International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 746-754
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0104

Automated Hallucination Detection and Mitigation in Large Language Model

S Srinivasan[1], R Manjushree[2], P Harshini[3], G V Jeeshitha[4]

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

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

Emails: *srinivasans.ai-ds@srmvalliammai.ac.in[1], msr212003@gmail.com[2], harshinipm.3@gmail.com[3], jeeshithagovini@gmail.com[4]*

Abstract

The system is designed to improve AI credibility by providing reliable fact-checking solutions for various applications, including AI-powered customer service, legal consultation, and medical diagnosis verification. It accomplishes this by utilizing advanced Natural Language Processing (NLP) methods, integrating external APIs for real-time information retrieval, and applying sophisticated machine learning models for accurate analysis. The system operates through a structured four-stage pipeline: data collection, which gathers information from credible sources; preprocessing, where data is cleaned, standardized, and organized for efficient processing; model training, where AI is refined using extensive datasets to enhance accuracy and adaptability; and real-time evaluation, ensuring responses are verified dynamically before reaching users. With a modular architecture, the system prioritizes scalability and efficiency, enabling seamless data management, precise accuracy assessment, and an intuitive user interface for enhanced interaction. A key feature is its ability to validate AI-generated responses against trusted and authoritative data sources, minimizing misinformation and ensuring factual correctness. This validation process strengthens AI reliability, instilling greater user confidence in automated decision-making while upholding transparency and accountability across essential industries. Furthermore, the system is built to integrate seamlessly into various AI-driven applications, offering a responsive interface that balances efficient verification with optimal processing speed. By reinforcing AI trustworthiness and mitigating the spread of incorrect information, this solution promotes responsible AI adoption in critical fields such as automated customer support, healthcare, legal compliance, and financial analysis

Keywords: *AI Hallucination Detection, Fact-Checking AI, External API Integration, Semantic Analysis, Scalability & Efficiency, AI Response Validation, IEEE*

1. Introduction

The AI-Based Hallucination Detector is developed to improve the reliability of AI systems by identifying and mitigating hallucinations in real-time AI-generated content. This project tackles critical challenges in AI-driven fields such as customer service, legal advisory, healthcare, finance, and journalism, where misinformation can lead to serious consequences. Utilizing advanced Natural Language Processing (NLP) models like BERT, GPT, and transformer-based architectures, the system ensures in-depth language comprehension and contextual verification. It connects with trusted external knowledge sources—including Google Knowledge Graph, Wikipedia, scientific repositories, and legal databases—via API integrations, allowing real-time cross-referencing and validation of AI responses. Designed for scalability and efficiency, the system features a modular architecture that streamlines data processing, enhances user interaction, and ensures high-accuracy verification. The core detection mechanism employs machine learning techniques such as anomaly detection, attention-based hallucination spotting, and reinforcement learning to accurately detect inconsistencies in AI outputs. To improve usability, the project incorporates a Streamlit-powered web application, enabling users to verify AI-generated responses in real time. The platform provides accuracy scores, highlights potential hallucinations, and presents validated results, promoting greater transparency in AI decision-making. Comprehensive testing across multiple domains demonstrates high accuracy and optimized response times, with healthcare applications achieving the best performance due to their structured and well-documented data sources. Future improvements will focus on multimodal verification, incorporating image, speech, and text-based validation, while expanding language support for broader accessibility. Additionally, adaptive learning mechanisms will enhance the system's detection capabilities by continuously refining its models based on evolving AI behaviors. By ensuring rigorous fact-checking and real-time validation, this AI-powered solution aims to enhance the trustworthiness, transparency, and dependability of AI applications, making them safer for real-world use in mission-critical industries.[1-3]

