The hallucination problem in Generative Artificial Intelligence: accuracy and trust in digital learning

Bogdan-Iulian CIUBOTARU

Military Equipment and Technologies Research Agency (METRA),
Ministry of National Defence, Clinceni, Romania

bciubotaru@acttm.ro

Abstract: *Generative Artificial Intelligence (GenAI) changes digital learning systems by offering personalized content, adaptive feedback, and interactive study materials. This technology can help fill gaps in resources and reduce educational inequality. The main issue is that it also brings challenges. One major issue is called "The Hallucination Problem." This is when GenAI models create content that sounds believable but is actually false. If these errors are not detected in time, they can reduce trust in the technology and spread misinformation. This paper looks at how GenAI works and how is transforming digital education. The reasons behind these hallucinations are examined, so mitigation risk strategies can be defined. In the end, using GenAI responsibly in digital education means that everyone, teachers, students, and institutions, must understand its risks and the need for systems which were build with human-in-the-loop methods. The best way of hallucination risks mitigation is hallucination awareness. This way, we can take advantage of new technology without sacrificing the quality of education.*

Keywords: artificial intelligence in e-learning, artificial intelligence hallucination, hallucination awareness.

1. Introduction

In an era defined by information, historian Yuval Noah Harari's observation that "having a lot of information doesn't in and of itself guarantee either truth or order" (Harari, 2024) emphasizes an interesting paradox related to modern education. Generative artificial intelligence is transforming the way we understand and develop e-learning systems, offering new opportunities for personalized learning, automated content creation, and adaptive learning systems. The integration of GenAI into educational frameworks introduces significant challenges. The most important risk associated with GenAI is AI "hallucinations", instances where models generate plausible content which is actually. This paper emphasizes how generative AI reshapes digital education, analyses the implications of hallucinations, and proposes strategies to mitigate the risks associated with hallucination.

https://doi.org/10.58503/icvl-v20y202503

1.1 The evolving needs of new-age students

Nowadays students live in a digital connected world, where digital skills are essential. Unlike previous generations, they demand instant access to interactive content and expect learning platforms to adapt to their individual pace and preferences. The need for instant gratification transforms their responsiveness to traditional e-learning models, which are build on static and precise curricula.

Digital readiness and digital skills are critical factors influencing student engagement in e-learning environments. Research indicates that students who possess strong digital skills have confidence in their ability to navigate online learning platforms and tend to achieve better academic outcomes (Kim, Hong & Song, 2019). Furthermore, the Covid-19 pandemic has highlighted the urgent need for both educators and students to develop robust digital competencies, enabling them to respond effectively to modern educational challenges (Saienko, Kurysh & Siliutina, 2022). The ability to adapt to digital learning environments not only enhances student engagement but also helps to create a more equitable educational experience (Li, 2024).

Students face cognitive overload from fragmented information sources (most of information is delivered through mobile phone from different social media platforms), inconsistent feedback mechanisms, and limited opportunities for hands-on practice. For instance, STEM (science, technology, engineering, and mathematics) learners often lack access to lab equipment or real-world simulations, while language students require immersive conversational practice beyond textbook exercises (Sheily Panwar, 2024). These problems make it harder for students to develop the skills they need and increase inequality among them. This is especially true for those who have limited resources — such as not enough computers, reliable internet, or proper study materials. Without these essentials, it becomes even more difficult for them to keep up and gain the hands-on experience and feedback they need to succeed, but GenAI can offer scalable and personalized solutions at lower costs than before, which might mitigate the associated risks.

1.2 The Changing roles of educators in the digital age

As students increasingly rely on digital platforms and tools, educators face significant challenges in adapting to these new ways of learning. For many teachers, especially those who are new to digital literacy, acquiring digital skills can be more demanding. Older adults often acquire digital skills at a slower pace than younger individuals, primarily due to factors such as limited prior exposure to technology, age-related cognitive changes, and distinct learning preferences. Not only the learning methods and tools must adapt to new-age students, but also adapt to the teachers. Many older adults did not grow up with digital technologies, resulting in less familiarity and foundational knowledge, which can make learning new digital skills more challenging. In their study, (Pihlainen, Korjonen-Kuusipuro & Kärnä, 2021) highlights the role of educators in the digital age, emphasizing the importance of adapting teaching methods to meet the unique needs of older learners. The conclusions of article points out the need for educators to develop digital skills and adopt flexible, learner-centered approaches to effectively facilitate digital literacy among older adults.

