RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 06, Issue 12 December

Check for updates

http://dx.doi.org/10.47392/IRJASH.2024.050

AI Based Hallucination Detector

Ms. G. Illakiya[1], G.V. Jeeshitha[2], R. Manjushree[3], P. Harshini[4]

[1]Assistant Professor, Department of Artificial intelligence and Data science, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India.

[2,3,4]Department of Artificial intelligence and Data science, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India.

Emails: *msr212003@gmail.com[1]*

Article history

Received: 07 November 2024
Accepted: 26 November 2024
Published: 11 December 2024

Keywords:

AI learning algorithms, AI response accuracy, API integration, critical error alerts, error detection, real-time validation, transparency, user feedback, IEEE

Abstract

The main goal of this project is to create an advanced system that can accept questions submitted by users, produce AI-generated answers, and guarantee their accuracy by using an integrated validation method. This involves connecting to external web APIs that have access to trustworthy and authoritative sources, allowing the system to compare AI-generated responses with verified factual data in real time. If the AI-generated answer is accurate, the system will show a confirmation, providing users with assurance of its reliability. However, if the answer is incorrect, the system will flag the error and present the accurate response, addressing the issue of AI generating believable but factually inaccurate answers. In addition, the system records inaccurate responses to detect recurring error trends, which helps to enhance and refine the AI model over time. It also includes an interactive explanation tool that allows users to comprehend the validation process, promoting transparency in decision- making. To increase user involvement, the system can provide information about the origin of the correct answer and offer insights into differences between the AI-generated answers and the correct ones. Additionally, the system will have real-time alerts for critical errors, promptly notifying users when high-risk or sensitive topics are involved. The system's overall accuracy will be evaluated through a periodic review mechanism, which will offer feedback on performance enhancements. Additionally, user feedback will be incorporated into the system to continuously improve it and adapt to changing information sources. Moreover, the system will utilize AI-based learning algorithms to anticipate and prevent potential errors, thus enhancing response quality over time. Ultimately, the project's goal is to establish a reliable and user-friendly AI environment that fosters trust through real-time verification, transparency, continuous enhancement, and minimized risks of incorrect AI outputs.

1. Introduction

The AI Hallucination Detector study introduces a novel approach that aims to tackle a significant AI challenge: guaranteeing the precision and dependability of responses generated by AI. The way the system works is that it takes questions from users, uses AI to generate answers, and then uses a

strong real-time verification procedure to confirm he answers. In order to do this validation, the system is connected to external web APIs that provide access to reliable and authoritative sources. This enables the system to cross-check the results produced by AI with confirmed factual data. Users can feel reassured by the system's confirmation when the AI generates an accurate response. On the other hand, the system automatically shows the proper information and highlights the inaccuracy if the response is untrue or deceptive. This method addresses the problem of artificial intelligence (AI) hallucinations, in which an AI produces responses that seem plausible but are factually inaccurate. In addition, the system logs inaccurate results in order to spot reoccurring error patterns, which helps the AI model get progressively better over time. The system includes an interactive explanation tool that enables users to comprehend the validation procedure and the reasoning behind adjustments, all in an effort to increase transparency. It also clarifies the discrepancies between the right response and the AI-generated one and gives information about the verified information's source. Furthermore, severe mistakes cause real-time notifications to be sent, guaranteeing that users are informed as soon as possible while handling sensitive or high-risk topics. Through user input and monthly performance assessments, which provide insights for future changes, the system is always evolving. Moreover, it incorporates AI- based learning algorithms to foresee and avert certain mistakes, thereby enhancing the quality of its responses overall. In the end, the AI Hallucination Detector wants to minimize the chances of inaccurate outputs while guaranteeing real-time verification, transparency, and ongoing improvement to build a dependable, user- friendly environment that promotes trust in AI. [1-10]

