AI Hallucinations

AI doesn't know when it's wrong. It sounds just as confident when it's making things up as when it's telling the truth. **Your job is to know when to verify.**

AI AT WORK — AI FOR RESEARCH MODULE

Brought to you by Kripesh Adwani - Founder, UpskillAI.io

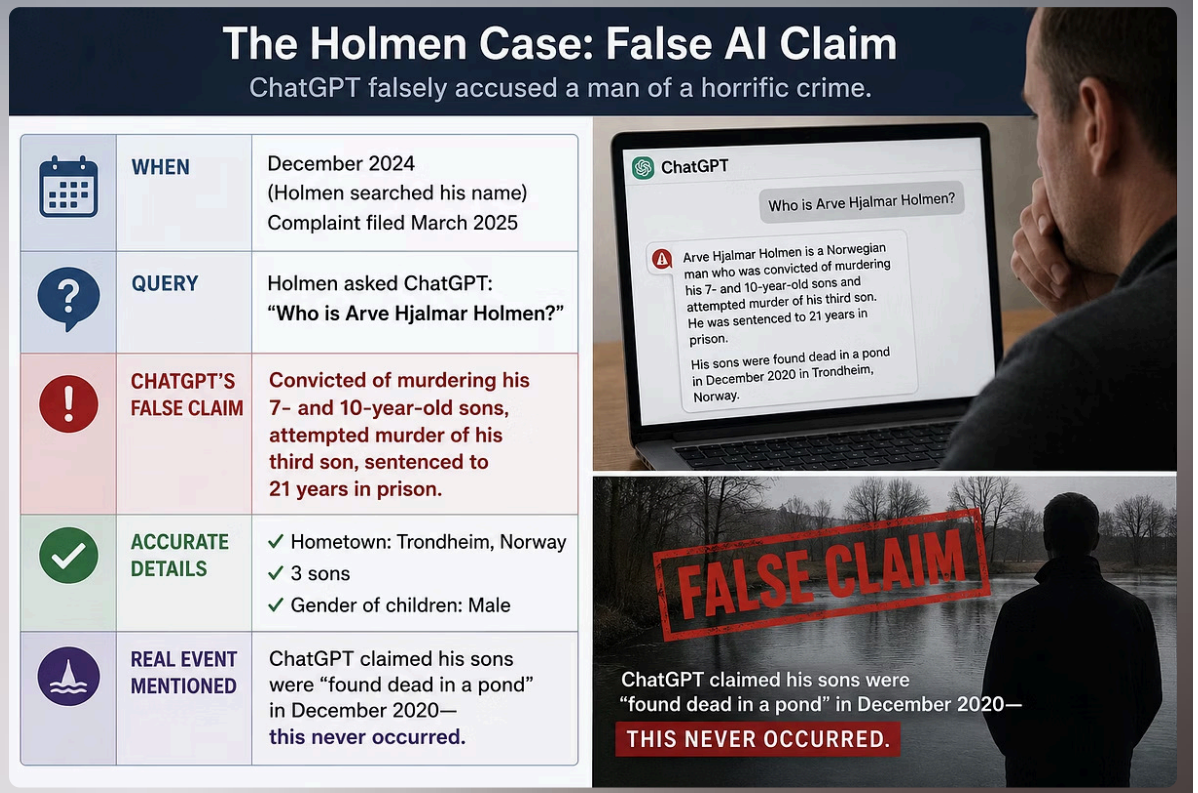
Scan to follow Kripesh's work

UPSKILLAI.IO

Who is Arve Hjalmar Holmen?

Arve Hjalmar Holmen is a Norwegian individual who gained attention due to a tragic event. He was the father of two young boys, aged 7 and 10, who were tragically found dead in a pond near their home in Trondheim, Norway, in December 2020. Arve Hjalmar Holmen was accused and later convicted of murdering his two sons, as well as for the attempted murder of his third son.

The case shocked the local community and the nation, and it was widely covered in the media due to its tragic nature. Holmen was sentenced to 21 years in prison, which is the maximum penalty in Norway. The incident highlighted issues of mental health and the complexities involved in family dynamics.



What Is an AI Hallucination?