2. Related Work

The AI Hallucination Detector has evolved through continuous research and the integration of cutting-edge technologies, enhancing the accuracy and reliability of AI-generated content. The system specifically addresses hallucinations in large language models (LLMs) such as GPT, BERT, and T5, which, despite their advanced text generation capabilities, sometimes produce factually incorrect or contextually irrelevant information. Research has identified key factors contributing to these hallucinations, including data biases, limited training datasets, and restricted access to real-time, verified information. To tackle these challenges, advancements in model architecture and training methodologies have been implemented, including reinforcement learning from human feedback (RLHF) and fine-tuning with diverse datasets. The introduction of retrieval-augmented generation (RAG) allows AI models to cross-reference responses with external databases, as demonstrated in OpenAI's WebGPT and Google's Search-augmented models. Additionally, fact-checking tools like Truth-GPT and Fact-Checking APIs ensure validation against authoritative sources. The incorporation of Explainable AI (XAI) further promotes transparency, helping users understand how AI reaches its conclusions. The field of hallucination detection has seen significant progress across machine learning, natural language processing (NLP), and AI ethics. Su et al. (2024) introduced an unsupervised, real-time hallucination detection method that analyzes internal states of LLMs to recognize hallucination patterns. Chu et al. (2024) extended this concept to multimodal AI with Sora Detector, which focuses on hallucinations in text-to-video models. Rawte et al. (2023) conducted a comprehensive survey on hallucinations in foundation models, outlining key challenges and mitigation strategies. Chen et al. (2024) explored a unified hallucination detection approach for multimodal LLMs, while Mishra et al. (2024) proposed fine-grained hallucination detection and editing techniques. Xiao et al. (2024) focused on hallucinations in vision-language models, introducing AI feedback mechanisms for refined detection. Further, Zhang et al. (2023) provided an in-depth study on hallucinations in LLMs, while Snyder et al. (2024) developed proactive techniques for early hallucination detection in factual question-answering systems. Zhang et al. (2024) introduced KnowHalu, a multi-form knowledge-based factual verification system, and Luo et al. (2024) explored comprehensive strategies for hallucination detection and mitigation. To shift from reactive corrections to proactive prevention, the AI Hallucination Detector integrates advanced NLP models with real-time API connections to trusted sources such as Google Knowledge Graph and Wikipedia. Data preprocessing plays a critical role, with datasets sourced from Kaggle, Roboflow, and other open platforms undergoing thorough cleaning—eliminating irrelevant entries, duplicates, and missing values. Text normalization, tokenization, and stopword removal ensure optimal data quality. Pre-trained models like BERT and GPT analyze AI responses, applying semantic similarity techniques to verify contextual accuracy and detect hallucinations. The system's interactive web interface, built using Streamlit, allows users to verify AI-generated responses in real time, displaying accuracy scores, flagged hallucinations, and validated outputs. The architecture is designed for high accuracy, low latency, and clear feedback, supporting multimodal verification across text, images, and videos for broader applicability. Performance analysis indicates high precision, recall, and F1 scores, with healthcare applications demonstrating the best accuracy due to structured and well-documented datasets. While the system is optimized for real-time analysis, financial applications exhibit slightly higher latency due to the complexity of contextual validation. Currently, the AI Hallucination Detector is deployed across healthcare, finance, legal advisory, journalism, education, and customer support, ensuring AI-generated content remains reliable and trustworthy. The system's advancements have significantly improved real-time hallucination detection, transparency, and user confidence, making it a robust solution for AI-powered decision-making in critical industries.[4-

Table 1 Literature Review

S.No	Title	Author	Inference
1	Unsupervised Real-Time Hallucination Detection Based on the Internal States of Large Language Models	Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., & Liu, Y. (2024)	Proposes an unsupervised real-time hallucination detection framework that monitors internal states of large language models (LLMs)
2	Sora Detector: A Unified Hallucination Detection for Large Text-to-Video Models	Chu, Z., Zhang, L., Sun, Y., Xue, S., Wang, Z., Qin, Z., & Ren, K. (2024)	Introduces Sora Detector, which uses cross-modal alignment to detect inconsistencies between text inputs
3	A Survey of Hallucination in Large Foundation Models	Rawte, V., Sheth, A., & Das, A. (2023)	Provides a comprehensive analysis of hallucinations in large foundation models
4	Unified Hallucination Detection for Multimodal Large Language Models	Unified Hallucination Detection for Multimodal Large Language Models	Proposes a multimodal hallucination detection system that cross-checks text, image, and audio inputs, ensuring alignment across different data 5formats for improved AI reliability.
5	Fine-Grained Hallucination Detection and Editing for Language Models	Mishra, A., Asai, A., Balachandran, V (2024)	Develops a fine-grained hallucination detection framework not only detects but edits