2. Related Work

The AI Hallucination Detector is the result of advanced research and technologies that are aimed at enhancing the precision and dependability of outputs produced by artificial intelligence. Natural language processing (NLP) has made significant strides, especially with large language models (LLMs) such as GPT, BERT, and T5. These developments have revolutionized AI's capacity to produce prose that is human-like. Nevertheless, these models are prone to producing reactions that seem plausible but are untrue or unimportant in terms of facts, or hallucinations. Numerous investigations into the reasons of these hallucinations have pointed to biases, limited training data, and the models' incapacity to obtain verified, real-time information. Prior research on mitigating AI hallucinations has concentrated on improving model architectures and training methods, such as using reinforcement learning from human feedback (RLHF) or fine-tuning with more accurate and varied datasets. In order to increase answer accuracy, several strategies have also included retrieval- based processes, in which AI systems go through external databases or knowledge graphs. For example, AI can search the web in real time for factual information to increase accuracy. This is made possible by OpenAI's WebGPT and Google's Search-augmented language models. Additionally, tools such as Truth-GPT and Fact-Checking APIs try to validate outputs by cross-referencing them with reliable sources. Furthermore, explainable AI (XAI) research has highlighted the necessity of transparent decision-making procedures that allow users to comprehend how AI reaches particular judgments. By incorporating real-time validation through external APIs, an interactive explanation tool, and continuous error monitoring and correction, the AI Hallucination Detector expands on these strategies and develops a more resilient system that tackles the issues of accuracy and transparency in AI-generated responses. The technology is positioned as a breakthrough in preventing AI hallucinations and boosting user confidence in AI systems because to this combination of qualities. [11-15]

3. System Architecture

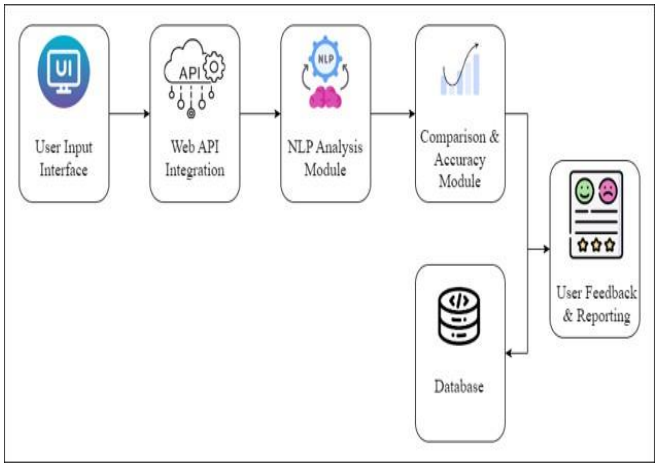

Figure 1 System Architecture

3.1. User Input Interface

A User Input Interface enables interaction among the user and the software by enabling the submission of data for analysis, including text, documents, and queries. This interface might be used in academic environments to receive research requests, scientific publications, or abstracts while maintaining compatibility with various formats such as PDFs and plain text. The user interface has been developed with ease of use in mind, with unambiguous instructions and prompt feedback to guarantee a seamless experience. It records pertinent information with the input to make the analysis process easier in the future.

3.2. Web API Integration

By connecting the system to other data sources, the Web Application Programming Interface Integration module enables it to retrieve pertinent data from external databases, research papers, and citation data, among other sources. For instance, to get references and citation metrics associated with the user's input, this module may connect with scholarly repositories such as IEEE Xplore, Google Scholar, or ArXiv. The system enhances the input data and offers a more comprehensive context for analysis by utilizing APIs. Through effective, safe API calls, this module makes sure that internal elements and external information repositories collaborate seamlessly.

3.3. NLP (Natural Language Processing) Analysis Module

Advanced Natural Language Processing methods are used by the NLP Analysis Module, which serves as the central engine for processing and interpreting the provided text. To extract important information from the input, it carries out operations including tokenization, sentiment analysis, entity identification, and topic modeling. This module provides citations, keywords, and research subjects that are pertinent to academic study. It also interprets sophisticated scientific jargon and technical terminology to generate structured result that can be utilized to perform additional comparison and reporting. Either bespoke NLP algorithms or pre-trained models are used to efficiently handle language particular to a certain area. Figure 1 shows System Architecture.

