CS-114 · AI Security

AI Hallucinations

When an AI system produces confident output that is false, invented, or unsupported.

Executive Summary

An AI hallucination is output that sounds confident and fluent but is factually wrong, made up, or not supported by any real source. Language models generate the most likely next words rather than looking up verified facts, so they can invent citations, names, figures, and events. Hallucinations are a reliability and governance risk that teams manage with grounding, verification, and human oversight rather than eliminate outright.

What It Is

A hallucination is a plausible-sounding but incorrect or fabricated answer from a generative AI system. It happens because a language model predicts likely text based on patterns it learned, not because it retrieves a checked fact from a database. When the model has a gap, it fills it smoothly, which is why a wrong answer can read just as confidently as a right one.

Hallucinations take several forms: inventing sources or quotes, stating false facts, misattributing information, or drifting away from a document the model was supposed to summarize. The fluent, authoritative tone is what makes them dangerous, because it invites trust the output has not earned.

Why It Matters

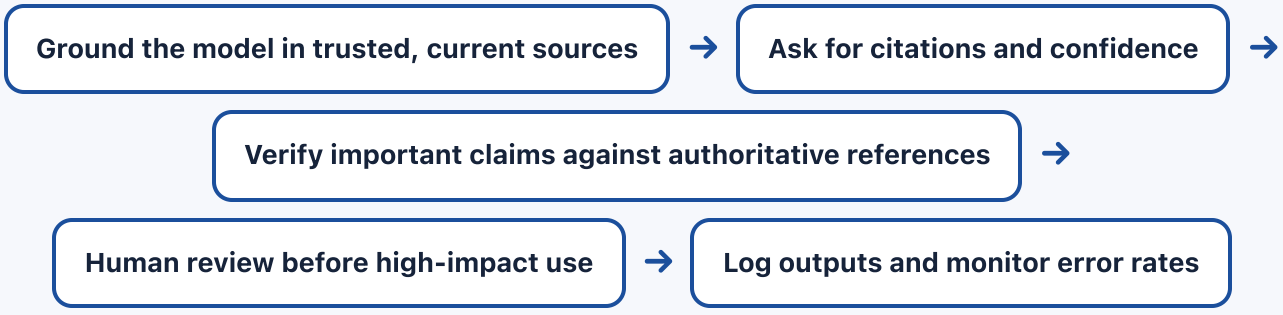
Organizations are using generative AI to draft contracts, answer customer questions, summarize records, write code, and support decisions. A confident falsehood in any of those settings can cause real harm: a wrong medical or legal statement, a fabricated citation in a filing, an insecure code snippet, or a misleading answer to a customer. Because the output looks polished, reviewers can be lulled into accepting it. In regulated work, a decision based on a hallucination

can create liability and compliance exposure, and it can be hard to explain after the fact. For governance, risk, and compliance professionals, hallucination is a core AI reliability risk that must be documented, controlled, and tied to how much a given use case can tolerate an error.

How It Works

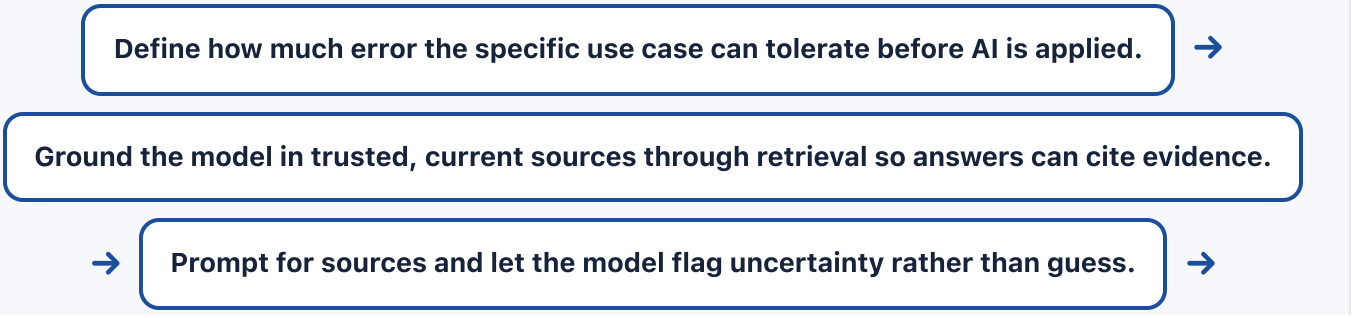
A language model turns an input into a sequence of likely words drawn from patterns in its training, with no built-in sense of truth. When the training data is thin, outdated, or contradictory, or when the question pushes beyond what the model reliably knows, it still produces a fluent answer, and that answer may be invented. The risk rises when users treat the model as a search engine or a source of record. Reducing hallucinations means grounding the model in trusted, current information it can cite, asking it to show its sources, keeping tasks within its reliable range, verifying important outputs against authoritative references, and keeping a human in the loop for anything consequential. The goal is to lower the rate and, more importantly, to catch errors before they are trusted.

