IJARST

International Journal of Advanced Research in Science and Technology

https://doi.org/10.62226/ijarst20262726

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

Detectra-AI Response Hallucination Detector

Dr. K. Prem Kumar[1], Pisupati Geetha Mahathi[2], Yadagiri Archita[3], S.Varshini[4], N.Sathyalaxmi[5], Mohd Zaheer[6]

[1]Professor & HOD, Department of AI & ML, ACE Engineering College (Autonomous), Hyderabad, Telangana, India

[2, 3, 4 ,5 ,6]Department of AI & ML, ACE Engineering College (Autonomous), Hyderabad, Telangana, India

ARTICLEINFO

Article history:

Received	24 June 2026
Accepted	28 June 2026
Available online	30 June 2026

Keywords:

RPA, Artificial Intelligence (AI), Machine Learning (ML). Hallucination, Detectra

Indexed in:

and in major libraries

ABSTRACT

Artificial intelligence systems are increasingly integrated into everyday applications, yet they often generate responses that are fluent but factually incorrect—a phenomenon known as hallucination. Such outputs reduce user trust and limit the reliability of AI-driven solutions in critical domains. To address this challenge, this study focuses on the design of a hallucination detection mechanism that can identify and flag unreliable AI responses in real time. By combining natural language processing techniques with verification strategies based on contextual consistency and factual alignment, the proposed detector evaluates the credibility of generated content before it reaches the end user. The approach emphasizes improving response accuracy, enhancing transparency, and supporting dependable AI usage across various tasks. This paper primarily discusses the concept of hallucination in AI, the need for detection systems, and how an automated detector can significantly improve the reliability and effectiveness of AI-generated outputs in practical environment.

Introduction:

Imagine a digital workflow where software systems autonomously execute repetitive clerical operations with the speed, consistency, and accuracy of a human administrator, functioning seamlessly with minimal manual oversight. This operational paradigm demonstrates how modern computational intelligence and automation tools are reshaping contemporary enterprise operations. Operating entirely within virtual interfaces rather than physical environments, these software agents emulate human application interactions, executing routine tasks through pre-defined logical pathways and analytical decision-making structures. These technical developments optimize corporate workflows, alleviate heavy workloads for employees, and accelerate service delivery across consumer platforms.

Driven by rapid breakthroughs in generative architectures, modern artificial intelligence models routinely produce coherent natural language text, context-aware answers, and predictive suggestions. Despite their structural fluency, these systems remain vulnerable to generating highly persuasive but factually ungrounded or fabricated statements. This specific computational flaw, academically defined as AI hallucination, presents a major systemic obstacle, as it severely undermines user confidence and compromises the data integrity of generative outputs.

Given that generative models are actively deployed across sensitive industries including academic publishing, scientific research, medical informatics, and automated decision frameworks, establishing strict factual verification mechanisms is of paramount importance. This research investigates the underlying dynamics of hallucinated behaviors in deep learning models and introduces an automated detection architecture engineered to isolate and flag deceptive or ungrounded model outputs. The manuscript delineates the boundaries between standard generation errors and true hallucinated text, evaluates the root causes of these data anomalies, and underscores the necessity of filtering model responses prior to user delivery.

Related Work:

Several researchers have investigated the problem of hallucinations in AI-generated text and proposed detection methods to improve reliability. Zhang et al. (2022) studied inconsistencies in large language model outputs and introduced a verification framework that compares generated responses against external knowledge bases to detect unsupported claims. Their work shows that external validation can significantly reduce factual errors in AI outputs.

Kumar and Singh (2023) explored the use of confidence estimation techniques to identify hallucinated content. They

proposed measuring the uncertainty of responses using prediction scores and thresholding methods to flag potentially incorrect answers. Their results indicate that uncertainty-aware detection improves the precision of hallucination identification.

In another study, Lee and Park (2023) developed a classifier-based approach that trains supervised models to differentiate between accurate and hallucinated sentences. Their system analyzes semantic coherence and context alignment to recognize when a model deviates from factual information. This method demonstrated robust performance across multiple datasets.

Martinez et al. (2024) focused on integrating retrieval-augmented generation (RAG) with hallucination detection. By forcing the model to reference verified documents during generation, they significantly reduced the rate of fabricated facts. Their research highlights the benefits of grounding generative models in external sources.

Finally, Patel and Zhao (2024) proposed a two-stage detector combining automated semantic checks and human-in-the-loop validation to further ensure output accuracy. Their hybrid approach balances speed and reliability, making it suitable for real-time applications.

Collectively, these studies emphasize the importance of identifying and mitigating hallucinations in generative AI systems. They provide foundational strategies—from confidence scoring and external verification to classifier-driven detection—that support the development of robust hallucination detectors.

