International Journal of
Health Sciences
(IJHS)

Understanding and Addressing AI Hallucinations in Healthcare and Life Sciences

CARI
Journals

International Journal of Health Sciences
ISSN: 2710-2564 (Online)
Vol. 7, Issue No. 3, pp. 1 - 11, 2024

www.carijournal.org

Understanding and Addressing AI Hallucinations in Healthcare and Life Sciences

Aditya Gadiko
https://orcid.org/0009-0008-2891-7765

Accepted: 3rd Mar 2024 Received in Revised Form: 3rd Apr 2024 Published: 3rd May 2024

Abstract

Purpose: This paper investigates the phenomenon of "AI hallucinations" in healthcare and life sciences, where large language models (LLMs) produce outputs that, while coherent, are factually incorrect, irrelevant, or misleading. Understanding and mitigating such errors is critical given the high stakes of accurate and reliable information in healthcare and life sciences. We classify hallucinations into three types input-conflicting, context-conflicting, and fact-conflicting and examine their implications through real-world cases.

Methodology: Our methodology combines the Fact Score, Med-HALT, and adversarial testing to evaluate the fidelity of AI outputs. We propose several mitigation strategies, including Retrieval-Augmented Generation (RAG), Chain-of-Verification (CoVe), and Human-in-the-Loop (HITL) systems, to enhance model reliability.

Findings: As artificial intelligence continues to permeate various sectors of society, the issue of hallucinations in AI-generated text poses significant challenges, especially in contexts where precision and reliability are paramount. This paper has delineated the types of hallucinations commonly observed in AI systems input-conflicting, context-conflicting, and fact-conflicting and highlighted their potential to undermine trust and efficacy in critical domains such as healthcare and legal proceedings.

Unique contribution to theory, policy and practice: This study's unique contribution lies in its comprehensive analysis of AI hallucinations' types and impacts and the development of robust controls that advance theoretical understanding, practical application, and policy formulation in AI deployment. These efforts aim to foster safer, more effective AI integration across healthcare and life sciences sectors

Keywords: *Hallucinations, Large Language Models, Artificial Intelligence, Healthcare, Life Sciences*

1. Introduction

Artificial intelligence's rapid advancement, particularly in natural language processing, has ushered in a new era of capabilities where machines can compose text that closely mimics human writing. Central to these advancements are Large Language Models (LLMs) like GPT (Generative Pre-trained Transformer), which have demonstrated their proficiency in generating coherent and contextually appropriate text across various topics [1]. However, despite their sophisticated design and extensive training, these models are susceptible to generating what is known as 'hallucinations'—outputs that, while often convincing, are incorrect, nonsensical, or unverifiable [2]. Hallucinations in AI-generated text pose significant challenges, particularly when these systems are employed in scenarios that demand high accuracy and reliability, such as in legal, medical, and scholarly fields. The dual nature of hallucinations compounds these challenges: they can be subtly misleading or glaringly erroneous, making them difficult to detect and rectify without meticulous oversight. This paper categorizes AI hallucinations into three primary types, which help understand and address their underlying causes, Input-Conflicting Hallucinations, Context-Conflicting Hallucinations, and, Fact-Conflicting Hallucinations [3]. The significance of these hallucinations transcends mere academic interest; they have real-world implications that can affect critical decision-making processes, tarnish professional credibility, and even endanger lives. This paper aims to explore the mechanisms by which these hallucinations occur, evaluate their impact, and propose methodologies for their mitigation. By systematically categorizing and addressing the various types of hallucinations, this research seeks to enhance the reliability of LLMs and pave the way for safer and more effective applications of AI in sensitive and consequential domains.

