What are AI hallucinations?

Preventing AI hallucinations

AI hallucination applications

Products

Resources

Take the next step

AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.

Generally, if a user makes a request of a generative AI tool, they desire an output that appropriately addresses the prompt (i.e., a correct answer to a question). However, sometimes AI algorithms produce outputs that are not based on training data, are incorrectly decoded by the transformer or do not follow any identifiable pattern. In other words, it "hallucinates" the response.

The term may seem paradoxical, given that hallucinations are typically associated with human or animal brains, not machines. But from a metaphorical standpoint, hallucination accurately describes these outputs, especially in the case of image and pattern recognition (where outputs can be truly surreal in appearance).

AI hallucinations are similar to how humans sometimes see figures in the clouds or faces on the moon. In the case of AI, these misinterpretations occur due to various factors, including overfitting, training data bias/inaccuracy and high model complexity.

Preventing issues with generative, open-source technologies can prove challenging. Some notable examples of AI hallucination include:

- Google's Bard chatbot incorrectly claiming that the James Webb Space Telescope had captured the world's first images of a planet outside our solar system.[1]
- Microsoft's chat AI, Sydney, admitting to falling in love with users and spying on Bing employees.[2]
- Meta pulling its Galactica LLM demo in 2022, after it provided users inaccurate information, sometimes rooted in prejudice.[3]

While many of these issues have since been addressed and resolved, it's easy to see how, even in the best of circumstances, the use of AI tools can have unforeseen and undesirable consequences.

Related content

Subscribe to the IBM newsletter →

Explore watsonx.ai →

Implications of AI hallucination

AI hallucination can have significant consequences for real-world applications. For example, a healthcare AI model might incorrectly identify a benign skin lesion as malignant, leading to unnecessary medical interventions. AI hallucination problems can also contribute to the spread of misinformation. If, for instance, hallucinating news bots respond to queries about a developing emergency with information that hasn't been fact-checked, it can quickly spread falsehoods that undermine mitigation efforts. One significant source of hallucination in machine learning algorithms is input bias. If an AI model is trained on a dataset comprising biased or unrepresentative data, it may hallucinate patterns or features that reflect these biases.

AI models can also be vulnerable to adversarial attack, wherein bad actors manipulate the output of an AI model by subtly tweaking the input data. In image recognition tasks, for example, an adversarial attack might involve adding a small amount of specially-crafted noise to an image, causing the AI to misclassify it. This can become a significant security concern, especially in sensitive areas like cybersecurity and autonomous vehicle technologies. AI researchers are constantly developing guardrails to protect AI tools against adversarial attacks. Techniques like adversarial training—where the model is

trained on a mixture of normal and adversarial examples—are shoring up security issues. But in the meantime, vigilance in the training and fact-checking phases is paramount.

Preventing AI hallucinations

The best way to mitigate the impact of AI hallucinations is to stop them before they happen. Here are some steps you can take to keep your AI models functioning optimally:

Use high-quality training data

Generative AI models rely on input data to complete tasks, so the quality and relevance of training datasets will dictate the model's behavior and the quality of its outputs. In order to prevent hallucinations, ensure that AI models are trained on diverse, balanced and well-structured data. This will help your model minimize output bias, better understand its tasks and yield more effective outputs.

Define the purpose your AI model will serve

Spelling out how you will use the AI model—as well as any limitations on the use of the model—will help reduce hallucinations. Your team or organization should establish the chosen AI system's responsibilities and limitations; this will help the system complete tasks more effectively and minimize irrelevant, "hallucinatory" results.

Use data templates

Data templates provide teams a predefined format, increasing the likelihood that an AI model will generate outputs that align with prescribed guidelines. Relying on data templates ensures output consistency and reduces the likelihood that the model will produce faulty results.

Limit responses

AI models often hallucinate because they lack constraints that limit possible outcomes. To prevent this issue and improve the overall consistency and accuracy of results, define boundaries for AI models using filtering tools and/or clear probabilistic thresholds.