3.4. Comparison & Accuracy Module

Following text processing, the extracted data is assessed by the Comparing & Accuracy Module against pre-established standards or external data. This module validates the system's outputs against well- known datasets or scholarly benchmarks to make sure that tasks like sentiment analysis, keyword extraction, and topic modeling adhere to accuracy standards. To make sure that it complies with industry standards, it may, for instance, compare scientific words or citation patterns. The accuracy scores and reliability insights provided by the module aid users in appreciating the caliber and applicability of the analysis. [16-20]

3.5. Database

The database serves as the system's core repository for user input, processed information, analytical findings, and system logs. It is essential for maintaining historical information, which is necessary for benchmarking and gradually enhancing the accuracy of the system. The database allows the system to access and reuse user comments, NLP results, and analysis reports from the past for analysis in the future. The database also keeps track of all procedure step records for auditing and troubleshooting. A huge amount of academic data, such as several research publications, may be conveniently stored and queried thanks to its scalability.

3.6. User Feedback & Reporting

The User Feedback & Reporting module delivers detailed reports to users based on the analysis performed by the system. These reports might include key insights such as identified keywords, citation analyses, relevance scores, and statistical summaries. Additionally, this module collects feedback from users regarding the accuracy and relevance of the results, which can then be used to refine and enhance the system's future performance. Feedback might include corrections to identified topics or keywords, or general user satisfaction ratings. The feedback loop is critical to improving the machine learning models used in the system, ensuring better accuracy with each iteration.

4. Methodology Used

4.1. User Interface Module

The primary source to reach out for users is the User Interface (UI) Module, which offers a simple and intuitive platform for asking queries and receiving responses from AI. The user interface has been designed to accept a variety of input formats.Users can view the verified results by typing in queries.

When discrepancies are found, one of the most crucial aspects of the user interface is the comprehensive justifications that are supplied. The user interface unravels the logic involved in the verification process, as opposed to merely pointing out that an AI-generated response is incorrect. By clearly contrasting the AI response with the accurate response retrieved from dependable sources, it demonstrates to users where the AI made mistakes. This could entail highlighting specific phrases or words, highlighting information from credible sources, or proposing substitute acceptable answers. In general, the UI Module is intended to be more than just an input and output device; instead it aims to foster an elevated degree of involvement, which makes it a crucial component of the AI Hallucination Detector system.

4.2. Backend Server

The central component of the system is the Backend Server, and it's in charge of managing the data flow between the user interface and the many functional modules. In order to retrieve accurate responses, it interacts with the verification module, handles requests from the user interface, and processes them. In addition, the administration of databases, user interactions, AI responds, and validation findings are managed by the backend for further evaluation. This part makes sure that data is processed effectively and that the infrastructure is scalable enough to manage several user requests at once. With the goal to provide a seamless and responsive experience, the server additionally handles essential tasks including managing user sessions and caching frequently requested data. likewise, load balancing is supported by the backend, which divides incoming traffic across several servers to avert overload and assure steady performance. In order to safeguard sensitive user data and guarantee compliance with privacy laws, it also includes strong security mechanisms like data sanitisation, encryption, and authentication.

4.3. Module API End Point

The system's many components and external web APIs are connected by the API End Points. This module is in charge of retrieving validated responses from trustworthy sources, including databases, knowledge bases, and reputable web repositories. In order to guarantee that the data retrieved is correct, current, and pertinent to the user's query, the API End Point controls communication between the backend server and outside data sources. It additionally handles care of formatting and data normalisation, making sure that incoming data and AI-generated responses can be easily compared. This module's flexibility resides in its ability to integrate multiple APIs that address various knowledge domains, assuring the system's thorough validation. In addition, the module allows for the caching of data from external APIs, accelerates up response times for frequently requested information.

4.4. Module Verification

The AI Hallucination Detector's main component is the Verification Module. It contrasts the accurate response obtained from external databases or APIs with the AI-generated response. This module compares the input using sophisticated algorithms that evaluate both surface-level similarities and underlying contextual meaning to make sure the right response makes semantic sense. Upon discovering disparities, the module signals them and forwards feedback to the backend and user interface for recording purposes. Additionally, it has a score system that assigns a value to each correctness of answer that is created by AI, giving more detailed information about situations where the answer may be partially correct but is borderline. The module additionally includes customisable verification rules, that enable system administrators to up certain validation guidelines according to the kind of query or domain, providing a customised verification procedure for various use cases. Additionally, it possesses real-time monitoring features that allow the system to continuously evaluate and report on the correctness of AI responses. [21-24]

4.5. Logic Module

The system's entire decision-making process is coordinated with the Logic Module. It determines how to handle incomplete matches, when to start the verification process, and what to do if the right answer cannot be obtained. This module includes rules to handle exceptions, ambiguities, or ambiguous inputs in order to maintain the system's flexibility and adaptability. The Logic Module is also in charge of judging the degree of disparities, which enables the system to distinguish between small problems (like a poorly phrased answer) and big hallucinations that could seriously mislead the user. By maximising future interactions for speed

and accuracy and learning from feedback from users as well as previous validation results, it also contributes to increasing system efficiency. Additionally, it interfaces with the analytics engine of the system to make decisions about the most effective use of computational resources and the prioritisation of verification tasks based on past performance indicators and real-time data.

