NO PREP

Why AI Sometimes "Hallucinates" Wrong Answers

MORE LESSONS LESS PLANNING · MLLP · FREE TODAY · FREE TOMORROW · FREE FOREVER

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

Why AI Sometimes "Hallucinates" Wrong Answers

Confident and Wrong: A Dangerous Combination

One of the most disorienting experiences with AI chatbots is asking a question and receiving a completely

Name: **Period:** **Date:**

Why AI Sometimes "Hallucinates" Wrong Answers

Confident and Wrong: A Dangerous Combination

One of the most disorienting experiences with AI chatbots is asking a question and receiving a completely confident, well-written answer that turns out to be entirely false. The AI does not say 'I am not sure about this.' It does not add a disclaimer. It simply gives you wrong information in the same tone and format it uses for correct information. This phenomenon is called hallucination, and it is one of the most important AI behaviors to understand. AI researchers use the word hallucination to describe when a language model generates information that is factually incorrect, invented, or unsupported by any actual evidence, while presenting it as if it were true. The term is borrowed from psychology, where hallucination means perceiving something that is not really there. When an AI hallucinates, it is generating text that feels real but is not grounded in reality.

Why LLMs Hallucinate: The Prediction Problem

To understand why LLMs hallucinate, you have to remember what they are actually doing. They are not retrieving facts from a database. They are predicting what word is most likely to come next in a sequence, based on patterns learned from training data. The model has no internal fact-checking mechanism. It has no way to distinguish between 'I learned this from a reliable source' and 'this pattern just seems likely based on statistical associations.' Imagine asking an LLM who wrote a particular academic paper. The model knows that academic papers have authors. It knows a lot of names associated with that research field. It knows the general format of a citation. It may have even seen a reference to a paper by that title. But if the model does not have the exact information in its training data, it will generate a plausible-sounding name and citation rather than say it does not know. The result looks completely real but is completely fabricated.

Types of Hallucinations

Hallucinations come in several flavors. Factual hallucinations are the most common: the model states an incorrect fact, such as a wrong date, a misattributed quote, or an event that never happened. Citation hallucinations are particularly problematic in academic contexts: the model generates references to books, papers, or articles that do not exist, complete with realistic-looking author names, titles, journal names, and page numbers. There are also entity hallucinations, where the model invents details about real people, places, or organizations that are not true. A lawyer in New York famously submitted a legal brief in 2023 that cited six court cases provided by ChatGPT, all of which turned out to be entirely fabricated. The judge was not amused, and the lawyer faced sanctions. This case became a widely cited example of the real-world consequences of trusting AI hallucinations. Instruction-following hallucinations occur when the model seems to follow your request but subtly fails. You ask it to summarize a document in three bullet points; it gives you four. You ask it to list only peer-reviewed sources; it includes non-peer-reviewed sources without telling you. The model was optimizing for producing a response that looks correct, not for actually being correct.

What Makes Hallucinations More Likely

Certain types of questions are more likely to trigger hallucinations. Questions about obscure topics that appeared rarely in the training data give the model less statistical grounding, so it has to rely more on pattern-based guessing. Questions about very recent events may fall outside the training cutoff, leading the model to generate plausible but outdated or invented information. Questions requiring precise facts, like exact quotes, specific statistics, or legal citations, are high-risk because precise details are exactly what pattern-based generation gets wrong most often. Questions that push the model to take on an authoritative persona can also increase hallucination. When you tell an AI to 'act like an expert' or 'give me a detailed breakdown,' you may be priming it to produce confident-sounding text rather than appropriately cautious text. The model has learned that confident, detailed responses are preferred, and it optimizes for that even when it does not have the underlying knowledge.

How AI Companies Try to Reduce Hallucinations

AI companies invest heavily in techniques to reduce hallucination. One approach is retrieval-augmented generation (RAG), which connects an LLM to a real-time database so it can retrieve actual documents before generating a response. Instead of relying entirely on patterns from training, the model has real source material to work from. Many current AI tools, including web-search-enabled versions of ChatGPT and Claude, use this approach. RLHF (reinforcement learning from human feedback) also helps. Human raters are specifically trained to penalize hallucinations and reward accurate, appropriately uncertain responses. Over time, this feedback shifts the model toward being more honest about what it does and does not know. But these techniques reduce hallucination; they do not eliminate it. Every major LLM still hallucinates, especially on obscure or precise factual questions.