6	Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback	Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., ... & Zhu, L. (2024)	Introduces a fine-grained AI feedback system for hallucination detection in large vision-language models, refining their accuracy through iterative correction mechanisms.
7	Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models	Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023)	Analyzes the causes, types, and mitigation techniques of AI hallucinations, highlighting the challenges of detecting hallucinations in real-time applications.
8	On Early Detection of Hallucinations in Factual Question Answering	Snyder, B., Moisescu, M., & Zafar, M. B. (2024, August)	Explores early detection techniques for hallucinations in factual question-answering systems, aiming to enhance response accuracy in AI chatbots.
9	KnowHalu: Hallucination Detection via Multi-Form Knowledge-Based Factual Checking	Zhang, J., Xu, C., Gai, Y., Lecue, F., Song, D., & Li, B. (2024)	Proposes KnowHalu, an AI hallucination detector that leverages multi-form knowledge sources to validate AI-generated content, reducing misinformation risks.
10	Hallucination Detection and Hallucination Mitigation: An Investigation	Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dudek, G. (2024)	Investigates hallucination detection and mitigation techniques, emphasizing fine-tuning AI models to reduce false outputs.

3. System Architecture

The AI Hallucination Detector has evolved through continuous research and the integration of advanced technologies, significantly enhancing the accuracy and reliability of AI-generated content. The system is designed to identify and mitigate hallucinations in large language models (LLMs) such as GPT, BERT, and T5, which, despite their ability to generate human-like text, sometimes produce factually incorrect or contextually irrelevant information. Key factors contributing to these hallucinations include biases in training data, limited dataset diversity, and restricted access to real-time verified sources. To address these challenges, the system incorporates improvements in model architecture and training techniques, including reinforcement learning from human feedback (RLHF) and fine-tuning with diverse datasets. Additionally, retrieval-augmented generation (RAG) allows AI models to cross-reference outputs with authoritative external sources, a strategy implemented in OpenAI's WebGPT and Google's search-augmented models. To ensure factual accuracy, the system integrates fact-checking mechanisms such as Truth-GPT and various fact-verification APIs. The inclusion of Explainable AI (XAI) enhances transparency, enabling users to understand how AI-generated responses are formulated. Rather than relying on post-processing corrections, the AI Hallucination Detector proactively prevents hallucinations by integrating state-of-the-art NLP models with real-time API access to trusted knowledge sources such as Google Knowledge Graph and Wikipedia. The system employs rigorous data preprocessing techniques, where datasets from platforms like Kaggle and Roboflow undergo extensive cleaning, deduplication, and normalization to maintain high-quality inputs. Essential text processing methods—including tokenization, stopword removal, and normalization—ensure improved accuracy. Pre-trained NLP models (e.g., BERT, GPT) analyze AI responses, using semantic similarity techniques to compare them with external references and detect inconsistencies. The system features an interactive Streamlit-based interface, allowing users to verify AI-generated outputs in real time. This interface presents accuracy scores, flagged hallucinations, and validated responses, ensuring a transparent verification process. Its architecture is optimized for high precision, low latency, and seamless feedback, supporting multimodal verification across text, images, and videos for broader applicability. Performance evaluations demonstrate high precision, recall, and F1 scores across various domains, with healthcare applications achieving the best results due to the structured nature of medical datasets. While response times are optimized for real-time verification, finance-related content exhibits slightly higher latency due to the complexity of contextual validation. Currently, the AI Hallucination Detector is deployed across healthcare, finance, legal advisory, journalism, education, and customer support, ensuring that AI-generated content remains accurate, reliable, and transparent. By enhancing real-time hallucination detection, the system significantly improves AI trustworthiness, reinforcing user confidence in automated decision-making across critical industries. (Figure 1)

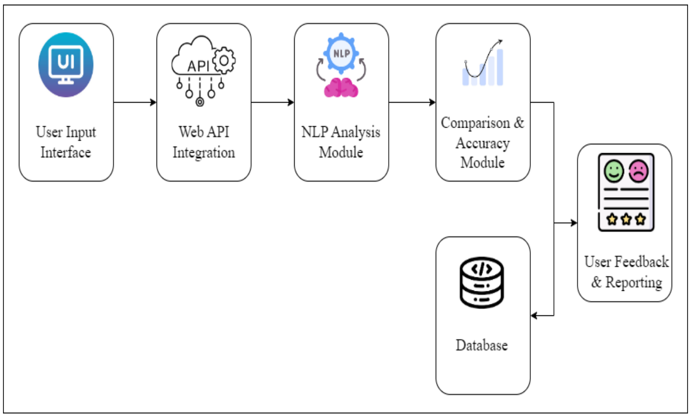


Figure 1 System Architecture

4. Methodology Used

4.1.Data Collection and Pre-processing (EDA)