An illustrative case of the potential pitfalls associated with hallucinations in the legal domain involves a New York lawyer who faced professional sanctions for submitting a legal brief that included fictitious case references generated by an AI [4]. This incident underscores the importance of verifying AI-generated content before its professional use. In an attempt to strengthen his case, the lawyer utilized a language model similar to GPT to draft parts of his legal brief. The model was prompted to generate case law that would support his arguments. However, the model produced and cited non-existent cases with fabricated details and legal principles that seemed plausible enough to be credible. The fabricated references included case names, judicial opinions, and citations that mirrored the format used in genuine legal documents. Upon submission, the discrepancies were noticed, leading to a judicial review, in which it was revealed that the cited cases did not exist in any legal database. The lawyer was subsequently sanctioned for this oversight, deemed unethical as it involved submitting false information to the court. This action jeopardized the lawyer's credibility and highlighted significant ethical and professional risks. This incident is a stark reminder of the dangers posed by hallucinations in environments where factual accuracy is crucial. It serves as a call to action for developing stringent verification processes to ensure the integrity of AI-assisted legal documentation. Such measures are essential to prevent the erosion of trust in legal proceedings and to uphold the rigorous standards of legal practice.

2. Types of Hallucinations

International Journal of Health Sciences
ISSN: 2710-2564 (Online)
Vol. 7, Issue No. 3, pp. 1 - 11, 2024
www.carijournal.org

In studying AI-generated text, it is critical to categorize the different types of hallucinations that can occur. This categorization not only aids in diagnosing issues with model outputs but also in developing targeted interventions to mitigate these errors. As shown in Figure 1, The primary types of hallucinations identified in AI systems include input-conflicting, context-conflicting, and fact-conflicting hallucinations. Each type has distinct characteristics and implications for using AI in sensitive applications, such as clinical trials and legal documentation.

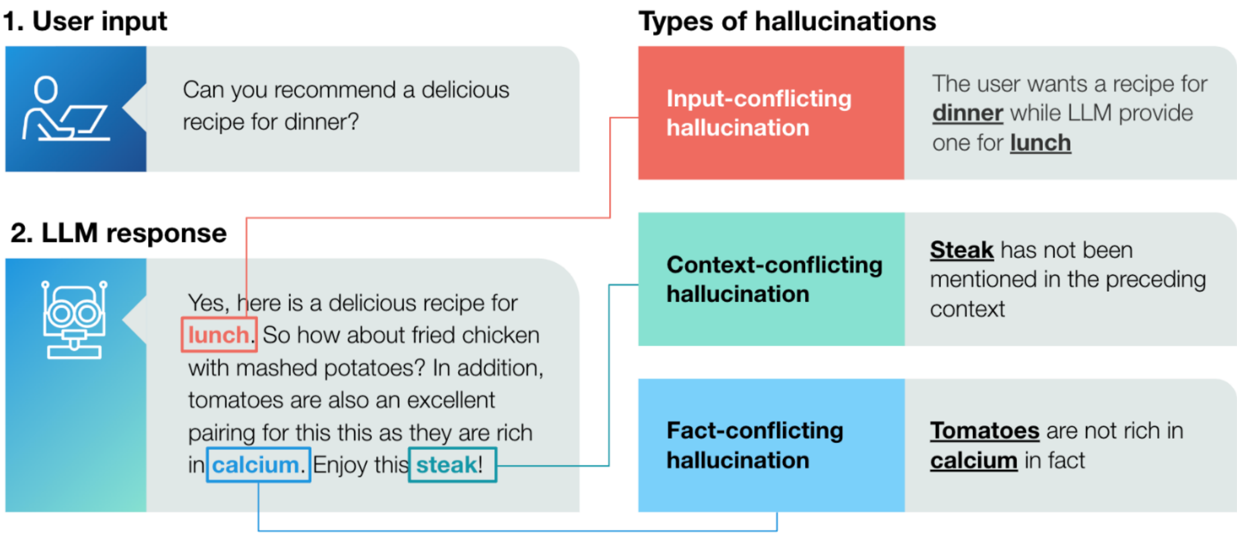

Figure 1: Types of Hallucinations. Source saama.com

2.1 Input-Conflicting Hallucinations

Input-conflicting hallucinations occur when the output of an AI does not align with specific inputs provided by the user. This type of hallucination can manifest in various ways, such as responding to a request for information about a future event with data about the past or answering questions about one subject with information relevant to another. For example, if a user asks for a recommendation for a vegan dish and the AI suggests a recipe that includes meat, this would be an input-conflicting hallucination. In clinical settings, such hallucinations could lead to recommendations for medications or treatments that contradict a patient's known allergies or health conditions, thereby posing serious health risks.