Cognitive changes associated with aging, such as reduced processing speed and memory retention, which leads to lack of confidence when using technology, can impact the learning process. Social and emotional barriers, like fear of making mistakes make the students feel intimidated, with a negative outcome their learning process. Instructors report that standard digital literacy programs are often ineffective, requiring more patient and personalized teaching methods, adapted to each individual (Vercruyssen et al., 2023).

The entire education system is being transformed by new technologies, reshaping teaching methods, student engagement, and access to learning resources. Also, the reduced attention span of students is another aspect worth considering. It is an ongoing struggle between the rapid emergence of new tools, teacher training, and timely adoption of these innovations to remain relevant. While technology offers opportunities for personalized learning and increased accessibility, its effectiveness depends on how well educators are prepared to integrate it into their teaching practices.

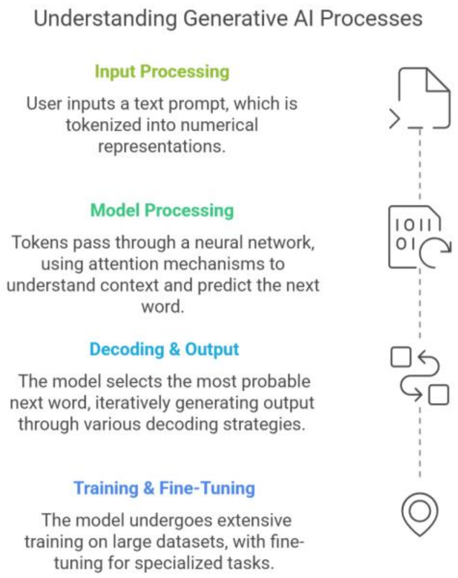
2. How generative AI works in e-learning?

Generative Artificial Intelligence represents a technology defined by the ability of machines to create new content, including text, images, audio, and video, by learning patterns from existing data. This technology uses complex algorithms, particularly deep learning models, which analyze different datasets to identify structures and patterns and generate outputs that is similar to human creativity. Generative AI have different forms, from Generative Adversarial Networks (GANs) used mostly for image generation, Variational Autoencoders (VAEs) used for data generation and dimensionality reduction, and transformer-based models like GPT (Generative Pre-trained Transformer), used for text generation based on user input named prompts (Ali et al., 2024). For instance, GANs consist of two neural networks, the generator and the discriminator, that work in the same time to produce realistic outputs, while transformer models utilize attention mechanisms to generate coherent and contextually relevant text (Sauvola et al., 2024).

Generative AI applications in e-learning are diverse and multiple, offering innovative solutions to improve educational experiences. One important example are the applications which are able to develop personalized learning environments. Generative AI can analyse individual learning patterns and preferences and create personalized content that meets the unique needs of each student (Rasul et al., 2023). For example, AI platforms can generate quizzes, exercises, and study materials that adapt in real-time based on a learner's progress, which promotes a more engaging and effective learning experience (Baidoo-Anu & Ansah, 2023). Also, Generative AI can make the learning systems to adapt and adjust the difficulty of tasks based on the learner's performance, ensuring that students are consistently challenged without becoming overwhelmed (Chan, 2023). Another significant application of generative AI in e-learning is the instant feedback and support which can be offered during the learning phase. AI systems can evaluate student submissions and provide constructive feedback immediately, which is crucial for a better learning, encouraged by continuous improvement (Rasul et al., 2023). For instance, tools like ChatGPT can assist students in understanding complex concepts by generating explanations and examples based on their prompt. For example, the usage of different prompts (like 80-20 rule, known as Pareto principle), can help students to understand complex subjects. This capability not only enhances the learning process but also empowers educators by allowing them to focus on more complex pedagogical tasks while AI handles routine assessments and feedback (Baidoo-Anu & Ansah, 2023), (Rasul et al., 2023).