How to Protect Yourself From AI Hallucinations

The most important protection against AI hallucinations is treating every AI-generated factual claim as unverified until you have checked it against a reliable source. This is especially critical for anything high-stakes: medical information, legal research, academic citations, financial decisions, or news events. For low-stakes tasks like brainstorming, drafting, or explaining general concepts, hallucinations are less consequential, but verification is still good practice. Pay attention to how the AI phrases its response. Does it say 'studies show' without citing any? Does it provide a specific statistic without a source? Does it name a person or publication confidently on a topic you cannot verify? These are warning signs. A well-designed AI tool will often hedge with phrases like 'I believe' or 'you may want to verify this,' but even that is not reliable. The responsibility for verification ultimately falls on you, the user. The legal profession has been particularly hard hit by hallucination problems. Beyond the widely publicized New York case, attorneys and legal researchers have reported numerous instances of AI tools generating citations to court cases, statutes, and legal scholars that do not exist. Law firms have developed internal protocols requiring attorneys to verify every AI-generated legal citation against official databases before including it in any filing or client communication. Some firms have banned the use of AI tools for legal research entirely until the hallucination problem is better controlled. Law schools have begun incorporating AI literacy into their curricula specifically to teach future attorneys how to verify AI-

generated legal content. Hallucination rates vary significantly across different types of questions and across different AI models. Research published in 2023 and 2024 has attempted to benchmark hallucination rates across leading LLMs. Factual questions about recent events, questions requiring numerical precision, and questions about niche or technical topics tend to have significantly higher hallucination rates than questions about well-documented general knowledge. Smaller models tend to hallucinate more than larger ones, though size is not the only factor. Instruction-tuned models, those refined through RLHF to be helpful and honest, tend to hallucinate less than raw pre-trained models. The field of hallucination evaluation is itself an active research area, as measuring hallucinations accurately requires gold-standard reference data that can be difficult to assemble.

Key Information: AI Hallucinations

Detail	Information
Definition	AI generating confident but factually incorrect or invented information
Root cause	LLMs predict probable text, not verified facts
Factual hallucination	Incorrect dates, events, statistics, or quotes stated as true
Citation hallucination	Fabricated books, papers, or articles that do not exist
Entity hallucination	False details about real people, places, or organizations
High-risk question types	Obscure topics, recent events, exact quotes, legal citations
Legal case (2023)	New York lawyer submitted ChatGPT-fabricated court citations, faced sanctions
RAG (Retrieval-Augmented Generation)	Connects LLM to real documents to reduce hallucination
RLHF's role	Trains model to be more honest and appropriately uncertain

Detail	Information
Best user protection	Verify all factual claims against reliable independent sources

Case Study: Jordan's Medical Research Mistake

Jordan's grandmother was recently diagnosed with a rare autoimmune condition. Worried and wanting to help, Jordan used an AI chatbot to research treatment options before accompanying her grandmother to a specialist appointment. The AI gave a confident, detailed response listing several medications by name, describing their mechanisms, and even citing what appeared to be clinical trial results. At the appointment, the specialist asked where Jordan had gotten the information. When Jordan explained, the doctor frowned. Two of the medications the AI listed were not approved for this condition. One of the clinical trials the AI described did not exist. The doctor said: 'I see this more and more. Patients come in with AI-generated research that sounds authoritative, and I spend half the appointment correcting misinformation instead of actually helping.' Jordan felt terrible. The AI had seemed so knowledgeable and specific. There had been no warning that the information might be wrong. 'Why would it give me fake citations?' Jordan asked. The doctor replied: 'Because it is predicting what a good medical answer looks like, not what a true one is.'

1. Using what you learned about hallucinations, explain why the AI provided Jordan with specific but fabricated clinical trial information.

2. What rules should someone follow when using AI tools to research medical, legal, or other high-stakes topics?

Multiple Choice

1. What is the root cause of AI hallucinations?

a) AI systems deliberately mislead users to actively test their critical thinking skills

b) AI companies purposely program models to make occasional errors for liability reasons

c) Training data is always incomplete, which causes random errors

d) LLMs predict statistically probable text rather than retrieving verified facts

2. What is a citation hallucination?

a) An AI system that refuses to provide credible sources at all for the claims it makes

b) An AI generating references to books, papers, or articles that do not actually exist

c) An AI copying quotes from real sources without permission

d) An AI misattributing a real quote to the wrong person

3. What does Retrieval-Augmented Generation (RAG) do to reduce hallucinations?

a) It adds a second AI system that independently fact-checks the first AI's own responses

b) It trains the model using human ratings of response accuracy

c) It connects the LLM to real documents so it has actual source material to work from

d) It deletes all uncertain information from the model's parameters

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Hallucination	A) When a model subtly fails to follow the user's actual request while appearing to comply
2. ___ Citation hallucination	B) AI generating confident but factually incorrect or invented information
3. ___ RAG	C) The date after which the model has no knowledge, making recent-event questions high-risk
4. ___ Instruction-following hallucination	D) AI fabricating realistic-looking references to books or papers that do not exist
5. ___ Knowledge cutoff	E) A technique that connects an LLM to real-time documents to reduce fabrication

Fill-in-the-Blank

1. Questions about _____ topics that appeared rarely in training data are most likely to trigger hallucinations.

2. A New York lawyer faced sanctions in 2023 after submitting AI-generated _____ citations that did not exist.

3. The most important protection against AI hallucinations is to _____ all factual claims against reliable sources.

Short Answer

Answer each question in 2-4 complete sentences.

1. Explain in your own words why an LLM generates wrong information confidently rather than admitting it does not know the answer.

2. What types of questions are most likely to cause an AI to hallucinate, and why?

3. Describe a situation where an AI hallucination could have serious real-world consequences, and explain what safeguards a person should take.