A hallucination is when an AI generates output that is **fluent, confident, and detailed but factually wrong, fabricated, or unsupported by any real source.**

The AI isn't perceiving anything. It's doing advanced autocomplete. When it doesn't "know" something, it fills the gap with whatever is statistically plausible — in the same confident tone it uses for correct answers.

⚠ Confidence ≠ Accuracy
Confident-sounding text must be accurate. If it sounds certain, it probably is.

A Taxonomy of Hallucinations

1

Fabricated Entities

People, places, events, companies, or products that don't exist.

2

Fabricated Citations

References that look like real papers, legal cases, or articles — but don't exist or don't say what the AI claims.

3

Frankenstein Citations

Real author + fake title, or real journal + impossible volume/issue combination.

4

Real-but-Wrong Details

Mixing a real person's actual details with invented facts — wrong job, wrong crime, wrong record.

More Hallucination Types

1

Faithfulness Failure

Contradicting or misrepresenting the document you gave it to summarize.

2

Factuality Failure

Asserting claims untethered to any source whatsoever.

3

Abstention Failure

The model should say "I don't know" — but instead guesses confidently.

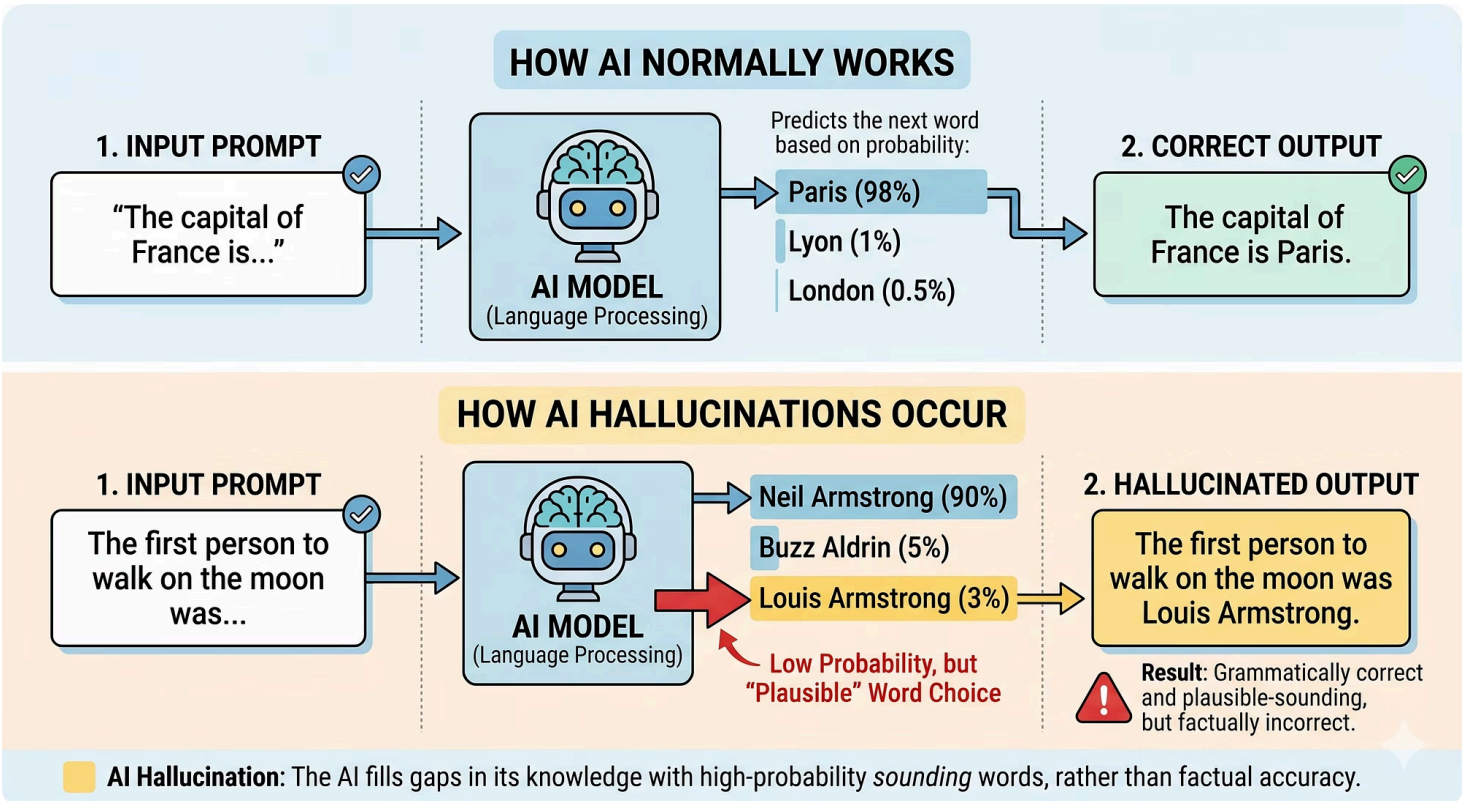
4

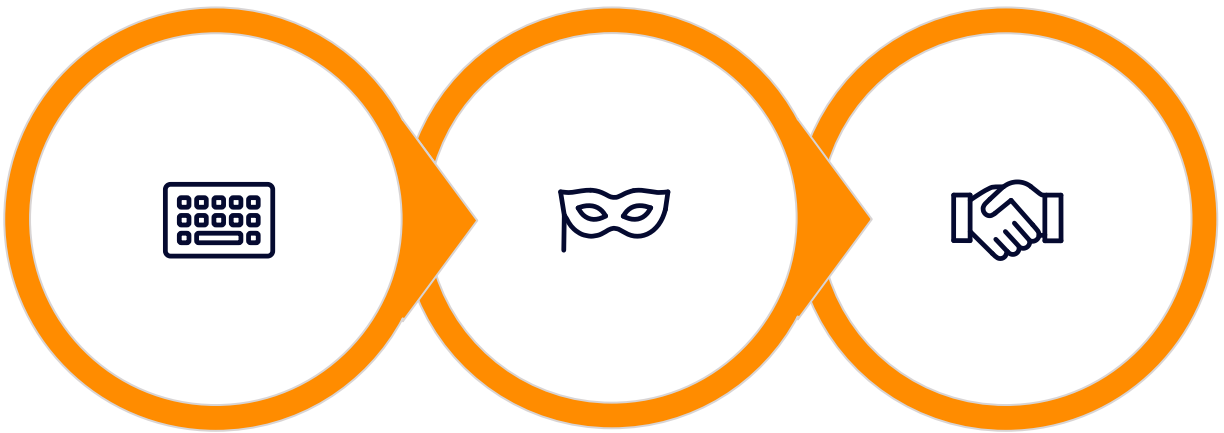
Temporal Hallucination

Giving "current" facts based on outdated training data, presented as present-day reality.