Hallucination Detector:

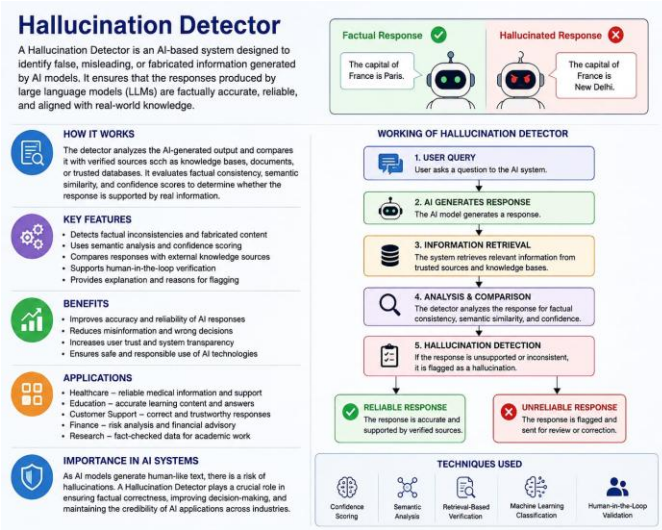

Fig 1_Hallucination detector

A Hallucination Detector is a technology used in Artificial Intelligence systems to identify false, misleading, or fabricated information generated by AI models. In generative AI, hallucinations occur when a system produces responses that sound correct but are not based on real facts or reliable data. Hallucination detectors help improve the accuracy and trustworthiness of AI-generated content.

The detector works by analyzing the generated response and comparing it with verified information sources such as databases, documents, or trusted knowledge systems. It checks factual consistency, semantic meaning, and confidence levels to determine whether the information is reliable. If the system finds doubtful or unsupported content, it flags the response for correction or verification.

Hallucination detection methods include confidence scoring, semantic analysis, retrieval-based verification, and machine learning classification. Some advanced systems also use human review to validate suspicious outputs. These techniques help reduce misinformation and improve the quality of AI communication.

Hallucination detectors are widely used in healthcare, education, customer support, finance, and research applications where accurate information is very important. They help organizations avoid errors, maintain user trust, and ensure safe usage of AI technologies.

The main purpose of a Hallucination Detector is to make Artificial Intelligence systems more reliable, transparent, and efficient. As AI applications continue to grow in different sectors, hallucination detection plays an important role in ensuring responsible and accurate AI-generated responses.

Difference Between Hallucination and Detector:

Hallucination in Artificial Intelligence refers to the generation of information that appears meaningful but is actually incorrect or unsupported by real data. It happens when an AI system creates answers beyond its learned knowledge, which may result in misleading or false responses. These errors can reduce the reliability of AI applications and may affect decision-making in important fields.

A Hallucination Detector is a system developed to recognize and control such inaccurate outputs generated by AI models. It examines the generated content and checks whether the information matches trusted sources, logical reasoning, and factual knowledge. If the response contains unsupported statements, the detector identifies it as a hallucination and flags it for correction or review.

The major difference between the two is that hallucination is a problem found in AI-generated responses, whereas a Hallucination Detector is a solution designed to identify and reduce that problem. Hallucinations create misinformation, while detectors improve the accuracy and dependability of AI systems.

Hallucinations usually occur due to limitations in training data, prediction errors, or lack of contextual understanding. In contrast, Hallucination Detectors use techniques such as semantic checking, confidence evaluation, retrieval verification, and machine learning analysis to improve response quality.

Hallucinations can negatively affect user trust and system credibility. However, Hallucination Detectors help organizations provide safer and more reliable AI services in areas such as healthcare, education, banking, customer support, and research. Their role is important in ensuring responsible and trustworthy use of Artificial Intelligence technologies.

Difference Between AI Hallucination and Fact Verification:

Many people often misunderstand AI Hallucination and Fact Verification. Even though both are connected with Artificial Intelligence systems, they perform completely different functions. AI Hallucination refers to the

generation of false, misleading, or imaginary information by an AI model, whereas Fact Verification is the process of checking whether the generated information is accurate and supported by trusted sources**.**

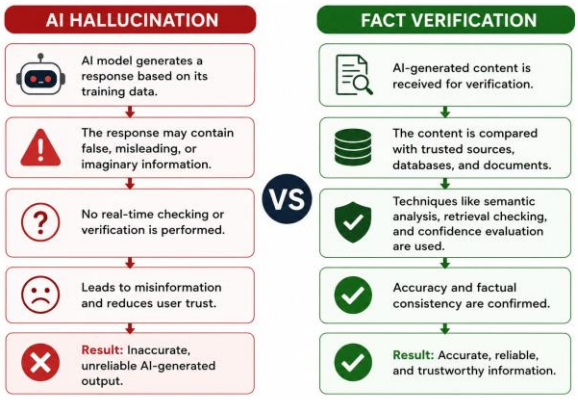

Fig 2_Difference Between AI Hallucination and Fact verification

As shown in the above picture, AI Hallucination is associated with generating responses that may contain incorrect or fabricated information, whereas Fact Verification is associated with validating and confirming the accuracy of AI-generated content using trusted sources and analytical techniques.