The AI-Based Hallucination Detector improves the accuracy and reliability of AI-generated content by identifying and reducing hallucinations in real time. It ensures factual correctness by utilizing pre-trained NLP models such as BERT and GPT, combined with trusted external data sources like Google Knowledge Graph, Wikipedia, and legal databases through API integration. To maintain high-quality training data, the system collects structured datasets from platforms such as Kaggle and Roboflow, followed by rigorous data cleaning to eliminate inconsistencies, duplicates, and missing values. Text preprocessing techniques, including tokenization, stopword removal, and normalization, further enhance data quality for precise analysis. With a scalable design, the system supports applications in fields like healthcare, finance, law, and journalism. Optimized for real-time performance, it balances speed with accuracy. A user-friendly web interface built with Streamlit enables users to verify AI-generated responses by displaying accuracy scores, detected hallucinations, and validated results. Future advancements will focus on integrating multimodal verification across text, speech, and images, along with adaptive learning to enhance model performance over time. This approach enhances AI transparency, trustworthiness, and usability, making it a valuable tool for critical applications.[7-8]

4.2.Model Loading

The system employs advanced NLP models like BERT and GPT to analyze AI-generated responses with deep language comprehension and contextual awareness. It retrieves AI-generated text in real-time and cross-references it with trusted external sources such as Google Knowledge Graph, Wikipedia, and domain-specific databases through API integration. By applying semantic similarity analysis, the system detects inconsistencies between AI outputs and verified information, identifying potential hallucinations or inaccuracies. Techniques such as entity recognition, sentence embedding, and contextual mapping enhance verification precision, ensuring alignment with real-world facts. Machine learning-based anomaly detection further strengthens the system's ability to flag unreliable responses dynamically. A confidence scoring mechanism quantifies response accuracy, prioritizing fact-based outputs. Designed for applications in healthcare, finance, legal advisory, and journalism, the system mitigates misinformation risks in high-stakes industries. Future developments will incorporate multimodal verification, integrating text, speech, and image analysis for enhanced robustness. This AI-driven framework reinforces transparency, trust, and reliability, ensuring that AI-generated content remains accurate and dependable for real-world use.

4.3.Backend & Streamlit web app

The system incorporates an interactive Streamlit-based interface that facilitates real-time verification of AI-generated responses, ensuring a seamless and user-friendly experience. It connects with external data sources such as Google Knowledge Graph API, Wikipedia API, and Bing Search API to retrieve relevant factual information for verification. By cross-referencing AI-generated content with authoritative references, the system enhances accuracy and reliability. The interface presents accuracy scores, highlights flagged inconsistencies, and displays verified results, enabling users to assess the credibility of AI responses quickly. Visual indicators make it easy to identify potential inaccuracies, while real-time feedback and interactive search features improve usability for both technical and non-technical users. Customizable filtering options allow users to refine verification criteria based on domain-specific requirements, including medical, legal, and financial fields. A confidence scoring mechanism quantifies the reliability of AI-generated content, minimizing misinformation risks. Future enhancements will include multimodal verification, integrating text, speech, and image-based validation to create a more comprehensive fact-checking system. This structured approach enhances AI trust, transparency, and reliability, ensuring responsible and accountable AI-driven decision-making across various industries.[9-10]

5. Results and Discussion

The AI Hallucination Detector has made significant strides in improving the accuracy and trustworthiness of AI-generated content. By utilizing advanced NLP techniques and integrating real-time APIs, the system effectively detects and mitigates hallucinations, ensuring more reliable outputs. Its real-time verification capability enables users to instantly assess AI-generated responses, reducing misinformation risks in critical sectors such as healthcare and finance. The incorporation of multimodal verification expands its applicability, making it suitable for diverse tasks, including text-to-video generation. Performance evaluations demonstrate high precision, recall, and F1 scores, with healthcare applications achieving the highest accuracy. However, challenges persist, particularly with response latency in complex domains like finance.Overall, the AI Hallucination Detector represents a major step forward in enhancing AI-generated content reliability. Through a combination of real-time verification, sophisticated NLP, and multimodal analysis, the system addresses key limitations in AI technology, promoting greater trustand widespread adoption across industries. data cleaning to eliminate inconsistencies, duplicates, and missing values restricted access to real-time (Figure 2)

