# Why AI Gets Things Wrong

> **Plain English:** AI produces confident-sounding text that is sometimes completely wrong. This isn't a bug — it's a fundamental feature of how these systems work. Knowing why it happens is the single most important thing you can learn about working with AI.

## What's actually happening

When AI writes something incorrect, people call it a "hallucination." The word is a bit misleading — it implies the AI is seeing things that aren't there, like a glitch. What's actually happening is simpler and more important to understand:

AI generates the most statistically likely next words based on patterns in its training data. It has no mechanism to check whether what it's producing is true. It doesn't "know" things the way you know your own phone number. It produces text that *looks and sounds like* correct text, because it learned from millions of examples of correct text.

This means:

- It can write a perfectly formatted citation for a paper that doesn't exist — because it's learned the *pattern* of what citations look like.
- It can confidently state a statistic that's completely fabricated — because it's generating a plausible number in a plausible context.
- It can describe a product feature that was never built — because the description sounds like something that *could* exist.

The AI isn't lying to you. Lying requires knowing the truth and choosing to say something different. AI doesn't know the truth. It's generating plausible text. That's a crucial distinction.

## When it's most likely to go wrong

Hallucinations aren't random. They follow predictable patterns:

**Specific facts, numbers, and dates.** Ask AI for a general explanation of how photosynthesis works and it'll be accurate. Ask it for the exact year a specific obscure paper was published and it might invent one. The more specific and verifiable the claim, the more you need to check it.

**Citations and sources.** AI is particularly bad at this. It will confidently produce author names, paper titles, journal names, and URLs that look real but don't exist. Never trust an AI-generated citation without verifying it.

**Recent events.** AI models have a training cutoff date. If you ask about something that happened after that date and the AI doesn't have search access, it may either say it doesn't know (good) or generate a plausible-sounding answer (dangerous).

**Niche or specialized domains.** AI performs best on topics that appeared frequently in its training data. Mainstream topics in English have dense coverage. Obscure or specialized topics — especially in other languages — have less, so the model has fewer patterns to draw from and is more likely to fill gaps with plausible-sounding fabrications.

**When you push it.** If you insist the AI answer a question it's uncertain about, or tell it "you must provide an answer," it will comply — by generating something. AI tools generally don't have a strong instinct to say "I don't know." Some are better than others, but the pressure to produce output is built into the system.

# Why it *sounds* so confident

This is the part that trips people up. When a human says something confidently, you assume they believe it and probably have some basis for it. When AI says something confidently, it means nothing — confidence is the default mode.

The model isn't more certain about accurate statements than inaccurate ones. It generates all text with the same fluent, authoritative tone because that's the pattern in its training data. Well-written text sounds confident. The model produces well-written text. Therefore, everything it produces sounds confident — including the wrong things.

This is why the philosopher Harry Frankfurt's essay "On Bullshit" is on our Further Reading page. Frankfurt distinguishes between lying (knowing the truth and hiding it) and bullshitting (not caring whether something is true). AI is, technically, the world's most sophisticated bullshit generator. Not because it's trying to deceive you, but because truth and falsehood aren't categories it operates in. It operates in plausibility.

# What you can do about it

The good news: once you understand this, you can work with it effectively. The Fact-Check Habit and Verification Checklist exercises build these skills in practice. Here's the framework:

**Treat AI output as a first draft, not a final answer.** This shift in mindset is the most important thing. AI gives you a starting point — fast, broad, often useful. Your job is to validate, refine, and

apply judgment.

**Cross-reference specific claims.** If the AI states a fact, a statistic, or a date that matters to your work, verify it with a primary source. This takes 30 seconds and prevents the kind of embarrassing errors that erode trust.

**Watch for the patterns above.** You now know when hallucinations are most likely. Apply extra scrutiny to specific facts, citations, recent events, and niche topics.

**Ask the AI to flag its uncertainty.** Adding "If you're not sure about something, say so" to your prompt doesn't guarantee honesty, but it does help some models hedge appropriately rather than fabricating with confidence.

**Use AI features that ground responses in sources.** Tools like Perplexity, Claude with web search, or ChatGPT with browsing can cite where they found information. This doesn't eliminate errors, but it gives you something to check against.

**Don't ask AI to be your only source.** AI is best when it's one input among several — when you're using it alongside your own expertise, your colleagues' perspectives, and verified data. The Multi-Source Brief exercise practices exactly this.

# The bigger picture

Understanding hallucinations isn't just about catching mistakes. It fundamentally shapes how you think about AI's role in your work:

- It's why the Ethical Prompting & Judgment pillar exists — because using AI responsibly requires knowing its limitations.
- It's why human oversight matters in any AI workflow — and why the most advanced exercises in this playbook always include human checkpoints.
- It's why AI fluency is more than just "knowing how to prompt" — it's knowing when to trust, when to verify, and when to override.

# Where to go next

- The Fact-Check Habit — practice catching AI mistakes in 15 minutes
- Prompt Engineering Basics — techniques that reduce (but never eliminate) hallucinations

---

Revision #1
Created 2026-03-16 15:47:58 UTC by Admin
Updated 2026-03-16 15:47:58 UTC by Admin