang, and K. Chen, "Fact Verification for Large Language Models Using Natural Language Inference," IEEE Access, vol. 13, 2025.

[8] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems, updated applications, 2024.

[9] Z. Huang, Y. Qin, and X. Li, "Knowledge Graph-Based Hallucination Detection in Large Language Models," IEEE Transactions on Artificial Intelligence, vol. 6, no. 1, pp. 45–58, 2026.

[10] X. Zhang, L. Chen, and Y. Zhao, "Recent Advances in Hallucination Detection for Large Language Models: A Comprehensive Review," IEEE Access, vol. 14, 2026.