Test and refine the system continually

Testing your AI model rigorously before use is vital to preventing hallucinations, as is evaluating the model on an ongoing basis. These processes improve the system's overall performance and enable users to adjust and/or retrain the model as data ages and evolves.

Rely on human oversight

Making sure a human being is validating and reviewing AI outputs is a final backstop measure to prevent hallucination. Involving human oversight ensures that, if the AI hallucinates, a human will be available to filter and correct it. A human reviewer can also offer subject matter expertise that enhances their ability to evaluate AI content for accuracy and relevance to the task.

AI hallucination applications

While AI hallucination is certainly an unwanted outcome in most cases, it also presents a range of intriguing use cases that can help organizations leverage its creative potential in positive ways. Examples include:

Art and design
AI hallucination offers a novel approach to artistic creation, providing artists, designers and other creatives a tool for generating visually stunning and imaginative imagery. With the hallucinatory capabilities of artificial intelligence, artists can produce surreal and dream-like images that can generate new art forms and styles.

Data visualization and interpretation
AI hallucination can streamline data visualization by exposing new connections and offering alternative perspectives on complex information. This can be particularly valuable in fields like finance, where visualizing intricate market trends and financial data facilitates more nuanced decision-making and risk analysis.

Gaming and virtual reality (VR)
AI hallucination also enhances immersive experiences in gaming and VR. Employing AI models to hallucinate and generate virtual environments can help game developers and VR designers imagine new worlds that take the user experience to the next level. Hallucination can also add an element of surprise, unpredictability and novelty to gaming experiences.

Related solutions

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows.

Explore watsonx.governance →

IBM watsonx Assistant

Deliver consistent and intelligent customer care across all channels and touchpoints with conversational AI.

Explore IBM watsonx Assistant →

watsonx.ai

Experiment with foundation models and build machine learning models automatically in our next-generation studio for AI builders.

Explore watsonx.ai →

View the interactive demo →

IBM Watson Discovery

Find critical answers and insights from your business data using AI-powered enterprise search technology.

Explore IBM Watson Discovery →

AI chatbots

Meet a natural language AI chatbot that understands human conversation and improves the customer experience.

Explore AI chatbots →

AI hallucination resources

Blog

Three ways to improve your conversation at AI

Explore three techniques to streamline your AI's dialogue, increase its effectiveness, and satisfy your users: asking for less, giving clear choices and copy editing.

Guide

IBM watsonx Assistant learning

Learn the skills you need to build robust conversational AI with help articles, tutorials, videos, and more.

Tutorial

How to build a chatbot

Check out our docs and resources to build a chatbot quickly and easily.

Market research

Magic Quadrant for Enterprise Conversational AI Platforms, 2023

IBM again recognized as a Leader in the 2023 Gartner® Magic Quadrant™ for Enterprise Conversational AI.

Thought leadership

IBV study: The value of virtual agent technology

Learn about how the COVID-19 pandemic rocketed the adoption of virtual agent technology (VAT) into hyperdrive.

Industry

AI for Customer Service

IBM watsonx users achieved a 337% ROI over three years. Improve the customer experience with conversational AI.

Take the next step

Accelerate responsible, transparent and explainable AI workflows across the lifecycle for both generative and machine learning models. Direct, manage, and monitor your organization's AI activities to better manage growing AI regulations and detect and mitigate risk.

Explore IBM® watsonx.governance™ →

Request a demo →

Footnotes

[1] What Makes A.I. Chatbots Go Wrong? (link resides outside ibm.com), The New York Times, 29 March 2023

[2] ChatGTP and the Generative AI Hallucinations (link resides outside ibm.com), Medium, 15 March 2023

[3] Why Meta's latest large language model survived only three days online (link resides outside ibm.com), MIT Technology Review, 18 November 2022