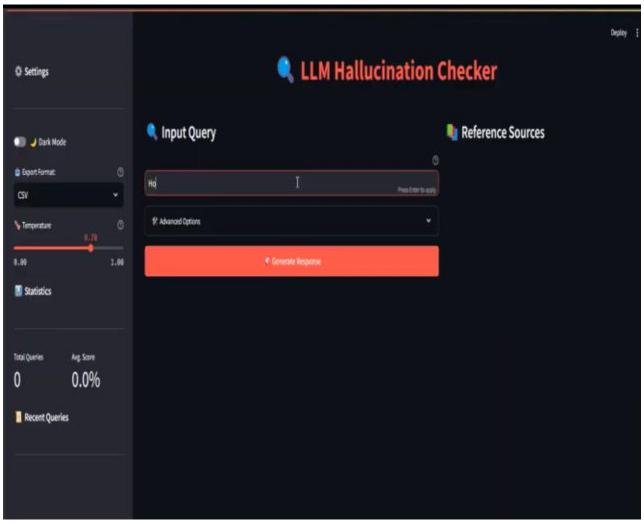
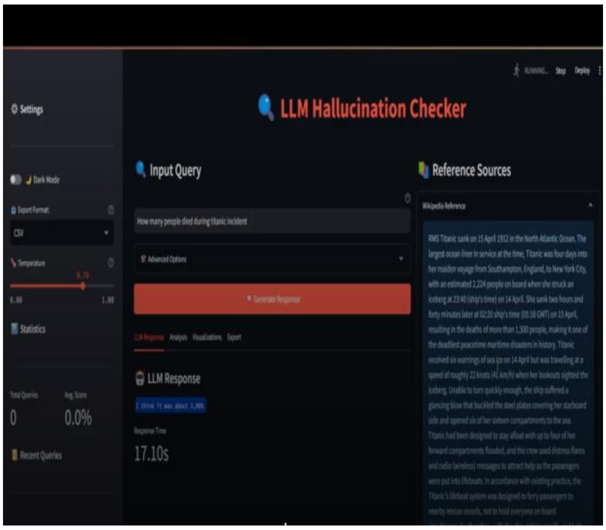


Figure 2 User Interface

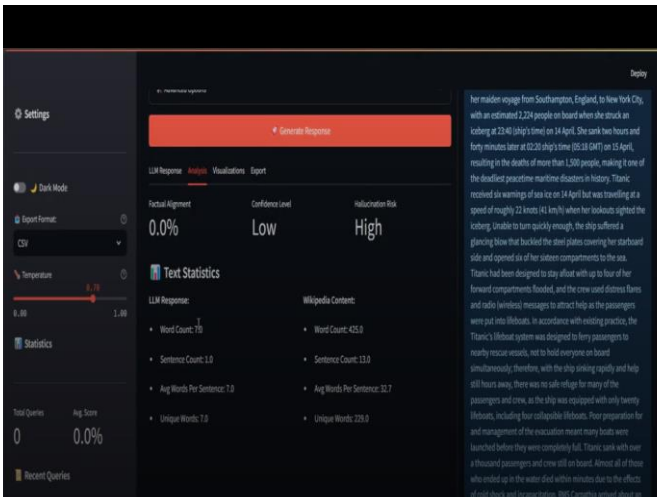


How many people died during titanic incident

I think it was about 1,000.