Architecture Diagram



Ground the model, ask for sources, verify important claims, and keep a human in the loop before output is trusted.

Visual Workflow



Verify important claims and any citations against authoritative references. →

Require human review before output is used in consequential or regulated work. →

Log outputs, track error rates, and refine the approach over time.

Common Attacks

- Fabricated citations, sources, or quotes that do not exist
- Confident false statements of fact presented without any caveat
- Misattribution that assigns real information to the wrong person or source
- Summaries that drift from or contradict the underlying document
- Invented figures, dates, or details that fill a gap in the model's knowledge

Common Mistakes

- Treating a generative model as a search engine or an authoritative source of record
- Accepting fluent, confident output without verifying the claims
- Skipping grounding and letting the model answer from memory alone
- Using AI for high-stakes decisions with no human review
- Never measuring how often the system is actually wrong

Best Practices

- Match the use case to the model's reliability and the tolerance for error
- Ground answers in trusted, current sources through retrieval
- Require citations and verify important claims against authoritative references
- Prompt the model to express uncertainty instead of guessing
- Keep a human in the loop for consequential or regulated output
- Measure and monitor error rates and improve prompts, sources, and controls over time

Quick Checklist

- [x] Error tolerance is defined for each AI use case
- [x] Important answers are grounded in trusted, current sources
- [x] Citations are required and verified for high-stakes output
- [x] Human review is required before consequential use
- [x] Users are told the output can be wrong and must be checked
- [x] Error rates are measured and tracked over time

Recommended Tools

Retrieval augmented generation (RAG) framework
Grounds answers in trusted documents the model can cite

Output validation and fact-check tooling
Checks claims and citations before output is trusted

LLM evaluation and testing framework
Measures accuracy and hallucination rate across cases

Human review workflow
Routes high-impact output to a person before it is used

Industry Standards

NIST AI Risk Management Framework (AI RMF, AI 100-1)
Its Measure and Manage functions guide testing and controlling AI reliability

ISO/IEC 42001
Standard for an AI management system covering quality and oversight of AI output

OWASP Top 10 for LLM Applications
Addresses overreliance on unverified model output as an application risk

Career Relevance

Hallucination is essential knowledge for AI risk managers, who define acceptable error and controls per use case, and for AI governance analysts, who document reliability and oversight. GRC analysts assess it when reviewing AI-assisted processes for accuracy and explainability, and AI security engineers help build grounding and verification into applications. Because it directly shapes whether an AI system can be trusted in regulated work, it is a recurring theme for the AI-Governance-Jobs.com audience.

Interview Questions

- What is an AI hallucination, and why do language models produce them?
- Why does confident, fluent output make hallucinations especially risky?
- How does grounding through retrieval reduce hallucinations?
- What role does human review play for high-stakes AI output?
- How would you measure the hallucination rate of an AI system in production?

Related Certifications

ISO/IEC 42001 lead implementer training (as available)

ISACA AI governance offerings (as available)

IAPP AI governance certification (as available)

Further Reading

- NIST AI Risk Management Framework
- OWASP Top 10 for LLM Applications
- NIST Trustworthy and Responsible AI

Key Takeaways

- Hallucinations are confident output that is false, invented, or unsupported.
- They happen because models predict likely text, not verified facts.
- Fluent, authoritative tone invites unearned trust, which is the real danger.
- Grounding, citations, verification, and human review reduce and catch them.

- It is a reliability risk that AI governance, risk, and GRC roles must manage per use case.

FAQ

▶ **Can hallucinations be eliminated completely?**

▶ **Does retrieval augmented generation fully fix hallucinations?**

▶ **How is a hallucination different from a normal software bug?**

Related Careers

RELATED ROLES

Ai Risk Manager

Ai Governance Analyst

Grc Analyst

Ai Security Engineer

RELATED CERTIFICATIONS

ISO/IEC 42001 lead implementer training (as available)

ISACA AI governance offerings (as available)

IAPP AI governance certification (as available)

CURRENT OPENINGS

Live roles are on the job board.

Browse all jobs

GUIDES & SALARY

Career guides

Role roadmaps

Salary data

Career tools

SUGGESTED LEARNING PATH

1. Ground the basics with CS-001 Cybersecurity
2. Study this sheet: **AI Hallucinations**
3. Go deeper: Secure AI Adoption

4. Go deeper: OWASP Top 10 for LLM Applications
5. Validate it: work toward **ISO/IEC 42001 lead implementer training (as available)**
6. Find the role: browse current openings

RELATED SHEETS

CS-115 Secure AI Adoption
CS-116 OWASP Top 10 for LLM Applications
CS-111 Prompt Injection
CS-001 Cybersecurity
CS-104 NIST Cybersecurity Framework (CSF)