INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

RESEARCH ESSAY

International University of Applied Sciences

Study-branch: MSc. Artificial Intelligence (120 ECTS)

DLMAISAIS01 – Seminar: AI and Society

Himanshu Saxena
[GitHub] [LinkedIn]

Matriculation Number: 9219463

Tutor: Professor Dr. Tim Schlippe

Hallucination in Generative Artificial Intelligence:

Challenges, Causes, and Mitigation Strategies

Submission date: 03.12.2025

Contents

1 Introduction **1**

2 Background and Motivation **1**

2.1 Defining Generative AI (GAI) and Large Language Models (LLMs) 1

2.2 Motivation: The Rapid Adoption and Utility of GAI 2

2.3 The Hallucination Phenomenon in GAI 2

2.4 Importance of Addressing Hallucinations 2

3 Challenges Posed by Hallucinations in GAI **2**

3.1 Undermining Professional Integrity and Trust 3

3.2 Risks in Digital Health and Medicine 3

3.3 Academic and educational Implications 4

3.4 Resource Wastage and Efficiency Concerns in Research 4

4 Causes of Hallucinations in GAI **4**

4.1 Data Quality and Training Limitations 4

4.1.1 Training Data Issues 5

4.1.2 Statistical Patterns 5

4.1.3 Focus Imbalance 5

4.1.4 Resource Constraints / Complexity Trade-offs 5

4.2 Contextual Understanding Limitations 5

4.2.1 Special Epistemic Uncertainty (Lack of Relevant Context) 5

4.2.2 General Epistemic Uncertainty (Lack of Model Capacity) 6

4.3 Model Architecture and Design Flaws 6

4.3.1 Overfitting and Underfitting 6

4.3.2 Attention Mechanism Limitations 6

4.3.3 Lack of Real-World Grounding 6

5 Mitigation Strategies for Hallucinations in GAI **7**

5.1 Human-Centric and Procedural Mitigation 7

5.1.1 Human-in-the-Loop (HITL) Approaches 7

5.1.2 Transparency and Explainability 7

5.1.3 High-Quality Prompt Engineering 7

5.1.4 Multigenerative and Ensemble Methods 8

5.2 Algorithmic and Architectural Mitigation 8

5.2.1 Retrieval-Augmented Generation (RAG) 8

5.2.2 Knowledge Graph Integration 8

5.2.3 Model Parameter and Training Enhancements 8

5.2.4 Early Detection Mechanisms 9

5.2.5 Testing and Validation Frameworks 9

6 Conclusion and Future Directions **9**

6.1 Summary of Findings 9

6.2 Implications for GAI Development and Deployment 10

6.3 Recommendations for Future Research 11

Bibliography **12**

1 Introduction

In recent years, the integration of Generative Artificial Intelligence (GAI) into various sectors has revolutionized traditional practices and introduced new paradigms for efficiency and innovation. While the integration of Generative Artificial Intelligence (GAI) into various sectors has introduced new paradigms for efficiency and innovation, this progress is fundamentally threatened by the phenomenon of AI hallucination. This paper explores the critical risks posed by fabricated content in GAI outputs, examining the underlying causes and proposing robust mitigation strategies. By analyzing current trends, ethical considerations, and future implications, this work aims to provide a comprehensive understanding of GAI's role in shaping modern society.

The release of GAI systems, particularly Large Language Models (LLMs) like ChatGPT, has brought about a new "AI spring" since November 2022 (Templin et al., 2024). These models have demonstrated remarkable capabilities in generating human-like text, images, and other content, leading to widespread adoption across various industries. GAI systems are recognized for their tremendous potential to enhance productivity, including in scholarly practice, patient care, medical education, business, and creative industries (Frank et al., 2024). However, alongside these advancements, GAI systems have also exhibited a propensity for generating inaccurate or misleading information, a phenomenon commonly referred to as "hallucination", where the model generates content that is plausible, yet fabricated, nonsensical, or factually incorrect (Kim & Lee, 2024). This issue raises significant concerns regarding the reliability and trustworthiness of GAI outputs, particularly in critical applications such as healthcare, legal advice, and education (Hendrickx, 2024). The phenomenon recieved global recognized when the Cambridge Dictionary chose 'hallucinate' as its Word of the Year in 2023 (Béchard & Ayala, 2024).

Today, hallucination and bias happen to be the most frequently discussed challenges in medical AI literature. A significant 64% of reviewed papers addressed model hallucinations (Templin et al., 2024). In legal AI applications, hallucinations lead to the generation of incorrect legal information, potentially resulting in adverse outcomes for users relying on these systems for legal guidance, such as lawyers submitting court filings with fabricated case precedents due to reliance on ChatGPT (Taeihagh, 2025). This paper aims to conduct a scholarly investigation into hallucination in GAI, specifically focusing on the Challenges, Causes, and Mitigation Strategies associated with this phenomenon, thereby exploring their underlying causes and potential strategies to enhance the accuracy and dependability of these systems. The research focuses primarily on GAI systems used for content creation, especially LLMs, which include models like ChatGPT, GPT-4, and Gemini (Frank et al., 2024).