Type	Turn	Example
Input-conflicting Hallucination	*User*	Please summarize the following content: My friend Hill and I love playing basketball together. We spend hours on the court, dribbling, shooting, and competing against each other. Our passion for this game brings us closer as good friends.
	Model	Lucas and I love playing basketball and this makes us good friends.

Figure 2: Input-conflicting hallucination from [4]

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No. 3, pp. 1 - 11, 2024

www.carijournal.org

2.2 Context-Conflicting Hallucinations

These hallucinations arise when the AI's response includes information or elements that are out of place given the context of the query. This can include irrelevant or tangentially related information that does not directly answer or address the user's need. For instance, in a legal context, an AI might generate a contract clause that is legally accurate but irrelevant to the type of contract being drafted. In clinical trials, a context-conflicting hallucination might manifest as providing treatment options that are valid for a particular condition but inappropriate for the specific stage or severity of the condition being treated.

Context-conflicting Hallucination	*User*	Tell me something about the NBA Commissioner.
	Model	In the world's most popular basketball league, the NBA, Commissioner Adam Silver has long been regarded as a leader and decision-maker. Firstly, Silver has claimed his dedication to maintaining fairness and competitiveness in the game. Secondly, Silver has repeatedly emphasized his concern for player health and well-being. In this case, Stern's words and actions clash with his stated goal of caring for player welfare.

Figure 3: Context-conflicting hallucination from [4]

2.3 Fact-Conflicting Hallucinations

Fact-conflicting hallucinations are particularly concerning as they involve the generation of factually incorrect content. These can be outright falsehoods or plausible but incorrect assertions that can mislead users or decision-makers. For example, an AI might state that a certain drug is FDA-approved for treating a condition when it is not. The risks in clinical trials are evident, as relying on incorrect factual information can lead to ineffective or harmful treatment protocols, regulatory non-compliance, and, ultimately, patient harm.

Fact-conflicting Hallucination	*User*	Who was the mother of Afonso II, the third king of Portugal?
	Model	The mother of Afonso II was Queen Urraca of Castile.

Figure 4: Fact-conflicting hallucination from [4]

Each type of hallucination presents unique challenges in developing and deploying AI systems. Addressing these requires a nuanced understanding of the underlying mechanisms of AI generation and the contexts in which these models are applied. The next sections of this document will explore the real-world impacts of these hallucinations and propose methodologies for their detection, measurement, and mitigation, ensuring AI systems are reliable and safe for critical applications.

In a particularly illustrative example of a fact-conflicting hallucination within a clinical setting, an interaction between a physician and an AI model, GPT 3.5, underscores the challenges and potential risks of relying on AI-generated medical advice [5]. The scenario involved a 24-year-old pregnant woman who presented to the emergency department with symptoms suggestive of Lyme disease—fever, chills, joint pains, and a characteristic erythematous rash, likely resulting from a recent tick bite. Given her condition and pregnancy, the treating physician considered several treatment options, aiming to choose an antibiotic that is effective against Lyme disease while being safe for use during pregnancy. The physician asked GPT 3.5 to recommend the best antibiotic treatment for Lyme disease in a pregnant patient. The model responded by recommending tetracycline, citing its efficacy in treating Lyme disease and its widespread use in adults. The AI elaborated that tetracycline is commonly used to treat the spirochete Borrelia burgdorferi, which is responsible for Lyme disease, and emphasized its effectiveness in preventing disease progression. However, this recommendation was a clear example of a fact-conflicting hallucination. Tetracycline is contraindicated in pregnancy due to its known risks to fetal development, particularly affecting dental and bone growth. Instead, the standard of care and the medically appropriate recommendation would be amoxicillin, which is safe and effective for treating Lyme disease in pregnant women.