Generative AI (GenAI) refers to artificial intelligence that can create new content — text, images, music, or even code — by learning from vast amounts of data. Unlike traditional AI models that classify or predict based on predefined categories, GenAI generates original outputs based on patterns it has learned. A particular application of GenAI is the Large Language Models (LLMs) which are specifically designed to process and generate human-like text. They are built using deep learning techniques, particularly transformer architectures, and are trained on massive datasets containing books, articles, and Internet scrapped text. LLMs are evolving rapidly, with improvements in efficiency, reasoning capabilities, and multimodal understanding (combining text, images, and audio). As AI research progresses, we may see models that are more aligned with human reasoning and less dependent on massive datasets.

In Figure 1, the process of how Large Language Models (LLMs) generate text follows four key stages: **Input Processing**: the user provides a text prompt, which serves as the starting point for the model's response. This text is tokenized, meaning it is broken down into smaller units (words or characters) and converted into numerical representations that the model can process. Each token is mapped to an embedding, a dense vector representation that captures its meaning in relation to other words. **Model Processing** is the step where tokenized input is fed into a neural network, typically based on the transformer architecture. Inside the transformer, self-attention mechanisms allow the model to analyse the entire context rather than just relying on immediate neighboring words. The model uses its billions of pre-trained parameters to *determine the most probable next token*, adjusting based on context. The deeper the model, the more refined its understanding of complex structures, enabling nuanced and coherent responses.



Understanding Generative AI Processes

Input Processing

User inputs a text prompt, which is tokenized into numerical representations.

Model Processing

Tokens pass through a neural network, using attention mechanisms to understand context and predict the next word.

Decoding & Output

The model selects the most probable next word, iteratively generating output through various decoding strategies.

Training & Fine-Tuning

The model undergoes extensive training on large datasets, with fine-tuning for specialized tasks.

Figure 1. How Generative AI process works

Decoding & Output Generation: the model iteratively generates words based on probability distributions, selecting the next most likely token each time. Different decoding strategies influence output quality: Greedy search selects the most probable token at each step (fast but can lack diversity), Beam search considers multiple possible continuations before selecting the best one (improves coherence), Temperature control adjusts randomness — higher values make responses more creative, lower values make them deterministic. This cycle repeats until the model generates a complete and meaningful response. In the end, the last step in the for understanding how an LLM works is **Training & Fine-Tuning**, which is a background process. LLMs are pre-trained on vast datasets containing books, articles, and web text using self-supervised learning — predicting missing words in a sentence to learn patterns. Later, they undergo fine-tuning for specific applications, such as legal, medical, or conversational AI tasks, using curated datasets. Reinforcement Learning from Human Feedback (RLHF) can further refine model behaviour, making responses more aligned with user expectations.

3. The hallucination problem in GenAI

In the previous chapter, we learned that LLMs pick their next word based on which word is most likely to come next. However, this process can lead to a big issue: the model may provide information that sounds convincing but is actually wrong or made up. This problem is called "The Hallucination Problem". The fact that LLMs can hallucinate is a significant concern in their development, particu-larly in the context of educational processes, where the accuracy is important.

3.1 The problem

The term "hallucination" in LLMs is used to describe outputs produced by different models (like Llama, BERT, Mistral) which do not correspond to any real-world facts or knowledge. A model can generate a statement about a historical event that never occurred or invent details about a scientific concept. This tendency is rooted in the way LLMs are trained and how they actually work; they learn to predict the next token in a sequence based on patterns in the training data, which can lead to the generation of fake information (Ji et al., 2023), (Maynez et al., 2020). Research has shown that hallucinations can manifest in various forms across different applications of LLMs. For example, in the domain of summarization, models often produce summaries that include fake facts or missunderstandings of the original content (Maynez et al., 2020). This is problematic in use-cases where users depend on these summaries for accurate information. Similarly, in question-answering tasks, LLMs may provide answers that are entirely made up, leading to a lack of trust in their outputs (Sadat et al., 2023).