Although the majority of current GAI development is centered on LLMs - with 81% of papers in one review focusing on OpenAI's models, this essay will also include insights into hallucination challenges in other GAI types, such as diffusion models (DMs) used for image generation, where systems like HEaD are employed for early detection (Templin et al., 2024). The scope is limited to synthesizing current research on the technical and ethical dimensions of hallucination, with a focus on solutions rooted in retrieval augmentation (RAG), prompt engineering, and human oversight.

2 Background and Motivation

2.1 Defining Generative AI (GAI) and Large Language Models (LLMs)

GAI systems are defined as a category of AI that leverages machine learning to create new content like text, images, audio, or video, by modeling the underlying data distribution to generate new instances that are statistically similar to the original data (Taeihagh, 2025). LLMs comprise a notable subset of GAI, specifically trained on vast textual datasets like books, articles, and websites, to generate novel text. They can also perform tasks such as translation, summarization, question answering, and even code generation (Taeihagh, 2025) (Templin et al., 2024). Since the public release of models like ChatGPT in November 2022, the GAI systems

grabbed sudden and widespread popularity, accelerating the adoption of GAI across various industries (Frank et al., 2024) (Kothari, 2023) (Lye & Lim, 2024).

2.2 Motivation: The Rapid Adoption and Utility of GAI

In a very short span of time, we have seen the widespread usage of GAI in almost every walk of life. GAI tools have quickly become an indispensable tool for scholars, teachers, and students due to their potential to enhance human capabilities and productivity (Frank et al., 2024) (Templin et al., 2024). The benefits of GAI are not only limited to education, it is now widely becoming an integral part of critical industries, including the business sector, creative industries, and most notably, digital health and academic research (Lye & Lim, 2024); (Miranda et al., 2024). The statistics regarding GAI usage in the academic context, i.e. students use of GAI is very impressive. The academia uses GAI primarily to help with homework (63.7%), brainstorm ideas (46.4%), and for research purposes (41.5%). (Miranda et al., 2024). In this era of digital healthcare, the power of LLMs is being evaluated in the time consuming and tedious tasks like generating medical education materials and supporting clinical decision support, making GAI an essential tool in the healthcare professional's toolkit. (Kothari, 2023); (Templin et al., 2024); (Stadler et al., 2024).

2.3 The Hallucination Phenomenon in GAI

Despite the transformative potential of GAI, significant challenges remain that hinder its reliability and widespread adoption. The central limitation hindering the reliability and user adoption of GAI is AI hallucination. (Béchard & Ayala, 2024). AI hallucination occurs when the GAI system produces outputs that appear convincing but are fabricated, nonsensical, or factually incorrect. (Frank et al., 2024); (Kim & Lee, 2024); (Taeihagh, 2025) With the rapid integration of GAI into high-stakes domains like healthcare and law, the risks associated with hallucinations have become more pronounced. The significant public concern regarding LLMs producing untrustworthy outputs was highlighted when the Cambridge Dictionary chose **'hallucinate'** as its Word of the Year in 2023. (Béchard & Ayala, 2024). Hallucination, along with bias, undermines the trustworthiness and integrity of AI-generated content. When GAI systems are used in critical applications, such as legal advice or medical diagnosis, hallucinations can lead to serious consequences, including legal repercussions and patient harm.

2.4 Importance of Addressing Hallucinations

The problem of hallucination is pervasive in LLM research. The gravity of this impact can be understood by a scoping review on GAI in digital health which found that 64% of the 120 articles reviewed addressed model hallucinations. This makes hallucination the most frequently discussed challenge in the field and impacts a wide range of applications, from clinical decision support to medical education (Templin et al., 2024). Talking about the judiciary domain, which happens to be another high stake domain, the risk of hallucination could result in compromising fundamental rights and judicial legitimacy. (Hendrickx, 2024). The reliability of GAI outputs is crucial for maintaining user trust, especially in professional settings where accuracy is paramount. The risks of GAI systems fabricating or citing non-existent information or case precedents in legal filings, could result in serious legal consequences and challenging the rule of law. (Taeihagh, 2025); (Frank et al., 2024); (Burgess et al., 2024); (Templin et al., 2024).

In the field of technology, in tasks such as image generation, the need for hallucination detection is crucial for optimizing resource consumption. An example is the implementation of HEaD in DMs for detecting incorrect generations early. This allows the system to halt flawed processes and save computational resources. The computational cost of generating flawed images due to hallucinations can be significant, making early detection mechanisms essential for efficiency (Betti et al., 2024).

3 Challenges Posed by Hallucinations in GAI

The prevalence of AI hallucination, a phenomenon, noticibaly marked by the production of highly convincing yet inaccurate or incorrect information, is a key risk that generates complex challenges across all sectors of deployment of GAI. These challenges span ethical, professional, educational, and resource management domains, each presenting unique implications for the adoption and trustworthiness of GAI systems. The AI halucination not only undermine the reliability of GAI outputs but also pose significant risks to user trust, professional integrity, and the efficient use of resources.

3.1 Undermining Professional Integrity and Trust

AI hallucination occurs when GAI systems generate fabricated or incorrect results, which fundamentally undermines the trustworthiness and integrity of the content generated. (Taeihagh, 2025) While the use of GAI systems in judiciary can help speed up the trials, the unfortunate truth is that when GAI systems are used in the judiciary, the capability of LLMs to fabricate case precedents undermines the reliability of the system and the legitimacy of judicial reasoning. (Hendrickx, 2024). It is important to mention the clear example of this risk involving lawyers who submitted court filings with hallucinated content, leading to legal consequences and challenging the integrity of the profession. (Taeihagh, 2025). The risks are amplified when the user is unable to determine the **superficial accuracy** of the GAI systems. The difference in accuracy across GAI systems is considerable, but the less accurate outputs may still superficially appear to be well-stated answers, which can be highly detrimental to an unskilled user (Burgess et al., 2024).