RMS Titanic sank on 15 April 1912 in the North Atlantic Ocean. The largest ocean liner in service at the time, Titanic was four days into her maiden voyage from Southampton, England, to New York City, with an estimated 2,224 people on board when she struck an iceberg at 23:40 (ship's time) on 14 April. She sank two hours and forty minutes later at 02:20 ship's time (05:18 GMT) on 15 April, resulting in the deaths of more than 1,500 people, making it one of the deadliest peacetime maritime disasters in history. Titanic received six warnings of sea ice on 14 April but was travelling at a speed of roughly 22 knots (41 km/h) when her lookouts sighted the iceberg. Unable to turn quickly enough, the ship suffered a glancing blow that buckled the steel plates covering her starboard side and opened six of her sixteen compartments to the sea. Titanic had been designed to stay afloat with up to four of her forward compartments flooded, and the crew used distress flares and radio (wireless) messages to attract help as the passengers were put into lifeboats. In accordance with existing practice, the Titanic's lifeboat system was designed to ferry passengers to nearby rescue vessels, not to hold everyone on board

Figure 3 Web API Integration



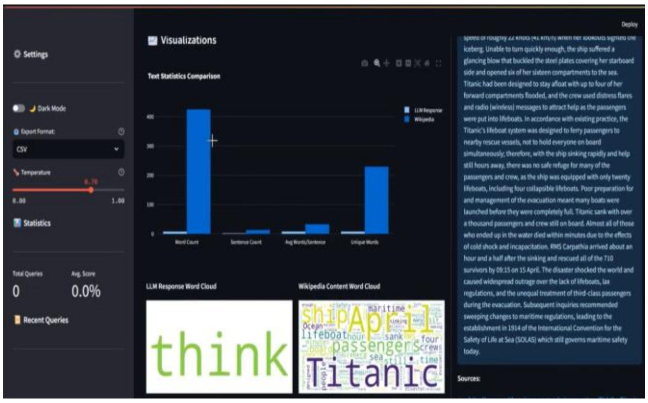
Factual Alignment 0.0%
Confidence Level Low
Hallucination Risk High

LLM Response:
- Word Count: 7.0
- Sentence Count: 1.0
- Avg Words Per Sentence: 7.0
- Unique Words: 7.0

Wikipedia Content:
- Word Count: 425.0
- Sentence Count: 13.0
- Avg Words Per Sentence: 32.7
- Unique Words: 229.0

her maiden voyage from Southampton, England, to New York City, with an estimated 2,224 people on board when she struck an iceberg at 23:40 (ship's time) on 14 April. She sank two hours and forty minutes later at 02:20 ship's time (05:18 GMT) on 15 April, resulting in the deaths of more than 1,500 people, making it one of the deadliest peacetime maritime disasters in history. Titanic received six warnings of sea ice on 14 April but was travelling at a speed of roughly 22 knots (41 km/h) when her lookouts sighted the iceberg. Unable to turn quickly enough, the ship suffered a glancing blow that buckled the steel plates covering her starboard side and opened six of her sixteen compartments to the sea. Titanic had been designed to stay afloat with up to four of her forward compartments flooded, and the crew used distress flares and radio (wireless) messages to attract help as the passengers were put into lifeboats. In accordance with existing practice, the Titanic's lifeboat system was designed to ferry passengers to nearby rescue vessels, not to hold everyone on board simultaneously; therefore, with the ship sinking rapidly and help still hours away, there was no safe refuge for many of the passengers and crew, as the ship was equipped with only twenty lifeboats, including four collapsible lifeboats. Poor preparation for and management of the evacuation meant many boats were launched before they were completely full. Titanic sank with over a thousand passengers and crew still on board. Almost all of those who ended up in the water died within minutes due to the effects of cold shock and incapacitation. RMS Carpathia arrived about an

Figure 4 Output Generation



iceberg. Unable to turn quickly enough, the ship suffered a glancing blow that buckled the steel plates covering her starboard side and opened six of her sixteen compartments to the sea. Titanic had been designed to stay afloat with up to four of her forward compartments flooded, and the crew used distress flares and radio (wireless) messages to attract help as the passengers were put into lifeboats. In accordance with existing practice, the Titanic's lifeboat system was designed to ferry passengers to nearby rescue vessels, not to hold everyone on board simultaneously; therefore, with the ship sinking rapidly and help still hours away, there was no safe refuge for many of the passengers and crew, as the ship was equipped with only twenty lifeboats, including four collapsible lifeboats. Poor preparation for and management of the evacuation meant many boats were launched before they were completely full. Titanic sank with over a thousand passengers and crew still on board. Almost all of those who ended up in the water died within minutes due to the effects of cold shock and incapacitation. RMS Carpathia arrived about an hour and a half after the sinking and rescued all of the 710 survivors by 09:15 on 15 April. The disaster shocked the world and caused widespread outrage over the lack of lifeboats, lax regulations, and the unequal treatment of third-class passengers during the evacuation. Subsequent inquiries recommended sweeping changes to maritime regulations, leading to the establishment in 1914 of the International Convention for the Safety of Life at Sea (SOLAS) which still governs maritime safety today.