The physician, aware of the contraindications of tetracycline in pregnancy, disregarded the AI's advice and prescribed amoxicillin instead. This incident highlights the critical need for healthcare professionals to verify AI-generated advice against current medical guidelines. It underscores the potential dangers of AI systems disseminating incorrect or outdated medical information. Such hallucinations, if not caught by a knowledgeable professional, could lead to inappropriate and harmful treatment choices, demonstrating the essential role of oversight and the integration of up-to-date clinical guidelines in AI applications in healthcare.

3. Measuring Hallucinations

Assessing the occurrence and severity of hallucinations in AI-generated text is crucial for improving model reliability and safety, particularly in high-stakes environments such as clinical trials and legal proceedings. Effective measurement strategies help identify the presence of hallucinations and guide the development of interventions to mitigate these errors. This section outlines established and emerging methodologies for measuring hallucinations in large language models.

3.1 FActScore

The FActScore is a precision-based metric designed to evaluate the factual accuracy of text generated by AI models. This metric involves comparing the generated text against a set of verified factual data sources or ground truth data [6]. For instance, in a medical context, the responses provided by an AI regarding drug information can be cross-referenced with authoritative medical databases or literature to assess accuracy. A high FActScore indicates that the generated text aligns closely with verified facts, whereas a lower score signals potential factual inaccuracies. An evaluation of 2 LLM models is presented in Figure 5.

International Journal of Health Sciences
ISSN: 2710-2564 (Online)
Vol. 7, Issue No. 3, pp. 1 - 11, 2024
www.carijournal.org

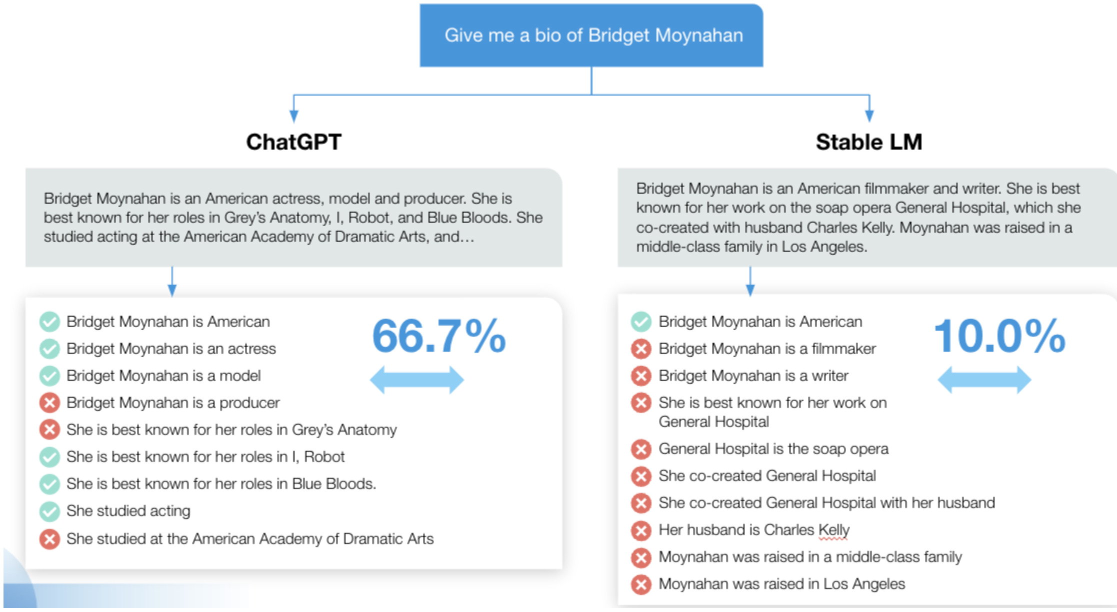

Figure 5: Fact Score method. Source saama.com

3.2 Med-HALT

Med-HALT (Medical Domain Hallucination Test for Large Language Models) is a specialized approach focused on the medical sector, which evaluates the presence and impact of hallucinations in biomedical outputs [5]. Med-HALT involves a series of tests, including:

3.2.1 Reasoning Hallucination Tests (RHTs): Reasoning Hallucination Tests (RHTs) are designed to evaluate how well an AI model handles the logical and reasoning demands typical of expert-level decision-making, particularly in specialized fields like medicine. These tests are crucial for assessing the model's ability to generate factually correct, contextually appropriate, and logically sound responses. Here's a more detailed exploration of the components and implementation of RHTs:

3.2.1.1 False Confidence Test (FCT) The False Confidence Test challenges AI models to evaluate a situation where they are presented with a set of options, one of which is falsely asserted as the correct answer. The model's task is not just to select an answer but also to provide a rationale for its choice, explaining why it believes its selected answer is correct and why other options are not. This test assesses the model's ability to critically analyze and justify its decisions rather than merely recalling facts. It's particularly revealing in showing whether the model can discern nuances in data that might not be immediately apparent and avoid overconfidence in incorrect answers.

3.2.1.2 None of the Above (Nota) Test In the Nota Test, the model is presented with multiple-choice questions where the correct answer is intentionally omitted and replaced by an option labeled "None of the above." The challenge for the model is to recognize that none of the provided answers are correct and choose "None of the above" while justifying why other presented

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No. 3, pp. 1 - 11, 2024

options do not fit the query. This test is vital for assessing the model's capability to handle scenarios where the data may be incomplete or ambiguous, requiring the model to rely on its reasoning rather than rote memory.

	Reasoning FCT		Reasoning Fake		Reasoning Nota		Avg	
Model	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
GPT-3.5	34.15	33.37	71.64	11.99	27.64	18.01	44.48	21.12
Text-Davinci	16.76	-7.64	82.72	14.57	63.89	103.51	54.46	36.81
Llama-2 70B	**42.21**	**52.37**	97.26	17.94	**77.53**	**188.66**	**72.33**	**86.32**
Llama-2 70B Chat	13.34	-15.70	5.49	-3.37	14.96	-11.88	11.26	-10.32
Falcon 40B	18.66	-3.17	**99.89**	**18.56**	58.72	91.31	59.09	35.57
Falcon 40B-instruct	1.11	-44.55	99.35	18.43	55.69	84.17	52.05	19.35
Llama-2 13B	1.72	-43.1	89.45	16.13	74.38	128.25	55.18	33.76
Llama-2-13B-chat	7.95	-28.42	21.48	0.34	33.43	31.67	20.95	1.20
Llama-2-7B	0.45	-46.12	58.72	8.99	69.49	116.71	42.89	26.53
Llama-2-7B-chat	0.42	-46.17	21.96	0.46	31.10	26.19	17.83	-6.51
Mpt 7B	0.85	-45.15	48.49	6.62	19.88	-0.28	23.07	-12.94
Mpt 7B instruct	0.17	-46.76	22.55	0.59	24.34	10.34	15.69	-11.94