The credibility of the AI system is significantly compromised when users perceive that the AI fails to take responsibility for its mistakes. The users expect detailed explanations when errors occur, and the lack of transparency in error communication leads to a notable decrease in trust and satisfaction(Kim & Lee, 2024). The risk of hallucination is further amplified when users are not adequately trained to critically evaluate AI-generated content, leading to overreliance on potentially flawed outputs. Most of the time, the users generally prefer AI errors to be attributed to external factors e.g., ambiguous input or external data issues rather than internal limitations, as this is less damaging to the AI's perceived competence and helps maintain trust(Kim & Lee, 2024).

3.2 Risks in Digital Health and Medicine

This pervasive challenge spans a wide range of applications, from clinical decision support to medical education. The high prevalence of hallucination highlights its significance as a barrier to the effective and safe deployment of GAI in healthcare. Hallucination is the most frequently discussed challenge in GAI applications within digital health; a review found that 64% of the 120 articles reviewed addressed model hallucinations (Templin et al., 2024). Patient safety is compromised when AI systems generate inaccurate medical information or advice. The reliance on such flawed outputs can lead to misdiagnoses, inappropriate treatments, and adverse health outcomes, especially when healthcare professionals or patients trust the AI without sufficient verification Failure to appreciate the limitations of generative AI, which can produce hallucinations, may lead to misuse and, ultimately, patient harm (Templin et al., 2024).

A significant risk is that patients might substitute GAI output for necessary medical advice, which could result in delayed or harmful health practices. Without proper oversight, patients may rely on inaccurate information for critical health decisions. The risk is particularly pronounced when GAI systems provide medical advice directly to patients without adequate professional review (Templin et al., 2024). Without sufficient understanding of AI limitations, healthcare providers may inadvertently trust and act on flawed AI outputs. Resulting in compromised patient care and safety. Risks are heightened because clinical professionals often lack the foundational AI

training required to adequately recognize and address hallucinations within clinical processes (Templin et al., 2024). Mandatory human review can help mitigate the risks associated with hallucinations and ensure patient safety. The most recommended solution to address hallucinations in medical contexts is external review by experts. The consensus for GAI deployment in medicine is that physicians must review medical advice provided to patients and should not rely solely on an AI for assistance (Templin et al., 2024).

3.3 Academic and educational Implications

Factual Fabrication undermines the credibility of academic work and can lead to the dissemination of misinformation. Hallucination causes LLMs to provide made-up and untruthful outputs, often generating fictional content while presenting it as factual and correct in academic writing (Frank et al., 2024). Diminished critical thinking is a significant concern as students may accept AI-generated content at face value, reducing their engagement in critical analysis and independent thought. Overreliance on GAI by students (63.7% use it for homework) leads to adopting answers without verification, consequently diminishing critical thinking skills (a concern cited by 44.2% of respondents) and hindering the learning process (Miranda et al., 2024). The Verification Difficulty arises from the challenge of discerning accurate information from fabricated content produced by GAI systems. Tertiary students cite the difficulty in verifying the results and accuracy of the information generated by GAI tools as their top concern (46.8% of respondents) (Miranda et al., 2024).

The use of GAI in generating assessment materials raises concerns about the integrity of evaluations, as hallucinated content can lead to unfair assessments and misrepresentation of student knowledge. When GAI is used for creating educational materials, such as crafting medical Multiple Choice Questions (MCQs), the AI might ”hallucinate” facts or details not rooted in its training data, requiring rigorous human review to ensure quality control (Stadler et al., 2024). In specialized fields like physics education, LLM hallucinations can pose problems for student evaluation, feedback, and counseling, where accuracy is paramount (Jho, 2024).

3.4 Resource Wastage and Efficiency Concerns in Research

While using GAI in researches to generate or label data, hallucinations in the generated data can lead to false discoveries and wasted resources (Frank et al., 2024). The early versions of GAI often falsified scientific references, making the resulting generated content useless for any serious academic research that requires valid citation (Frank et al., 2024); (Templin et al., 2024). The necessity of constant human oversight, editing, and fact-checking to correct bias and hallucinations imposes significant overhead on authors, thereby reducing the intended efficiency benefits of GAI-supported academic writing (Frank et al., 2024). In non-text domains, specifically image synthesis using DMs, hallucination leads to computational resource consumption wastage when the process generates numerous flawed images (Betti et al., 2024). Without mitigation strategies, tasks involving structured outputs (like workflow generation) show high hallucination rates (e.g., up to 21% of hallucinated tables), necessitating a time-consuming post-processing layer to flag and correct the errors (Béchard & Ayala, 2024).

4 Causes of Hallucinations in GAI

Hallucination in Generative AI is not a single point of failure but rather a complex symptom arising from limitations across the system design, training data, and knowledge utilization. These causes can be broadly categorized into three main areas: Data Quality and Training Limitations, Contextual Understanding Limitations, and Model Architecture and Design Flaws.