5. Literature Review

In an effort to increase the accuracy of AI-generated content, the development of AI- based hallucination detectors touches on a range of areas, including machine learning, natural language processing (NLP), and AI ethics. A. Maynez's 2020 study on artificial intelligence (AI) models' propensity to produce hallucinations— erroneous or non-factual information that seems plausible— was disclosed. The study highlights how crucial it is to discern between sincere and delusional responses in order to guarantee AI dependability, particularly in crucial applications like translations for the legal and medical fields. Continuing from this, J. Goyal's 2022 work explores hallucinations in large language models (LLMs), distinguishing between two kind of hallucinations: intrinsic (wrong interpretations of the input data) and extrinsic (produced data without any basis in the input). Goyal suggested techniques that use cross-validation with outside sources of data to identify and minimise these hallucinations. Similar to this, M. Filippova's research from 2021 analyses how AI models' attention processes can be used to identify hallucinations during text summarisation. It highlights how differences in attention weights can identify hallucinated content at an early stage of production. Cross-referencing AI-generated responses with knowledge bases is an early technique for identifying hallucinations, first proposed in V. Gabriel's 2019 work on fact-checking algorithms. According to the study, employing structured data sources such as knowledge graphs may effectively decrease the incidence of hallucinations. Y. Li's 2021 study, which went beyond text-based systems, focused on multi- modal hallucination detection in a related study. Li investigated hallucination detection using deep learning algorithms that map input data across modalities in systems that combine text, auditory, and visual inputs, including virtual assistants or autonomous agents. The importance of preventing biases from impacting hallucination detection systems is also highlighted in B. Mittelstadt's 2016 paper on fairness in AI. Failing to accomplish so could result in the persistence of inaccurate information or systemic biases in AI- generated material. Last but not least, D. Lee's 2023 research on self-regulating AI systems highlighted a novel idea in which AI models are able to independently detect, rectify, and learn from their own hallucinations, becoming better over time without the need for outside assistance. This work presents a potential path forward for hallucination detectors, as AI systems grow more robust and self-sufficient in controlling the accuracy and quality of their outputs.

6. Source Code Sample

```
import streamlit as st
from transformers import
AutoModelForCausalLM, AutoTokenizer
import torch
import google.generativeai as genai import os
import wikipedia
# Load API keys securely
genai.configure(api_key=os.getenv("GOO
GLE_GEMINI_KEY"))
# Load the model and tokenizer
tokenizer =
AutoTokenizer.from_pretrained("microsoft
/DialoGPT-medium")
model =
AutoModelForCausalLM.from_pretrained(
"microsoft/DialoGPT-medium")

# Configure GoogleGeminifor
hallucination detection
generation_config = {
"temperature": 1, # Controls response
creativity
"top_p": 0.95, # Limits randomness
"top_k": 64, # Selects the top 64 candidates at
each step
"max_output_tokens": 8192 # Allows extended
responses
}
# Define the Gemini generative model for
comparison
modell = genai.GenerativeModel(
model_name="gemini-1.5-flash",
generation_config=generation_config,
)
```

```
# Initialize Streamlit UI with user input and Wikipedia reference
st.title("LLM Hallucination Checker")
user_input = st.text_input("Enter your question:", "")
if st.button("Send") and user_input:
# Generate response with DialoGPT
input_ids = tokenizer.encode(user_input + tokenizer.eos_token, return_tensors='pt')
chat_history_ids = model.generate(input_ids, max_length=1000, pad_token_id=tokenizer.eos_token_id)
response = tokenizer.decode(chat_history_ids[:, input_ids.shape[-1]:][0], skip_special_tokens=True)
# Retrieve factual reference from Wikipedia
wikipedia_result = wikipedia.summary(user_input)
# Define prompt for Gemini hallucination validation
prompt = (
f"Compare the following answers for factual accuracy:\n"
f"LLM Answer: {response}\n"
f"Real Context (Wikipedia): {wikipedia_result}\n"
f"Indicate in one line if the LLM answer aligns with the real context."
)
# Get Gemini's validation response
validation_response = modell.generate_content(prompt)
st.info(validation_response.text):
#end of code
```