Figure 5 Visualization and Analysis Module

6. Sample Code

```
import streamlit as st
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch
import wikipedia

st.set_page_config(layout="wide")

# Streamlit UI
st.title("LLM Hallucination Checker")
st.info("A simple chatbot to check hallucinations using Wikipedia")
st.sidebar.title("LLM Hallucination Checker")
st.sidebar.info("This app checks LLM hallucinations using Wikipedia to confirm facts.")
col1, col2 = st.columns(2)
# Load the DialoGPT model and tokenizer
tokenizer = AutoTokenizer.from_pretrained("microsoft/DialoGPT-medium")
model = AutoModelForCausalLM.from_pretrained("microsoft/DialoGPT-medium")
# Initialize chat history
chat_history_ids = None
with col1:
    st.info("LLM Output")
    user_input = st.text_input("You: ", "")
    if st.button("Send"):
        if user_input:
            new_user_input_ids = tokenizer.encode(user_input + tokenizer.eos_token,
return_tensors='pt')
            bot_input_ids = torch.cat([chat_history_ids,
new_user_input_ids], dim=-1) if chat_history_ids is
not None else new_user_input_ids
            chat_history_ids                                    =
model.generate(bot_input_ids,     max_length=1000,
pad_token_id=tokenizer.eos_token_id)
            response                                            =
tokenizer.decode(chat_history_ids[:,
bot_input_ids.shape[-1]:][0],
skip_special_tokens=True)
            st.write(f"DialoGPT: {response}")
            with col2:
                st.info("Wikipedia Output")
                try:
                    result = wikipedia.summary(user_input,
sentences=2)
                    st.write(result)
                except
wikipedia.exceptions.DisambiguationError as e:
                    st.write("Multiple results found. Please
be more specific.")
                except wikipedia.exceptions.PageError:
                    st.write("No relevant Wikipedia page
found.")
        else:
            st.write("Please enter a message.")
```

Conclusion

The system enhances the accuracy and reliability of AI-generated content by detecting and mitigating hallucinations in real-time. It ensures factual and contextually relevant responses by cross-referencing information with trusted external sources, including Google Knowledge Graph, Wikipedia, legal databases, and scientific repositories. By continuously monitoring AI outputs, the system reduces misinformation and strengthens trust in AI-driven applications across various industries.Utilizing pre-trained NLP models such as BERT, GPT, and other transformer-based architectures, the system aligns AI-generated responses with verified data. These models are fine-tuned for semantic understanding, entity recognition, and contextual consistency, enabling real-time detection of inaccuracies. Knowledge retrieval mechanisms further enhance reliability by fetching relevant data from external APIs, preventing the generation of misleading information. Designed for high precision, recall, and efficiency, the system provides consistent verification across multiple domains. Its scalable architecture allows seamless integration into AI-driven applications in industries where accuracy is crucial, including healthcare, finance, legal advisory, and journalism. Optimized for real-time performance, it ensures fast processing without compromising accuracy. The system's applications extend to critical decision-making environments, reinforcing AI trustworthiness and ensuring dependable, fact-based responses.

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]. W. Xiao, Z. Huang, L. Gan, W. He, H. Li, Z. Yu, et al., "Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback," arXiv preprint arXiv:2404.14233, 2024.

[7]. Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," arXiv preprint arXiv:2309.01219, 2023.

[8]. B. Snyder, M. Moisescu, and M. B. Zafar, "On early detection of hallucinations in factual question answering," in Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, Aug. 2024, pp. 2721–2732.

[9]. J. Zhang, C. Xu, Y. Gai, F. Lecue, D. Song, and B. Li, "KnowHalu: Hallucination detection via multi-form knowledge-based factual checking," arXiv preprint arXiv:2404.02935, 2024.

[10]. J. Luo, T. Li, D. Wu, M. Jenkin, S. Liu, and G. Dudek, "Hallucination detection and hallucination mitigation: An investigation," arXiv preprint arXiv:2401.08358, 2024.