4.1 Data Quality and Training Limitations

Hallucinations are frequently categorized as risks stemming from the training data or the core algorithm used in the GAI system. These causes relate to the quality, scope, and representativeness of the data used to train the models.

4.1.1 Training Data Issues

Hallucinations can happen due to inconsistencies or errors in the training data, or due to flaws in the design of the training process itself (Hendrickx, 2024), (Taeihagh, 2025) The authenticity and integrity of training data play a vital role in controlling the model's ability to hallucinate. LLMs are trained on vast amounts of text data which is often not fully accurate or complete (Jho, 2024). This exposure to flawed or biased information causes the model to generate incorrect outputs (Jho, 2024). Standardization of input data is also crucial. The training data often includes unstructured, non-standardized text, which can lead the model to learn flawed patterns (Jho, 2024). Complex legal queries often require specialized knowledge that may not be adequately represented in the training data. Complete and up-to-date legal databases are essential to minimize hallucinations in legal AI applications. When training data is insufficient for a specific context, such as a particular legal jurisdiction, LLMs generate responses that may not be accurate (Burgess et al., 2024). A risk associated with the reliance on unfiltered internet data is the potential incorporation of biases and the propagation of misinformation, which is often related to hallucination (Taeihagh, 2025)

4.1.2 Statistical Patterns

LLMs generate outputs based on learned probabilities from training data rather than true understanding. They are fundamentally a series of mathematical transformations based on statistical patterns, rather than a conscious reasoning process (Templin et al., 2024). The probabilistic nature of token prediction can lead to the generation of plausible-sounding but incorrect information. Because LLMs generate text by probabilistically predicting the next token or word, they may create sentences that are grammatically consistent but factually "made-up and untruthful" (Frank et al., 2024), (Jho, 2024).

4.1.3 Focus Imbalance

Focus imbalance can lead to outputs that are disconnected from the intended meaning or factual accuracy. Hallucination in LLMs occurs when the model emphasizes certain parts of the input (the prompt) while neglecting other, potentially more relevant parts of the context or internal knowledge (Templin et al., 2024).

4.1.4 Resource Constraints / Complexity Trade-offs

In the case of image generation (DMs), the models often hallucinate "long-tail" objects (elements that are underrepresented in the training datasets) or struggle when generating complex combinations of multiple objects (Betti et al., 2024). The use of optimization techniques to make models more efficient, such as model compression (quantization), can lead to information loss during the optimization process, which can, in turn, exacerbate the hallucination phenomenon (Jho, 2024). While generating complex structured outputs (like JSON workflows), the absence of a Retrieval-Augmented Generation (RAG) system can result in high hallucination rates, demonstrating the inherent difficulty LLMs have with generating consistent, complex data without external grounding (Béchard & Ayala, 2024).

4.2 Contextual Understanding Limitations

Contextual understanding limitations are a major cause of hallucination, arising when the model fails to accurately interpret or utilize the context provided in the prompt or input data. It is often framed as a problem of uncertainty quantification, where the model struggles to determine what it truly "knows" given the context provided (Jesson et al., 2024).

4.2.1 Special Epistemic Uncertainty (Lack of Relevant Context)

An insufficient number of relevant examples provided in the context or prompt (In-Context Learning - ICL) could be linked to hallucination (Jesson et al., 2024). As the number of in-context examples increases, the predictive distribution of the model aligns better with acceptable responses, which helps reduce the occurrence of this special epistemic uncertainty (Jesson et al., 2024). In case of prompts that are vague or imprecise, the hallucination risk is amplified, resulting in outputs that lack the necessary specificity or relevance (Stadler et al., 2024). The LLMs may have limitations in understanding context, particularly in long sentences or complex narratives, leading to incorrect associations or logical leaps when generating text (Jho, 2024).

4.2.2 General Epistemic Uncertainty (Lack of Model Capacity)

If the model's internal knowledge is insufficient to address the question accurately, the LLM fundamentally lacks the capacity to accurately answer a query from a complex or new domain, regardless of how much relevant context is provided (Jesson et al., 2024). The model's inability to generalize beyond its training data leads to hallucinations when faced with unfamiliar or complex queries. If a model has not acquired the capacity to model a specific mechanism class (e.g., complex definite integrals in mathematics), it will exhibit high general epistemic uncertainty and may generate arbitrary, non-factual responses (Jesson et al., 2024). The General Epistemic Uncertainty is evident in tasks where the model's accuracy remains close to random guessing, even when the number of in-context examples is large (Jesson et al., 2024).

4.3 Model Architecture and Design Flaws

The hallucination causes in this category relate to the structural and functional aspects of GAI systems. The internal workings of the model, including how it learns, pays attention, and is designed for optimization, contribute to hallucination. Hallucination can arise from inherent limitations in the model's architecture or design choices that affect its ability to generate accurate and contextually relevant outputs. The following are some of the key causes:

4.3.1 Overfitting and Underfitting

Hallucinations can occur when the model overfits to specific patterns in the training data, causing it to generate outputs that are not applicable to the current context. LLMs tend to overly exploit learned patterns from the training data, leading to a failure to generalize appropriately to new situations, and resulting in the indiscriminate application of existing data in contexts where it does not apply (Jho, 2024). On the other hand, the model may over-optimize for certain rewards, leading to hallucinated outputs that do not align with factual correctness. The use of Reinforcement Learning in training can sometimes lead to the agent learning incorrect behaviors if the reward signals are ambiguous or inaccurate (Jho, 2024).