To illustrate the impact of hallucination, consider a scenario in which university students conducting research on Romania's education system use an LLM to gather statistics about school enrolment rates in rural areas. The AI confidently provides outdated or entirely fabricated numbers, such as a "35% dropout rate in the Teleorman region," without citing real sources. Relying on these fictional figures can skew the conclusions of their research project, leading to flawed recommendations for policymakers or NGOs working to improve rural education.

Hallucination is mainly seen at inference time (when the model generates its outputs) but can be influenced by every stage of a model's life cycle, as described below:

- Pre-Training Phase: if the training data is incorrect or biased, the model may learn misleading correlations. Large, diverse datasets often include errors or unverified information, setting the stage for the model to generate incorrect answers later;
- Fine-Tuning & Reinforcement Learning from Human Feedback: if fine-tuning data or human feedback is inadequately vetted, the model might amplify inaccuracies rather than correct them. During RLHF, evaluators may unintentionally reward responses that sound correct instead of those that are correct;
- Inference (Text Generation) Stage: here is where hallucination becomes visible, as the model selects its next words based on probability rather than factual verification. When the model lacks sufficient context or attempts to answer beyond its knowledge, it tends to "fill in the gaps" with invented facts;
- Deployment & User Interaction: continuous user queries in the real world can expose new gaps in the model's training. Without proper safeguards

— such as fact-checking tools, citations, or human review —incorrect information can spread to end-users.

Hallucination starts from training process and occurs during inference, especially when the model encounters knowledge gaps. It presents a significant challenge, particularly in educational contexts where accuracy is essential.

3.2 The solutions

The hallucination problem has various approaches aimed at reducing its risks; however, from a personal perspective, one of the best ways to reduce its impact is to promote and raise awareness of this phenomenon. Informing the learners, educators, and the public at large about the concept of hallucination makes them to question, verify, and critically think about AI-generated outputs. By highlighting this problem - for example, through workshops, clear guidelines, or policy changes - people can be encouraged to think carefully about AI-generated information and make AI-based educational tools more reliable.

Addressing hallucination is crucial for enhancing the trustworthiness of GenAI usage in digital learning. The strategies to mitigate hallucination are focusing on data curation and pre-training improvements, model architecture and fine-tuning techniques, tools and metrics for detecting and monitoring hallucinations, as if follows:

Data Curation and Pre-Training Improvements

The training data quality is important for biases and hallucinations reducing in LLMs. An effective strategy is to implement data curation processes that prioritize high-quality representative datasets. It is important to have comprehensive data collection methods from various sources, able to minimize biases and improve the accuracy of generated content (Ji et al., 2023). Bias in training data can lead to wrong outputs and reinforce different stereotypes which nowadays society is trying to avoid. Maynez et al. found that training AI with a method called maximum likelihood can sometimes make the text it produces sound less natural or human-like. They suggest that by changing the training process a bit, we can get better results. They also point out that using fair methods to choose training data can help include voices that are often ignored, reducing bias in the final output (Maynez et al., 2020).

Model architecture and fine-tuning techniques

Using human feedback to train language models (RLHF) can make them more accurate and reliable. This method involves people reviewing the model’s outputs during training, helping it focus on being correct rather than just sounding smooth. For example, research (Johnson et al., 2023) showed that this approach improved ChatGPT’s accuracy in medical responses. Adding fact-checking steps directly into the model can also help catch errors as they happen.

Tools and metrics for detecting and monitoring hallucinations

Output checking of a language model for hallucinations is important. Researchers have created tools and tests to see if the generated text is correct. For example, the HaluEval test is designed to find errors or made-up facts in text (Li et al., 2023). While automated tests are useful, it is important for humans to check the output. (Svikhnushina & Pu, 2023) shows that online human reviews can provide detailed feedback on the model's performance. By using both automated tools and human reviews, researchers can better understand and reduce errors, which helps users have more confidence in the model.