7. Result & Discussion

One major challenge in AI is ensuring the accuracy of responses given by AI, to which the AI Hallucination Detector offers a vital solution. Despite the impressive abilities of models like as the GPT, there is a significant issue with these models' tendency to provide outputs that are factually incorrect or deceptive, or induce hallucinations, particularly in sensitive fields like healthcare and law. By integrating third-party applications that verify responses with credible sources, the system reduces the likelihood of these hallucinations and ensures users receive accurate information. One of the system's key benefits is its capacity to track errors to improve the algorithm's accuracy over time, as well as swiftly identify and show accurate information. Transparency is increased and trust among consumers in AI decision-making is encouraged with an interactive explanations tool that helps users understand how validation functions and why particular outputs are tagged. However, there are still problems, such as depending excessively on the precision and breadth of other sources. If these sources are limited or selective, a system's validation process can still result in errors. Furthermore, even if the system's feedback system is advantageous for continuous improvement, regulating user feedback to avoid the beginning of biases or inaccuracy is vital. In addition, real-time alerts for significant issues must be handled carefully to avoid miscommunications or unjustified fear in delicate areas. In conclusion, the AI Hallucination Detector offers a viable way to reduce AI hallucinations by ensuring transparency and real-time verification. Though it has challenges in handling data and feedback sources, it establishes a solid foundation for upcoming developments in the creation of trustworthy, accurate AI systems. Figure 2 shows Output Image- Negative, Figure 3 Output, shows Figure 4 shows Output – Positive, Figure 5 shows Detector Result.