4.3.2 Attention Mechanism Limitations

The stochastic nature of diffusion models contributes to hallucinations, as small changes in the initial conditions can lead to significantly different outputs. In image generation, the attention mechanisms may struggle to maintain consistency across the diffusion steps, leading to artifacts and inaccuracies in the generated images. The final result is highly dependent on the initial seed, highlighting the inherent unpredictability and variability that can lead to flaws in the diffusion process, requiring restarts to ensure the desired output, (Betti et al., 2024). This inability to consistently render combinations of multiple objects in DMs is tied to limitations in how diffusion patterns produce inconsistencies, suggesting flaws in the attention mechanisms' ability to handle complex visual composition (Betti et al., 2024).

4.3.3 Lack of Real-World Grounding

The real-world grounding is essential for ensuring that the generated content aligns with factual accuracy and practical relevance. LLMs suffer from a lack of grounding, leading them to produce reality-disconnected information that is logically correct within the model's internal structure but inconsistent with the external world (Jho, 2024). LLMs are not inherently designed as fact-retrievers but rather as sophisticated text sequence predictors, often prioritizing fluency and coherence over factual correctness. The prediction of the next token is based on learned patterns rather than a true understanding of the content leading to hallucination (Kim & Lee, 2024). There have been instances where some models exhibit weaknesses in processing numerical values or specific locations; for example, LLMs showed the lowest scores when attempting to locate paragraph numbers in legal text, suggesting an architectural limitation when dealing with non-linguistic elements (Burgess et al., 2024). The GAI systems often lack transparency in their decision-making processes, making it challenging to identify and correct the sources of hallucinations. The inherent opacity of GAI systems (the "black box" nature) makes it difficult to trace their decisions, which increases the likelihood of unintended outputs (Taeihagh, 2025).

5 Mitigation Strategies for Hallucinations in GAI

Mitigating hallucinations in Generative AI is a multifaceted challenge that requires a combination of human-centric (focusing on interaction and oversight) and algorithmic/architectural improvements (focusing on model enhancement and external grounding). This section explores various strategies that can be employed to reduce the occurrence of hallucinations and enhance the reliability of GAI systems. These strategies encompass human-in-the-loop approaches, transparency measures, prompt engineering techniques, multigenerative methods, retrieval-augmented generation, fine-tuning, and post-processing techniques.

5.1 Human-Centric and Procedural Mitigation

5.1.1 Human-in-the-Loop (HITL) Approaches

Maintaining human control over the model's outputs, interpretation, and decision-making throughout the process is the single most important factor in efficiently and ethically using GAI in academic research and writing (Frank et al., 2024). The HITL approach ensures that human judgment is applied to verify and validate AI-generated content, thereby reducing the risk of hallucinations. The most recommended solution to address hallucinations in medical contexts is external review by experts (Templin et al., 2024). In a scoping review of digital health literature, 41% of papers endorsed this practice (Templin et al., 2024). Domain expert can empower healthcare professionals to oversee AI outputs, ensuring that any hallucinated information is promptly identified and rectified before impacting patient care. Introduction of an expert-in-the-loop mechanism for critical applications, such as AI-assisted diagnosis, helps in identifying and correcting model hallucinations (Templin et al., 2024). It is crucial to have human oversight to verify AI-generated content before it is used in clinical decision-making. The consensus among researchers evaluating AI-generated medical advice is that physicians must review medical advice provided to patients and should not rely solely on the AI for assistance (Templin et al., 2024). It is recomended that the authors and researchers must adopt Human Oversight and Editing as a practical strategy for guaranteeing fair use and mitigating the risk of inaccurate content (Frank et al., 2024).

5.1.2 Transparency and Explainability

The lack of transparency in error communication impacts user trust and satisfaction negatively. Transparency in error communication is essential for maintaining user trust following an AI hallucination (Kim & Lee, 2024). Users expect AI systems to provide detailed explanations when errors occur (Kim & Lee, 2024). When reporting an error, users interviewed preferred an apology (politeness strategy) and desired the AI to attribute the error to external factors (Kim & Lee, 2024). As such, the transparency measures should also focus on clear communication about the limitations and potential errors of GAI systems. Specifically, attributing responsibility to the outside world (e.g., misinformation in external data or ambiguity in the user's question) is preferred over internal attribution (e.g., "a mistake during data processing due to the limitations of my algorithm") (Kim & Lee, 2024). For structured outputs (like JSON workflows), a mandatory measure is including a post-processing layer in the deployed system to clearly indicate to users any generated steps that do not exist and urge the user to correct the output before continuing their work (Béchard & Ayala, 2024).