3.3 The interest

To investigate the interest in "AI Hallucination," Google Scholar was used as data source. A keyword search for "Artificial Intelligence hallucination" was performed within specific publication years, ranging from 2016 to 2024. For each year, the number of search results returned was recorded, using these counts as a broad measure of academic and research attention on the phenomenon. While some publications might be counted more than once or might not be entirely about AI Hallucination, the overall trend provides a useful snapshot of how the term's usage and academic exploration have evolved. From Figure 2, it can be observed that the release of ChatGPT in 2022 is a likely inflection point: as large language models reached a broader audience, the phenomenon of hallucination garnered significant media and academic attention. Researchers began urgently exploring (in the Figure 2, it can be observed an exponential interest) how to identify, quantify, and mitigate erroneous outputs generated by AI tools now accessible to millions of users. As more people use tools like ChatGPT, deepseek, Claude, and others, the issue of AI making up information has become very noticeable. This has sparked both interest and worry among users.

Researchers and professionals in the tech industry have started to look closely at how AI "hallucinations" - instances when AI systems generate incorrect information - impact our trust and decision-making. The interest is in how these mistakes affect different domains, including education, and how these tools can be used in digital learning.

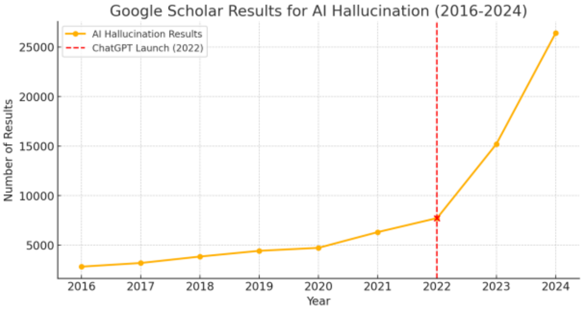


Figure 2. Google Scholar results

4. Conclusions

Generative AI is transforming digital education, creating new teaching methods that adapt to the needs of each student. By processing large amounts of data, new-era digital systems can create learning materials for each student's skill level and interest. In that way, education will become more equitable for students who have access to limited resources. Compared to traditional methods where the information is provided by teachers, generative AI can engage with learners by creating personalized study plans, and offer immediate help and feedback when they are required. This creates a real student-centered learning environment, making education more dynamic and responsive. When used correctly, these tools can work completentary with human teachers' capabilities, rather than replacing them. However, the benefits depend on how well teachers, schools, and students understand and control what AI can do.

A major concern is the AI capability to hallucinate. This means that AI models can create answers that sound real but are actually wrong or not based on evidence. If these errors go unnoticed, they can affect learners' trust and proliferate misinformation. Hallucinations have implications in different life cycle stages of a model:

- During training: the data used to train the AI can have errors or biases;
- In the fine-tuning phase: if the feedback used to improve the AI is incorrect;
- When used in real life: students may take a confident answer from the AI as fact, even if it is wrong. Also, if the models are learning from their interaction with users, they can be manipulated so they can learnig incorrect.

As digital education is globally adopted, the risks associated with hallucination can have bad impact, especially in areas such as academic research where accuracy is essential. To reduce these risks, educators and students should double-check the information generated by AI. Schools can also add fact-checking tools or use peer review, performed by real teachers, to detect errors. It is important for both teachers and students to learn how generative AI works and what are the risks associated. Teachers must understand both the strengths and weaknesses of these tools, and students should question and verify the information they create by GenAI usage.

Schools and universities can include training on AI tools in teacher training. Also, considering that GenAI is not used only for school assesements, courses on how GenAI works should be designed and teached in schools, considering that they can address to students from primary schools to high schools and universities. Training should cover how to use AI effectively, identify potential errors, and understand how the data behind AI could influence output. In that way, educational institutions can create a culture of critical thinking.

In summary, when combined with human-in-the-loop methods, generative AI can be a trusted partner in modern education system. These tools will not replace human teachers, but will enhance their capabilities and help students learn as effectively as possible.

REFERENCES

Ali, S., Ravi, P., Williams, R., DiPaola, D. & Breazeal, C. (2024) Constructing dreams using generative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*. 38(21), 23268-23275. doi: 10.1609/aaai.v38i21.30374.