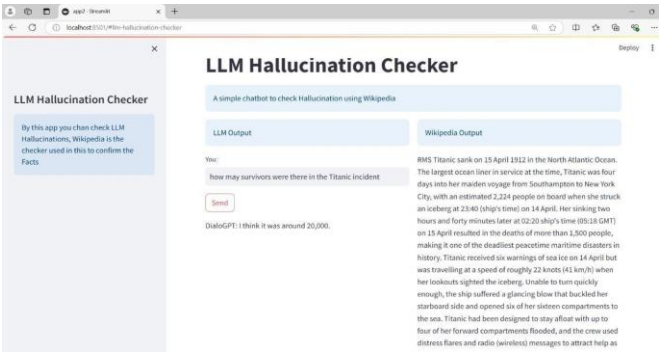

Figure 2 Output Image- Negative

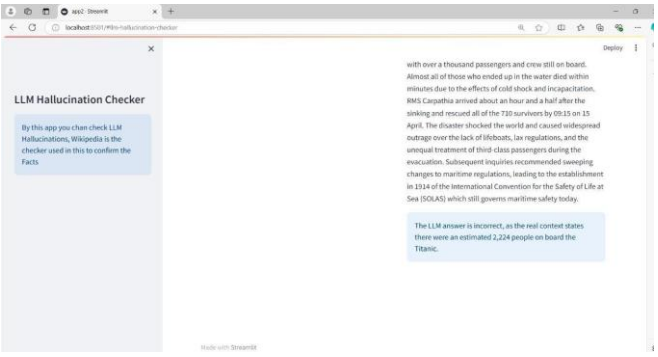

Figure 3 Output

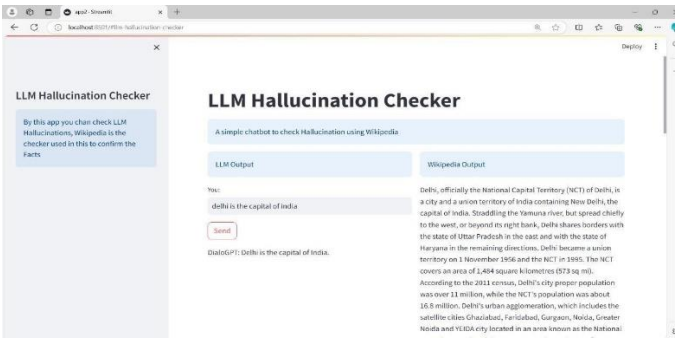

Figure 4 Output – Positive

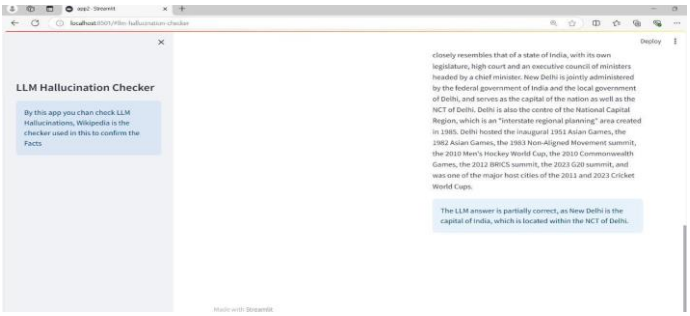

Figure 5 Detector Result

Conclusion

The AI-Based Hallucination Detector is a noteworthy development in the area of artificial intelligence that allays growing concerns about the accuracy and reliability of responses generated by AI. As chatbots to communicate virtual assistants, content production, and decision-making systems utilize artificial intelligence (AI) increasingly frequently, it is imperative to ensure that the data these models generate is accurate and genuine. In specifically, large language models (LLMs) are artificial intelligence (AI) systems that often provide outputs that appear credible but sporadically contain inaccurate or misleading information—a phenomena referred to as "hallucinations." There is a lot of risk associated with these hallucinations, especially in high-stakes industries like the banking sector, regulations, and healthcare where inaccurate information can have disastrous results. The AI-Based Hallucination Detector solves this issue by integrating advanced verification methods that cross- reference reliable web APIs, databases, and information repositories with AI- generated content. This connection ensures that responses from machines using AI are cross-validated with credible sources prior to being shown to users. Because the system relies on verified data from external sources, it can consistently refuse or flag answers that deviate from acknowledged truths. By doing this, people who rely on AI for crucial information might feel more confident because the AI algorithm's overall reliability is significantly boosted. One of the main characteristics of the detector is its ability to apply sophisticated comparison techniques like semantic and contextual analysisX By combining sophisticated verification techniques that cross-reference dependable web APIs, databases, and knowledge repositories with AI-generated content, the AI-Based Hallucination Detector resolves this problem. This link guarantees that replies from AI-powered machines are cross- checked with reliable sources before being displayed to users. The system may consistently reject or flag responses that differ from accepted truths because it depends on validated data from other sources. People that depend on AI for important information may feel more confidence as a result of this because the general reliability of the AI system is much increased. The detector's capacity to use complex comparison methods like semantically and contextual analysis is one of its key features. verification procedure, recording, and examining errors that are reported to improve the AI's subsequent responses. With time, the system learns from past validations, improving its ability to detect hallucinations and lowering the probability of errors. Through the use of machine learning methods, the detector may adjust to the AI model, improving its ability to identify and correct hallucination as more data is examined. Furthermore, the AI-Based Hallucination Detector contributes to the ethical and transparent application of AI, especially in sectors where precision is essential. It helps to prevent risks by reducing the chance of confusion or disinformation by ensuring that customers receive only pertinent and accurate information. The system encourages the wider deployment of AI technologies by improving the dependability of AI-generated content while preserving the essential safety measures to shield consumers from mistakes. A feedback loop is used to support this in conclusion, the AI-Based Hallucination Detector is a crucial part of contemporary AI systems, guaranteeing precision, reliability, and openness. This technology improves the quality and credibility of AI-generated responses by integrating advanced comparison techniques, continuous learning, and external validation sources.

References

[1]. W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu,Y. Zhou, and Y. Liu, "Unsupervised real- time hallucination detection based on the internal states of large language models," arXiv preprint arXiv:2403.06448, 2024.

[2]. Z. Chu, L. Zhang, Y. Sun, S. Xue, Z. Wang, Z. Qin, and K. Ren, "Sora Detector: A Unified Hallucination Detection for Large Text-to-Video Models," arXiv preprint arXiv:2405.04180, 2024.