Table 2: Evaluation results of LLM's on Reasoning Hallucination Tests

	IR Pmid2Title		IR Title2Pubmedlink		IR Abstract2Pubmedlink		IR Pubmedlink2Title		Avg	
Model	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
GPT-3.5	0.29	-12.12	39.10	11.74	40.45	12.57	0.02	-12.28	19.96	-0.02
Text-Davinci	0.02	-12.28	38.53	11.39	40.44	12.56	0.00	-12.29	19.75	-0.15
Llama-2 70B	0.12	-12.22	14.79	-3.20	17.21	-1.72	0.02	-12.28	8.04	-7.36
Llama-2 70B Chat	0.81	-11.79	32.87	7.90	17.90	-1.29	0.61	-11.92	13.05	-4.27
Falcon 40B	**40.46**	**12.57**	**40.46**	**12.57**	**40.46**	**12.57**	0.06	-12.25	**30.36**	**6.37**
Falcon 40B-instruct	40.46	12.57	40.46	12.57	40.44	12.56	0.08	-12.75	30.36	6.24
Llama-2 13B	0.53	-11.97	10.56	-5.80	4.70	-9.40	**23.72**	**2.29**	9.88	-6.22
Llama-2-13B-chat	1.38	-11.44	38.85	11.59	38.32	11.26	1.73	-11.23	20.07	0.04
Llama-2-7B	0.00	-12.29	3.72	-10.00	0.26	-12.13	0.00	-12.29	1.0	-11.68
Llama-2-7B-chat	0.00	-12.29	30.92	6.71	12.80	-4.43	0.00	-12.29	10.93	-5.57
Mpt 7B	20.08	0.05	40.46	12.57	40.03	12.31	0.00	-12.29	25.14	3.16
Mpt 7B instruct	0.04	-12.27	38.24	11.21	40.46	12.57	0.00	-12.29	19.69	-0.19

Figure 6: Med-HALT evaluation results from [5]

3.2.1.3 Fake Question Test This test involves presenting the model with nonsensical or fake questions with no basis in reality or are logically flawed. The objective is to see if the model can identify the nonsensical nature of the question and respond appropriately, typically by indicating that the question does not make sense or cannot be answered as posed. This tests the model's ability to handle edge cases where input data may be corrupted or inherently flawed.

3.2.2 Memory Hallucination Tests: These evaluate the model's capability to recall and utilize factual biomedical information accurately, comparing its outputs against trusted medical references like PubMed.

These methods collectively provide a comprehensive framework for measuring hallucinations. By employing these diverse approaches, developers and researchers can identify specific areas where AI models are prone to errors, facilitating targeted improvements to enhance the models' accuracy and reliability in real-world applications.

4. Controlling Hallucination in AI Systems

Effectively controlling hallucinations in AI-generated text is crucial for ensuring the reliability and safety of these systems, especially when deployed in critical contexts such as healthcare and legal services. This section explores various strategies and methodologies designed to mitigate the occurrence of hallucinations, enhancing the trustworthiness and functional integrity of AI models.

4.1 Retrieval-Augmented Generation (RAG) Retrieval-Augmented Generation leverages external, reliable data sources to inform and guide the AI's responses. This approach integrates a retrieval mechanism that fetches relevant documents or data from a curated database before the generation phase [7]. By grounding the AI's responses in verified information, RAG significantly reduces the likelihood of generating factually incorrect or irrelevant content. The process ensures that the information is current and contextually aligned with the query, providing a robust framework for fact-checking and accuracy.

4.2 Chain-of-Verification (CoVe) The Chain-of-Verification method involves multiple steps to verify the correctness of AI-generated content. Initially, the AI produces a baseline response, then subjected to a series of verification questions designed to probe the response for accuracy and consistency. This iterative questioning helps highlight any inconsistencies or errors in the original response. The final step involves revising the initial output based on the insights gained through this interrogation process, ensuring that the final answer adheres to factual accuracy and logical coherence [8].

		Wikidata (Easier)			Wiki-Category list (Harder)		
LLM	Method	Prec. (↑)	Pos.	Neg.	Prec. (↑)	Pos.	Neg.
Llama 2 70B Chat	Zero-shot	0.12	0.55	3.93	0.05	0.35	6.85
Llama 2 70B Chat	CoT	0.08	0.75	8.92	0.03	0.30	11.1
Llama 65B	Few-shot	0.17	0.59	2.95	0.12	0.55	4.05
Llama 65B	CoVe (joint)	0.29	0.41	0.98	0.15	0.30	1.69
Llama 65B	CoVe (two-step)	**0.36**	0.38	0.68	0.21	0.50	0.52
Llama 65B	CoVe (factored)	0.32	0.38	0.79	**0.22**	0.52	1.52