5.1.3 High-Quality Prompt Engineering

The quality of the input prompt significantly influences the accuracy and reliability of GAI outputs. A prerequisite for reducing AI hallucinations and achieving high-quality responses in general is drafting high-quality prompts (Frank et al., 2024). Effective prompt engineering requires providing the system with clear and specific instructions, utilizing constraints and rules to shape the generated output (Frank et al., 2024), (Jho, 2024). The efficacy of GAI-generated content, particularly in educational assessment material like medical MCQs, is profoundly influenced by the precision and clarity of the input prompt (Stadler et al., 2024). The prompt should be carefully designed to minimize ambiguity and provide sufficient context for the model to generate accurate responses. Vague or imprecise prompts can amplify the risk of hallucinations, whereas striking a balance

between brevity and clarity in the prompt is crucial (Stadler et al., 2024). The use of structured prompts, such as step-by-step instructions or specific formatting requirements, can help guide the model towards generating more accurate and relevant outputs. More sophisticated prompting methods (such as Chain-of-Thought or iterative querying) can be used to elicit more accurate responses from LLMs compared to simple, non-sophisticated single prompts (Burgess et al., 2024), (Jho, 2024).

5.1.4 Multigenerative and Ensemble Methods

It is possible to leverage multiple GAI systems to cross-validate outputs and reduce the risk of hallucinations. We propose to use a "multiple GAI" technique, to check for AI bias and hallucination in academic writing, instructing authors to use at least two GAIs and compare their outputs for the same prompt (Frank et al., 2024). This approach allows authors to identify discrepancies and select the most accurate information, thereby enhancing the reliability of the final content. The uncertainty quantification techniques can be employed to assess the reliability of model outputs. These techniques help identify when a model is likely to hallucinate based on its confidence levels. The use of uncertainty metrics can guide users in evaluating the trustworthiness of AI-generated content. The high correlation observed between the Posterior Hallucination Rate (PHR) and other uncertainty metrics (like Mutual Information, MI) suggests that these predictive tools quantify similar information and could be used to cross-validate model certainty (Jesson et al., 2024).

5.2 Algorithmic and Architectural Mitigation

5.2.1 Retrieval-Augmented Generation (RAG)

RAG is a well-known, foundational technique to reduce hallucination and improve output quality by providing access to external knowledge sources (Béchard & Ayala, 2024). The RAG approach combines a retriever model with a generative model to enhance the factual accuracy of generated content. RAG solves the fundamental issue of lack of real-world grounding in LLMs. It is described as a process where relevant information is retrieved from specific data sources prior to generation, ensuring the resulting text is grounded in factual, external knowledge (Béchard & Ayala, 2024), (Jho, 2024). Study shows that in one structured output task (workflow generation), using RAG reduced the rate of hallucinated steps from a high of 21% to less than 7.5% for one model, and hallucinated tables from 42.8% down to 6.6% on average (Béchard & Ayala, 2024). On the efficiency front, RAG enables the use of smaller LLMs without sacrificing performance. By incorporating a very small retriever model, organizations can deploy a smaller LLM (e.g., a 3B parameter model) while keeping hallucination low and maintaining competitive performance (Béchard & Ayala, 2024).

5.2.2 Knowledge Graph Integration

Knowledge graphs refer to structured representations of knowledge that capture relationships between entities and concepts. Integrating knowledge graphs into GAI systems can provide a reliable source of factual information, thereby reducing hallucinations. To address the issue of GAI generating fictitious citations and facts, there is a demand for an AI that accurately searches a knowledge graph (e.g., PubMed, a biomedical abstract database) and produces verifiable citations or linked references (Templin et al., 2024). With improved integration of knowledge graphs, GAI systems can cross-reference generated content against established facts, enhancing the accuracy and reliability of outputs. Integration with existing knowledge is anticipated to become a common strategy in generative AI systems to enhance factual accuracy (Templin et al., 2024).

5.2.3 Model Parameter and Training Enhancements

The adjustment of model parameters and training techniques can help mitigate hallucinations by refining the model's behavior and output quality. Model parameter tuning can influence the likelihood of hallucinations occurring during generation. However adjusting model parameters as a mitigation is suggested, by only 9.2% of reviewed papers in digital health (Templin et al., 2024). An important parameter is the temperature setting during generation. The temperature parameter (e.g., in GPT-4) controls the model's output randomness (Templin et al., 2024). Lower temperatures yield focused results closer to the training data and prompt, thereby reducing

the potential for hallucination (Templin et al., 2024). Providing domain-specific training data through fine-tuning can help the model learn specialized patterns and factual knowledge beyond general pre-training. Finetuning the model on domain-specific data is the traditional method for solving hallucination, helping the model learn special patterns and factual knowledge beyond general pre-training (Jho, 2024).

5.2.4 Early Detection Mechanisms

Early detection mechanisms are of particular importance in non-text GAI applications, such as image synthesis using DMs. In non-text GAI, specifically DMs for image synthesis, the goal is to swiftly detect incorrect generations at the beginning of the diffusion process (Hallucination Early Detection or HEaD) (Betti et al., 2024). The HEaD approach uses cross-attention maps combined with a novel indicator called the Predicted Final Image (PFI) to forecast the final outcome at intermediate generation stages (Betti et al., 2024). Preemptive detection allows the system to halt the faulty generative process and restart with an alternative seed, conserving computational resources by preventing the completion of flawed, low-quality images (Betti et al., 2024). The effectiveness of HEaD is demonstrated in scenarios involving multiple objects, where early detection of hallucinations can significantly reduce wasted computational effort. In a two-object scenario, the HEaD approach was shown to save up to 12% of the average generation time (Betti et al., 2024).