Baidoo-Anu, D. & Ansah, L. (2023). Education in the era of generative artificial intelligence (ai): understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal.* doi: 10.2139/ssrn.4337484.

Chan, C. (2023) A comprehensive ai policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*. 20(1). doi: 10.1186/s41239-023-00408-3.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023) Survey of hallucination in natural language generation. *ACM Computing Surveys*. 55(12), 1-38. doi: 10.1145/3571730.

Johnson, D. B., Goodman, R., Patrinely, J. R., Stone, C. A., Zimmerman, E. E., Donald, R., … & Wheless, L. (2023) *Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model*. doi: 10.21203/rs.3.rs-2566942/v1.

Kim, H. J., Hong, A. J. & Song, H. D. (2019) The roles of academic engagement and digital readiness in students' achievements in university e-learning environments. *International Journal of Educational Technology in Higher Education*. 16, 21. doi: 10.1186/s41239-019-0152-3.

Li, J., Cheng, X., Zhao, X., Nie, J., & Wen, J. (2023) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. *Halueval: a large-scale hallucination evaluation benchmark for large language models*. doi: 10.18653/v1/2023.emnlp-main.397.

Li, Y. (2024) Adapting to the Digital Learning Environment: The Impact on Student Learning and Outcomes. *Lecture Notes in Education Psychology and Public Media*. 37, 65-71. doi: 10.54254/2753-7048/37/20240504.

Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. (2020) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. *On faithfulness and factuality in abstractive summarization*. doi: 10.18653/v1/2020.acl-main.173.

Mindsmith (2024) *Leveraging Generative AI in eLearning* https://www.mindsmith.ai/blog/leveraging-generative-ai-in-elearning [Accessed 12th February 2025].

Pihlainen, K., Korjonen-Kuusipuro, K. & Kärnä, E. (2021) Perceived benefits from non-formal digital training sessions in later life: views of older adult learners, peer tutors, and teachers. *International Journal of Lifelong Education*. 40(2), 155–169. doi: 10.1080/02601370.2021.1919768.

Rasul, T., Nair, S., Kalendra, D., Robin, M., Santini, F., Ladeira, W., … & Heathcote, L. (2023) The role of chatgpt in higher education: benefits, challenges, and future research directions. *Journal of Applied Learning & Teaching*. 6(1). Doi: 10.37074/jalt.2023.6.1.29.

Sadat, M., Zhou, Z., Lange, L., Araki, J., Gundroo, A., Wang, B., … & Feng, Z. (2023). Delucionqa: detecting hallucinations in domain-specific question answering. *Findings of the Association for Computational Linguistics: EMNLP*. doi: 10.18653/v1/2023.findings-emnlp.59.

Saienko, V., Kurysh, N. & Siliutina, I. (2022) Digital Competence of Higher Education Applicants: New Opportunities and Challenges for Future Education. *Futurity Education*. 2(1), 45–54. doi: 10.57125/FED/2022.10.11.23.

Sauvola, J., Tarkoma, S., Klemettinen, M., Riekki, J. & Doermann, D. (2024) Future of software development with generative AI. *Springer Automated Software Engineering*. 31(1). doi: 10.1007/s10515-024-00426-z.

Sheily Panwar (2024) *Exploring Generative AI in Education with Innovative Applications and Transformative Potential in Classroom Settings*. https://cuc-ulster.edu.qa/post/exploring-generative-ai-in-education-withinnovative-applications-and-transformative/

Svikhnushina, E. & Pu, P. (2023) Approximating online human evaluation of social chatbots with prompting. *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*. doi: 10.18653/v1/2023.sigdial-1.25.

Vercruyssen, A., Schirmer, W., Geerts, N., Mortelmans, D. (2023) How “basic” is basic digital literacy for older adults? Insights from digital skills instructors. *Journal Frontiers in Education*. doi: 10.3389/feduc.2023.1231701.

Yuval Noah Harari. (2024) *Nexus: A Brief History of Information Networks from the Stone Age to AI.* Random House Publishing Group.