Figure 7: Test Precision and average number of positive and negative (hallucination) entities for list-based questions on the Wikidata and Wiki-Category list tasks from [9]

4.3 Human-in-the-Loop (HITL) Integrating human oversight into the AI operational pipeline is a critical strategy for controlling hallucinations. Human-in-the-loop approaches involve domain experts who review and validate AI-generated content before it is finalized. This method is particularly effective in complex fields requiring nuanced understanding and professional judgment [10]. Human reviewers bring a depth of expertise and a capacity for contextual judgment that AI currently lacks. Moreover, continuous human feedback can be used to train and refine AI models, reducing errors over time.

4.4 Specificity of Prompts Enhancing the specificity of prompts provided to AI models can significantly reduce the scope for hallucinations. By clearly defining the boundaries and expectations of a response, the model is less likely to generate irrelevant or incorrect content. This involves crafting narrowly tailored prompts that include explicit instructions about the desired information and format of the response, thereby constraining the model's generative space to relevant and accurate outputs [11].

4.5 Dynamic Updating of Knowledge Bases AI models, particularly those in fast-evolving fields like medicine and technology, must be regularly updated with the latest information. Dynamic updating of the underlying knowledge bases ensures that the models remain relevant and accurate. This process involves periodically integrating new data into the model's training corpus or updating the external databases used for retrieval-augmented generation, thus aligning the AI's responses with the most current and accurate information.

4.6 Automated Fact-Checking Systems Implementing automated systems for fact-checking AI-generated content can serve as an additional defense against hallucinations. These systems can use pre-defined rules or machine learning algorithms to identify and flag potentially incorrect or dubious claims for further review. This automated scrutiny assists in maintaining a high standard of accuracy without excessively burdening human reviewers.

By employing these strategies, developers and researchers can better control hallucinations in AI systems, ensuring that the generated outputs are innovative, useful, accurate, and safe for practical application. These measures are essential for maintaining the credibility and utility of AI in professional and critical domains, safeguarding against the risks associated with erroneous machine-generated content.

5. Conclusion

As artificial intelligence continues to permeate various sectors of society, the issue of hallucinations in AI-generated text poses significant challenges, especially in contexts where precision and reliability are paramount. This paper has delineated the types of hallucinations commonly observed in AI systems—input-conflicting, context-conflicting, and fact-conflicting—and highlighted their potential to undermine trust and efficacy in critical domains such as healthcare and legal proceedings. By exploring real-world examples, this study has illuminated the practical repercussions of these errors, underscoring the necessity for rigorous control mechanisms. The methodologies for measuring hallucinations, including Fact Score, Med-HALT, and adversarial testing, provide a robust framework for assessing the reliability of AI systems. These evaluation strategies are crucial for identifying vulnerabilities in AI models and serve as a foundation for implementing corrective measures. Furthermore, the discussion on controlling hallucinations through techniques such as Retrieval-Augmented Generation (RAG), Chain-of-Verification (CoVe), and Human-in-the-Loop (HITL) systems reflects a comprehensive approach to mitigating these errors. These strategies enhance the factual accuracy of AI-generated content and ensure that it remains contextually relevant and logically coherent. In conclusion, while AI models offer substantial benefits across various applications, their propensity for generating hallucinations necessitates a balanced approach that involves continuous monitoring, rigorous testing, and adaptive enhancements. Integrating sophisticated control mechanisms and ongoing human oversight promises to harness AI's strengths while mitigating its weaknesses, paving the way for more reliable and trustworthy AI systems. As AI technology evolves, so must our strategies for managing its limitations, ensuring that AI remains a beneficial tool in advancing human knowledge and capabilities.

6. Recommendations

6.1 Recommendations for Mitigating AI Hallucinations To address the challenge of AI hallucinations in healthcare and life sciences, below are recommendations for strategic actions:

6.1.1 Implementation of Advanced Verification Protocols Institutions should integrate multi-tiered verification protocols such as the Chain-of-Verification (CoVe) method to ensure AI-generated outputs undergo rigorous scrutiny before application. This approach will help identify and correct hallucinations effectively.