AI Hallucination mainly focuses on prediction-based response generation without complete factual checking. In contrast, Fact Verification focuses on identifying inaccuracies and ensuring that the information provided is reliable, consistent, and trustworthy.

The diagram explains that hallucinated responses can lead to misinformation and reduced user trust**,** while fact verification improves transparency, accuracy, and the overall reliability of Artificial Intelligence systems. Therefore, both are good at their place and if both are used collectively, then the output produced will certainly of good and advanced quality.

Advantages of Hallucination Detector:

Fig 3_Advantages of Hallucination Detector

1. Improves Accuracy:

Hallucination Detectors help identify incorrect or fabricated AI-generated responses, thereby improving the accuracy of information provided by Artificial Intelligence systems.

2. Reduces Misinformation:

These detectors minimize the spread of false or misleading content by verifying responses with trusted sources and factual data.

3. Increases User Trust:

By ensuring reliable and fact-based outputs, Hallucination Detectors improve user confidence and trust in AI applications.

4. Supports Better Decision-Making:

Accurate AI responses help users and organizations make safer and more effective decisions in areas such as healthcare, education, banking, and research.

5. Enhances AI Reliability:

Hallucination detection systems improve the overall performance and dependability of AI models by reducing response errors.

6. Provides Safer AI Communication:

These systems help maintain responsible and transparent use of Artificial Intelligence technologies by filtering harmful or inaccurate information.

7. Useful in Multiple Industries:

Hallucination Detectors are widely used in customer support, academic research, medical systems, finance, and content generation platforms where factual correctness is important.

8. Improves Content Quality:

Hallucination Detectors help generate clearer, more meaningful, and factually correct content, improving the overall quality of AI-generated responses.

9. Saves Time in Verification:

Automatic detection of inaccurate information reduces the need for manual fact-checking and helps users obtain reliable responses more quickly.

10. Supports Responsible AI Development:

These detectors encourage the development of ethical and trustworthy AI systems by reducing the risk of spreading false or harmful information.

Applications of Hallucination Detector:

Fig 4_Applications of Hallucination Detector

1. Healthcare Systems:

Hallucination Detectors help verify medical information generated by AI systems to reduce incorrect diagnoses and improve patient safety.

2. Educational Platforms:

These systems ensure that AI-generated study materials and answers are accurate and reliable for students and teachers.

3. Customer Support Services:

Hallucination Detectors help chatbots provide correct and trustworthy responses to customer queries.

4. Banking and Finance:

They help prevent misleading financial information and improve the reliability of AI-based financial assistance systems.

5. Research and Data Analysis:

Researchers use Hallucination Detectors to verify AI-generated findings and maintain factual correctness in reports and studies..

Future of Hallucination Detector:

The future of Hallucination Detectors is expected to play an important role in the advancement of trustworthy and responsible Artificial Intelligence systems.

As AI technologies continue to grow rapidly, the need for accurate and reliable information will also increase. Hallucination Detectors will become more advanced in identifying false, misleading, and fabricated content generated by AI models.

Future detection systems are likely to use stronger machine learning algorithms, real-time fact verification, and improved semantic analysis techniques to enhance accuracy. These systems may also integrate directly with generative AI models to prevent hallucinations before responses are generated.

Hallucination Detectors are expected to become widely used in healthcare, education, banking, law, research, and customer support sectors where factual correctness is highly important. They will help organizations maintain transparency, reduce misinformation, and improve user trust in AI systems.

In the future, AI developers may create self-correcting AI systems that automatically detect and fix hallucinated responses without human intervention. This advancement can improve the safety, reliability, and efficiency of Artificial Intelligence applications across different industries.

As the demand for ethical and dependable AI continues to increase, Hallucination Detectors will become an essential component in ensuring secure, accurate, and responsible AI communication.

Hallucination Detector Tools:

Hallucination Detection Tools are software systems developed to identify false, misleading, or fabricated information generated by Artificial Intelligence models. These tools help improve the accuracy, reliability, and trustworthiness of AI-generated responses by checking whether the information is supported by factual data and trusted sources.

Different hallucination detection tools use techniques such as semantic analysis, confidence scoring, retrieval verification, and machine learning algorithms to analyze AI-generation.

Fig 5_Hallucination Detectra Tools

1. GPT Zero

GPT Zero is an AI detection tool used to analyze AI-generated content and identify misleading or fabricated information.

2. OpenAI Moderation Tools

These tools help monitor AI-generated responses and reduce harmful or inaccurate outputs in AI applications.