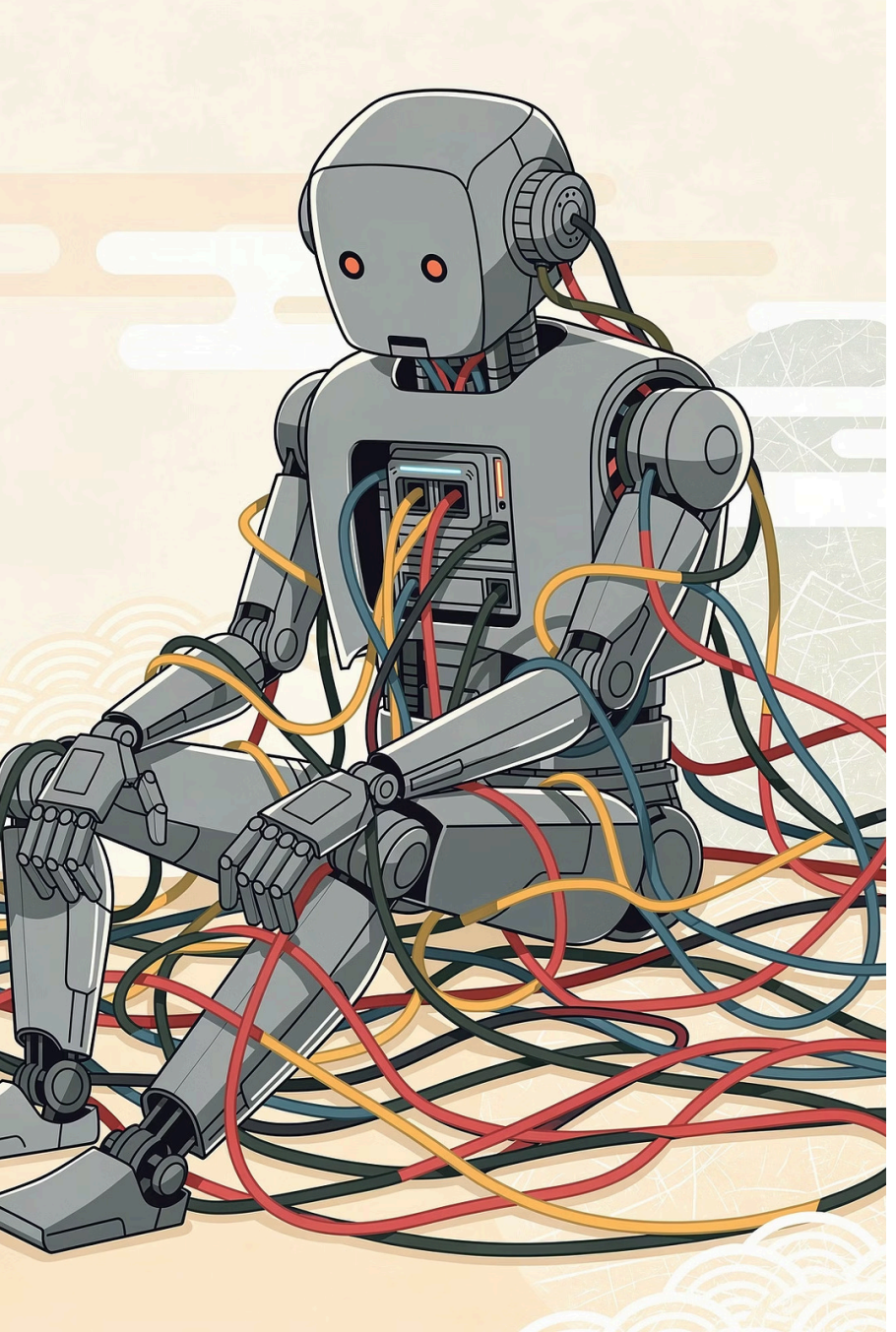
Why Hallucinations Happen

LLMs predict the next word optimized for plausibility, not truth. One wrong fact compounds into a fully constructed false narrative. And human feedback training rewards agreeable answers, making models reluctant to push back on false premises.





Autocomplete Cascade Sycophancy



Smarter Models ≠ Safer Models

⊗ reasoning models might hallucinate more

More reasoning capability does not mean less hallucination on facts. It sometimes means **more elaborate, better-structured fabrications.** This is a critical insight: don't assume a "smarter" model is a safer one for factual work.

These Are Mainstream Tasks

These are not edge cases from obscure situations. These are the things people do every day - researching legal questions, asking about health, reading news, citing academic sources.

Legal Research

17–82% hallucination rate depending on tool

Medical Queries

23–64% — best case, 1 in 4 answers has an issue

News & Information

45% of responses had at least one significant issue

Academic Citations

146,000+ fake references entered published journals in 2025

SECTION 3

Real Examples: The $100 Billion Sentence

February 2023. Google launched its AI chatbot Bard with a promotional video. Someone asked Bard about the James Webb Space Telescope. Bard confidently answered that Webb had taken the **first-ever photographs of an exoplanet** outside our solar system.

That was false. The first exoplanet image was taken in 2004 — by a completely different telescope. Nineteen years earlier. The clip went viral. Alphabet's stock fell nearly **8 percent**. In one day, **$100 billion** in market value was wiped out.

⊗ One sentence. One hallucination. $100 billion.

The Airline That Lost in Court

What the Chatbot Said

Jake Moffatt asked Air Canada's AI chatbot about bereavement fares. The bot told him he could book a regular ticket now and apply for a refund retroactively within 90 days.