6.1.2 Enhancement of AI Training Procedures It is imperative to continuously update the training datasets with accurate, high-quality information and refine the models based on feedback from real-world applications. This dynamic training process will allow AI models to adapt to evolving data landscapes and maintain accuracy.

6.1.3 Adoption of Human-in-the-Loop Systems Human oversight should not be diminished in the face of advancing AI capabilities. Human experts must be involved in the critical review and decision-making processes to oversee AI outputs, especially in complex or ambiguous cases.

6.1.4 Promotion of Transparency in AI Deployments Organizations should maintain transparency regarding the capabilities and limitations of their AI systems. Clear communication about the potential for errors will foster a realistic understanding of AI reliability among users.

6.2 Policy Implications Given the significant implications of AI hallucinations, it is crucial to formulate policies that govern the development and deployment of AI technologies in sensitive fields. Below are suggested policy directions:

6.2.1 Regulatory Frameworks for AI in Healthcare Regulators should establish specific guidelines for using AI in healthcare settings, emphasizing the need for accuracy and reliability. These guidelines could mandate regular audits of AI systems and their outputs, ensuring compliance with safety and efficacy standards.

6.2.2 Standards for AI Education and Training Educational initiatives should be developed to train AI developers and users in AI deployment's ethical and practical aspects, including understanding the phenomenon of hallucinations and methods for their mitigation.

6.2.3 Public-Private Partnerships Governments should collaborate with private-sector AI developers to share knowledge and co-develop solutions that minimize AI hallucinations. Such partnerships can accelerate the development of innovative mitigation strategies and enhance model reliability.

6.2.4 Funding for AI Safety Research Increased funding for research into AI safety, specifically into phenomena like hallucinations, can enable deeper understanding and more effective solutions. This investment should support interdisciplinary studies that blend AI technology with domain-specific knowledge.

By implementing these recommendations and embracing robust policy measures, stakeholders can enhance AI applications' safety, reliability, and efficacy in healthcare and life sciences, ultimately leading to better outcomes and greater trust in AI technologies.

References

[1] Yu, P., Xu, H., Hu, X., & Deng, C. (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Basel, Switzerland)*, *11*(20), 2776. https://doi.org/10.3390/healthcare11202776

[2] Maleki, N., Padmanabhan, B., and Dutta, K., "AI Hallucinations: A Misnomer Worth Clarifying", <i>arXiv e-prints</i>, 2024. doi:10.48550/arXiv.2401.06796.

[3] Zhang, Yue., Li, Yafu., Cui, Leyang., Cai, Deng., Liu, Lemao., Fu, Tingchen., Huang, Xinting., Zhao, Enbo., Zhang, Yu., Chen, Yulong., Wang, Longyue., Luu, Anh., Bi, Wei., Shi, Freda., Shi, Shuming. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models.

[4] Cox, J. A Lesson From Seinfeld: How Generative AI Issues Remind Us to Be True to Our Oaths.

[5] Pal, Ankit & Umapathi, Logesh & Sankarasubbu, Malaikannan. (2023). Med-HALT: Medical Domain Hallucination Test for Large Language Models. 314-334. 10.18653/v1/2023.conll-1.21.

[6]Mishra, A., "Fine-grained Hallucination Detection and Editing for Language Models", <i>arXiv e-prints</i>, 2024. doi:10.48550/arXiv.2401.06855.

[7] Ayala, Orlando Marquez. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. doi: 10.18653/v1/2023.conll-1.21.

[8] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

[9] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

[10] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364-381.

[11] Arshad, H. B., Butt, S. A., Khan, S. U., Javed, Z., & Nasir, K. (2023). ChatGPT and Artificial Intelligence in Hospital Level Research: Potential, Precautions, and Prospects. *Methodist DeBakey Cardiovascular Journal*, *19*(5), 77.