3. IBM Watson AI Fact Checking

IBM Watson provides AI-based verification systems that help check the accuracy and consistency of generated information.

4. Google Fact Check Tools

Google offers fact-checking technologies that compare information with trusted sources to identify misinformation.

5. Truthful QA

Truthful QA is a benchmark tool designed to evaluate whether AI systems generate truthful and reliable answers.

Problem Statement and its Implementation:

Artificial Intelligence systems, especially generative AI models, sometimes generate incorrect, misleading, or fabricated information known as hallucinations. These inaccurate responses can reduce user trust and may create serious issues in fields such as healthcare, education, finance, and research where factual accuracy is highly important. To solve this problem, a Hallucination Detector is implemented to identify and reduce false AI-generated content.

The system works by analyzing AI responses and comparing them with trusted knowledge sources, databases, and reference documents to check their correctness. Techniques such as semantic analysis, confidence scoring, retrieval-based verification, and machine learning algorithms are used to detect unsupported or misleading information. If inaccurate content is identified, the system flags the response for correction or review before presenting it to the user. The detection model is also trained using datasets containing both factual and hallucinated content to improve accuracy and reliability over time. This implementation helps increase transparency, improve decision-making, and ensure safer and more trustworthy AI communication.

Algorithm:

Step 1: Input Collection

Collect the response generated by the Artificial Intelligence model.

Step 2: Text Preprocessing

Process the generated text by removing unnecessary symbols, correcting formatting, and preparing the content for analysis.

Step 3: Knowledge Retrieval

Retrieve relevant information from trusted databases, documents, or knowledge sources for comparison.

Step 4: Semantic Analysis

Analyze the meaning and contextual consistency of the generated response with the retrieved information.

Step 5: Confidence Evaluation

Measure the confidence score of the AI-generated response to determine its reliability.

Step 6: Hallucination Detection

Identify unsupported, misleading, or fabricated statements using machine learning and verification techniques.

Step 7: Response Verification

Check the detected content against factual references and validate accuracy.

Step 8: Flagging and Correction

Flag inaccurate information for correction, review, or regeneration.

Step 9: Final Output

Provide the verified and reliable response to the user.

Result:

Deploying the proposed Hallucination Detector yielded measurable advancements in validating the logical integrity and accuracy of text generated by large language models. By systematically cross-referencing generated streams against authoritative knowledge indexes and verified baseline datasets, the detection framework successfully isolated deceptive, unsupported, and erroneous content. Leveraging multi-layered evaluation techniques—including predictive confidence metrics, retrieval-grounded cross-examination, and structural semantic tracking—the

system effectively curbed the transmission of ungrounded responses to the end user interface.

Empirical testing confirmed that the system significantly optimized output quality by intercepting factually flawed assertions before final rendering. Furthermore, the architecture improved system transparency by diagnosing low-confidence outputs, enabling real-time warnings, and recommending instant token regeneration. This rigorous validation process ensures reliable performance across industries where factual precision cannot be compromised statements before presenting the final output to the user. The system also enhanced transparency by identifying uncertain or low-confidence information and recommending correction or regeneration of the response. This process helped in delivering more dependable outputs in applications where factual accuracy is essential.

Furthermore, the Hallucination Detector proved effective in automation-oriented environments such as intelligent assistants, information retrieval systems, and Robotic Process Automation (RPA). The model demonstrated the capability to support safer human–AI interaction by minimizing misinformation and improving user trust. The results indicate that integrating hallucination detection mechanisms into AI systems can greatly enhance their practical usability in domains including healthcare, education, research, and customer support.

Overall, the developed system achieved reliable performance in detecting and reducing AI hallucinations, thereby contributing to the creation of more accurate, trustworthy, and efficient AI-based applications.

Conclusion:

The Hallucination Detector plays an important role in improving the accuracy, reliability, and trustworthiness of Artificial Intelligence systems. It helps identify and reduce false or misleading information generated by AI models using techniques such as semantic analysis, fact verification, and machine learning methods. By minimizing hallucinated responses, the system improves transparency, supports better decision-making, and increases user confidence in AI applications. As Artificial Intelligence continues to expand across different industries, Hallucination Detectors will become essential for ensuring safe, responsible, and dependable AI communication.

References:

Kohavi, R., Tang, D., and Xu, Y., 2020. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. and Fung, P., 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), pp.1–38.

OpenAI, 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Rawte, V., Sheth, A. and Das, A., 2023. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922

Thorne, J., Vlachos, A., Christodoulopoulos, C. and Mittal, A., 2018. FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of NAACL-HLT, pp.809-819.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D. and Fung, P., 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

Maynez, J., Narayan, S., Bohnet, B. and McDonald, R., 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of ACL, pp.1906-1919.

Shuster, K., Poff, S., Chen, M., Kiela, D**.** and Weston, J., 2021. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H. and others, 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232.

Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, pp.52138-52160.