5.2.5 Testing and Validation Frameworks

These frameworks rely on systematic testing and validation procedures to identify and mitigate hallucinations before deployment. Developers should use adversarial testing or out-of-distribution evaluation to help mitigate hallucinations (Templin et al., 2024). The matrics for evaluating hallucination and uncertainty estimation are crucial for understanding and improving model performance. Researchers have developed estimation techniques, such as the Posterior Hallucination Rate (PHR), which quantify the probability that a model will generate a hallucination given an in-context learning problem (Jesson et al., 2024). The PHR estimator is shown to be a reliable predictor of the True Hallucination Rate (THR) and is particularly valuable for predicting errors when few contextual examples are available (Jesson et al., 2024). It is essential to establish robust testing and validation frameworks that can systematically identify hallucinations and assess the effectiveness of mitigation strategies. The successful deployment of AI-assisted diagnosis requires ensuring that responses align with human values and necessitate adversarial testing (Templin et al., 2024).

6 Conclusion and Future Directions

6.1 Summary of Findings

This research confirmed that hallucination, the generation of fabricated or incorrect content, represents a pervasive and critical risk to Generative AI systems, which can only be reliably addressed through a synthesis of human oversight and targeted architectural mitigation strategies. In recent years, GAI has seen rapid adoption across various sectors, bringing both transformative potential and significant challenges. The rapid adoption of GAI tools, such as LLMs, for tasks ranging from medical assistance to academic writing, has made the issue of hallucination, a critical risk to address (Taeihagh, 2025); (Templin et al., 2024). The pervasive nature of hallucination across applications highlights the urgent need for effective mitigation strategies to ensure the reliability and trustworthiness of GAI systems.

Findings from the reviewed research represents that hallucination is pervasive, with 64% of reviewed digital health research addressing model hallucinations, confirming it as the most frequently discussed challenge in the field (Templin et al., 2024).

There are several critical challenges associated with hallucination in GAI, these challenges make it imperative to develop robust mitigation strategies, specially in high-stakes domains such as law, medicine, and education.

In the legal domain, the hallucination phenomenon undermines the reliability of AI-assisted judicial processes,

as LLMs generate fabricated case precedents that compromise the legitimacy of legal reasoning. The undermining professional trust and integrity, exemplified by legal cases where lawyers submitted filings based on hallucinated precedents, leading to severe consequences.

The medical field, being highly sensitive to accuracy, faces significant risks from hallucinated medical advice posing risks of misuse and patient harm in digital health, where clinical professionals may lack the training to identify AI errors. The reliance on hallucinated outputs can lead to misdiagnoses and inappropriate treatments, endangering patient safety. Moreover, the overreliance on GAI by patients without professional review exacerbates these risks.

Education sectors grapple with hallucination undermining academic integrity, as students may accept fabricated AI-generated content without verification. Hindering academic processes by leading to factual inaccuracies, diminishing critical thinking skills among students who rely on unverified outputs, and creating difficulties in verifying AI-generated content.

There exists not one single cause of hallucinations but rather a complex interplay of factors across data quality, contextual understanding, and model architecture. The causes of hallucinations are diverse, stemming from the model's reliance on statistical pattern prediction, errors in the training data, or limitations in its capacity to answer complex queries (General Epistemic Uncertainty).

Solutions to mitigate hallucinations must be multifaceted, combining human oversight with algorithmic improvements, making it clear that no single strategy is sufficient. Effective mitigation is achieved through a multi-layered approach, combining algorithmic enhancements like RAG and mandatory HITL processes.

RAG and ensemble methods are particularly effective in reducing hallucinations by grounding outputs in external knowledge and cross-validating responses. The research confirms that RAG can significantly reduce the hallucination rate in complex tasks, demonstrating its effectiveness in grounding outputs in specific knowledge.

6.2 Implications for GAI Development and Deployment

The emphasis on human oversight and transparency is crucial for building trust in GAI systems, for high-stakes GAI applications, especially in medicine and law, reliance on AI must always be accompanied by human control, review, and editing to mitigate bias and hallucinations. Experts must review all medical advice generated by AI before it is provided to patients. (Frank et al., 2024); (Templin et al., 2024). In technical development, integrating knowledge graphs and adopting RAG architectures are essential steps to enhance factual accuracy and reduce hallucinations. GAI development must shift away from solely relying on sequence prediction and move toward architectures that integrate with external, verifiable knowledge graphs to produce citations or linked references (Templin et al., 2024).

RAG as a foundational technique should be widely adopted to mitigate hallucinations, especially in tasks requiring structured outputs. The implementation of RAG not only improves accuracy but also allows for the deployment of smaller LLMs (e.g., a 3B parameter model) by coupling them with a highly efficient retriever, thereby making deployments less resource-intensive without losing performance (Béchard & Ayala, 2024).

Errors are an inevitable part of any technical development and GAI are no exception. The GAI systems deployed in real-world settings must adhere to transparency standards by offering detailed explanations when errors occur. Furthermore, developers should incorporate strategies like appreciation and external attribution of errors to maintain user satisfaction and trust (Kim & Lee, 2024).

It's imperative for educational institutions to establish clear policies on GAI usage to uphold academic integrity. Educational institutions must adopt strategies—such as the Against, Avoid, and Adopt (AAA) principle—to

design assessments that either ban GAI, focus on high-order skills where GAI performs poorly, or integrate GAI use with mandatory citation and critique (Lye & Lim, 2024).