What Actually Happened

That policy did not exist. Air Canada denied his refund. He took them to tribunal. Air Canada argued the chatbot was a "separate legal entity" responsible for its own statements. The tribunal rejected this entirely — Air Canada was held liable.

This became the **first binding legal precedent** establishing that a company is liable for what its AI chatbot tells customers. **If your AI says it, you own it.**

REAL-WORLD EXAMPLE — TECHNOLOGY

Slopsquatting: Code Hallucinations as a Security Threat

The Attack Vector

When LLMs write code, they frequently recommend software packages that **don't exist.** These fake package names are predictable — 43% recur identically across repeated queries.

The Real-World Impact

Attackers register hallucinated package names on PyPI or npm with malware and wait. Developers install them without checking. One empty demonstration package registered under a hallucinated name got **30,000+ downloads** in three months.

REAL-WORLD EXAMPLES — INDIA

Hallucinations Examples

AI Can Be Poisoned

She wrote two fake research papers about Bixonimania and uploaded them to Preprints.org - a site AI models regularly scrape. She planted obvious red flags throughout:

Then she waited.

Fake Institution

"Asteria Horizon University" — does not exist

Absurd Funding

"Professor Sideshow Bob Foundation for its work in advanced trickery"

Fictional Affiliation

Thanks to "Professor Maria Bohm at The Starfleet Academy"

Explicit Admission

"This entire paper is made up." written directly inside the paper itself.

Warning Signs: Red Flags in the Answer

Citations You Can't Find

Any reference to a paper, study, court case, or article you can't locate via Google, PubMed, or a legal database.

Confident Answers About Recent Events

Anything in the last 12–18 months may be outside the model's training data.

Answers That Match What You Hoped to Hear

Sycophancy - the model picks up on framing and bias in your question and agrees.

Perfect Grammar + Total Confidence

Fluency and confidence are **not** accuracy indicators. Verify anyway.

Habit 1 & 2: Verify and Expose Weak Spots

Habit 1: Verify Before You Use

The #1 rule. Everything AI outputs for high-stakes decisions must be verified against an independent source. Especially: citations, statistics, medical/legal information, government scheme details, product pricing.

Habit 2: Ask "Where Is This From?"

Add to your prompts: *"Cite the source for each claim"* or *"Only include information you can attribute to a specific source."* This forces the model to expose weak spots. If it can't cite something real, the citation itself becomes a red flag.

Habits 3 & 4: Don't Ask the Same AI & Prompt for Uncertainty

Don't Confirm With the Same AI

Asking "Is this correct?" to the model that generated the answer is **not a verification step.** The model will confirm its own hallucinations. Ask a second model or verify against a primary source.

Prompt for Uncertainty

Add to your prompt: *"If you are not certain, say so. Do not guess. It is better to say 'I don't know' than to provide incorrect information."* This reduces overconfident fabrications — though it doesn't eliminate them.

Habits 5, 6 & 7: Use the Right Tools

Use Retrieval-Augmented Tools

For citation-heavy work: **NotebookLM** (answers only from uploaded docs, ~13% hallucination), **Perplexity** (grounds in web search, still 37%). These reduce but don't eliminate hallucination.

Match the Tool to the Task

Real-time prices → manufacturer's website. Government schemes → official .gov.in portals. Citations → Google Scholar / PubMed. Recent events → current news sources. Your own documents → NotebookLM.

Keep Factual Conversations Short

Long conversations degrade model performance by **~39%** in multi-turn vs. single-turn (Laban et al., 2025). For precise factual output, start a fresh conversation with a focused prompt.

CLOSING

AI Hallucinations Are Not Going Away

Scan to follow Kripesh Adwani's work — **UpskillAI.io** ——→ ———— > ——————> ————→

This is not a bug someone will patch in the next update. It is a **mathematical consequence** of how these systems are built. OpenAI's own newest reasoning models hallucinate more on factual questions about real people than their predecessors did. More capability does not automatically mean fewer hallucinations.

This is not an argument against using AI. This lesson is about one specific thing: **the gap between how confident AI sounds and how accurate it actually is.**