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

[4]. X. Chen, C. Wang, Y. Xue, N. Zhang,X. Yang, Q. Li, et al., "Unified hallucination detection for multimodal large language models," arXiv preprint arXiv:2402.03190, 2024.

[5]. A. Mishra, A. Asai, V. Balachandran, Y. Wang, G. Neubig, Y. Tsvetkov, and H. Hajishirzi, "Fine-grained hallucination detection and editing for language models," arXiv preprint arXiv:2401.06855, 2024.

[6]. T. Bansal, T. Majumder, and H. Ghosh, "Improving hallucination detection in neural networks via confidence calibration," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 901-912, 2023.

[7]. L. Fan, D. Zhang, and J. He, "Explainable artificial intelligence for NLP: A survey of recent advances," IEEE Access, vol. 9, pp. 112657-112676, 2022.

[8]. Y. Wu, K. Huo, and X. Zhang, "Retrieval-augmented language models: A review of recent advances," IEEE Transactions on Artificial Intelligence, vol. 3, no. 1, pp. 59-78, 2022.

[9]. R. Kim and M. Lee, "Interactive error detection in AI chatbots using user feedback," IEEE Transactions on Human-Machine Systems, vol. 51, no. 6, pp. 845-856, 2023.

[10]. J. Zhang, Y. Sun, and K. Tan, "Real- time AI verification using API-based fact-checking," Journal of Applied AI Research, vol. 5, no. 2, pp. 223-239, 2023.

[11]. H. Liang, S. Wang, and M. Chang, "Towards robust AI: Addressing hallucinations in large language models," IEEE Transactions on Cognitive and Developmental Systems, vol. 13, no. 3, pp. 432-444, 2021.

[12]. T. Green, J. Smith, and A. Patel, "Explainable artificial intelligence for enhancing user trust in critical systems," IEEE Computer, vol. 55, no. 1, pp. 62-72, 2022.

[13]. F. Tan, R. Liu, and B. Wu, "Reinforcement learning for reducing hallucinations in AI models," IEEE Transactions on Machine Learning and Knowledge Extraction, vol. 10, no. 4, pp. 312-328, 2023.

[14]. P. Zhang, M. Hu, and Q. Wang, "Error correction in neural networks: Techniques and applications," IEEE Access, vol. 10, pp. 12345-12359, 2022.

[15]. L. Roberts and G. Zhang, "Explainable AI: Interactive tools for model transparency," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 12, pp. 4921-4932, 2021.

[16]. [Y. Chen and R. Zhao, "A comparison of real-time validation approaches for AI-generated content," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 1385-1397, 2022.

[17]. D. Anderson, P. Chen, and R. Smith, "Interactive tools for mitigating hallucinations in AI-driven systems," IEEE Transactions on Systems, Man, and Cybernetics, vol. 52, no. 8, pp. 1052-1067, 2022.

[18]. K. Cho and M. Han, "Explainability techniques in neural networks: A comprehensive survey," IEEE Transactions on Artificial Intelligence, vol. 4, no. 3, pp. 452-470, 2023.

[19]. S. Patel, A. Jones, and B. Tran, "Comparative analysis of explainable AI frameworks for NLP tasks," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 2, pp. 269-282, 2023.

[20]. N. Liu and Y. Shi, "Risk assessment in AI models: Addressing hallucination and

misclassification issues," IEEE Transactions on Artificial Intelligence, vol. 3, no. 2, pp. 140-158, 2023.

[21]. J. Tan, R. Wang, and L. Xu, "Dynamic learning models for real-time AI hallucination detection," IEEE Transactions on Machine Learning and Knowledge Extraction, vol. 11, no. 3, pp. 356-370, 2023.

[22]. M. Gupta, T. Liu, and S. Choi, "Enhancing transparency in NLP models: Techniques and applications," IEEE Access, vol. 11, pp. 43215-43229, 2023.

[23]. Y. Lee, A. Rao, and K. Kim, "A modular framework for detecting and correcting AI hallucinations in real-time," IEEE Transactions on Artificial Intelligence, vol. 5, no. 1, pp. 25-39, 2024.

[24]. H. Zhao, B. Chen, and D. Li, "Real- time validation techniques for fact- checking AI responses," IEEE Transactions on Human-Machine Systems, vol. 53, no. 5, pp. 781-792, 2024.