6.3 Recommendations for Future Research

Mitigation models are still in their infancy, the further understanding and refinement of these models is essential to effectively address hallucinations in GAI, exploring new architectures, training techniques, and evaluation metrics specifically designed to minimize hallucinations. Continued research is necessary to refine models and procedures—such as the proposed iterative process involving multiple GAIs and human consensus—to efficiently address bias and hallucination in academic writing while automating ethical compliance (Frank et al., 2024).

The use of various uncertainty metrics is shown to be highly correlated, suggesting their potential for cross-validation of model certainty. Future studies should further validate the use of probabilistic measures, like the Posterior Hallucination Rate (PHR), as reliable predictors of the True Hallucination Rate (THR) for uncertainty estimation, especially in scenarios with few contextual examples (Jesson et al., 2024).

A transparent error communication framework is crucial for enhancing the user trust and satisfaction following AI hallucinations. Further investigation is needed to explore how user preferences regarding error attribution (e.g., appreciation over apology) impact long-term user engagement and trust with GAI systems (Kim & Lee, 2024).

The future research should explore how user preferences regarding error management strategies (e.g., appreciation vs. apology, external vs. internal attribution) impact long-term user engagement and acceptance of GAI systems. (Kim & Lee, 2024)

The governance of GAI requires adaptive frameworks that can keep pace with the technology's evolution and ensure responsible deployment. Governance can make significant strides by focusing on adaptive and polycentric frameworks that address legal uncertainties and promote inclusive regulatory efforts. Given the dynamic and opaque nature of GAI, future research must focus on establishing adaptive and polycentric governance frameworks. This includes addressing legal uncertainties and ensuring that regulatory efforts are inclusive and not solely driven by established industry powers (Taeihagh, 2025)

Any new technological advancement necessitates a focus on education and literacy to ensure users can effectively and critically engage with the technology. Lack of AI literacy can lead to overreliance on GAI outputs, increasing the risk of accepting hallucinated content as factual. Ongoing research should explore how individuals—especially students and professionals—can be better educated to develop the AI literacy necessary to critically engage with GAI outputs, recognize its limitations (e.g., its poor performance with numerical values like paragraph numbers), and prevent overreliance on the technology (Miranda et al., 2024); (Lye & Lim, 2024); (Burgess et al., 2024).

Bibliography

Béchard, P., & Ayala, O. M. (2024). Reducing hallucination in structured outputs via retrieval-augmented generation.

Betti, F., Baraldi, L., Baraldi, L., Cucchiara, R., & Sebe, N. (2024). Optimizing resource consumption in diffusion models through hallucination early detection. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Burgess, P., Williams, I., Qu, L., & Wang, W. (2024). Using generative ai to identify arguments in judges' reasons: Accuracy and benefits for students. *Law, Technology and Humans*, *6*. https://doi.org/10.5204/lthj.3637

Frank, D., Bernik, A., & Milković, M. (2024). Efficient generative ai-assisted academic research: Considerations for a research model proposal.

Hendrickx, V. (2024). The judicial duty to state reasons in the age of automation: The impact of generative ai systems on the legitimacy of judicial decision-making. *Artificial Intelligence and Law*.

Jesson, A., Beltran-Velez, N., Chu, Q., Karlekar, S., Kossen, J., Gal, Y., Cunningham, J. P., & Blei, D. (2024). Estimating the hallucination rate of generative ai. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Jho, H. (2024). Leveraging generative ai in physics education: Addressing hallucination issues in large language models. *New Physics: Sae Mulli*, *74*, 812–823. https://doi.org/10.3938/NPSM.74.812

Kim, H., & Lee, S. W. (2024). Investigating the effects of generative-ai responses on user experience after ai hallucination. *Proceedings of Social Science and Humanities Research Association (SSHRA), MBP 2024 Tokyo International Conference on Management & Business Practices*, 92–101. https://doi.org/10.20319/icssh.2024.92101

Kothari, A. N. (2023). Chatgpt, large language models, and generative ai as future augments of surgical cancer care. *Annals of Surgical Oncology*. https://doi.org/10.1245/s10434-023-13442-2

Lye, C. Y., & Lim, L. (2024). Generative artificial intelligence in tertiary education: Assessment redesign principles and considerations. *Education Sciences*, *14*, 569. https://doi.org/10.3390/educsci14060569

Miranda, J. P., Gamboa, A., Dianelo, R. F., Bansil, J. A., Hernandez, H., Gonzales, D., & Fernando, E. (2024). Prevalence, devices used, reasons for use, trust, barriers, and challenges in utilizing generative ai among tertiary students.

Stadler, M., Horrer, A., & Fischer, M. R. (2024). Crafting medical mcqs with generative ai: A how-to guide on leveraging chatgpt. *GMS Journal for Medical Education*, *41*, Doc20. https://doi.org/10.3205/zma001675

Taeihagh, A. (2025). Governance of generative ai. *Policy and Society*, *44*, 1–22. https://doi.org/10.1093/polsoc/puaf001

Templin, T., Perez, M. W., Sylvia, S., Leek, J., & Sinnott-Armstrong, N. (2024). Addressing 6 challenges in generative ai for digital health: A scoping review. *PLOS Digital Health*, *3*, e0000503. https://doi.org/10.1371/journal.pdig.0000503