WORLD ASSESSMENT COUNCIL

MODULE 1 · HOW AI ACTUALLY WORKS

Why AI makes things up, and how to stop it fooling you

Six court cases with perfect names and quotes. None of them existed.

Lesson 1.6 | About 10 minutes | Centre for AI-Powered Learning

DEFINITION

What is a hallucination?

- A **confident** statement that is **false**
- Invented facts, fake quotes, **non-existent citations**
- Wrong dates, quietly wrong **calculations**
- Danger: it sounds **exactly as sure** as when it's right

Invented facts
Wrong birthplace, awards

Fake quotes
Words never said

Fake citations
Papers that don't exist

Wrong numbers
Quietly off by a step

WHY IT HAPPENS

It predicts what is plausible, not what is true.

A fake case name that looks like real ones is extremely plausible. Nothing in the loop checks truth.

DO YOU KNOW?

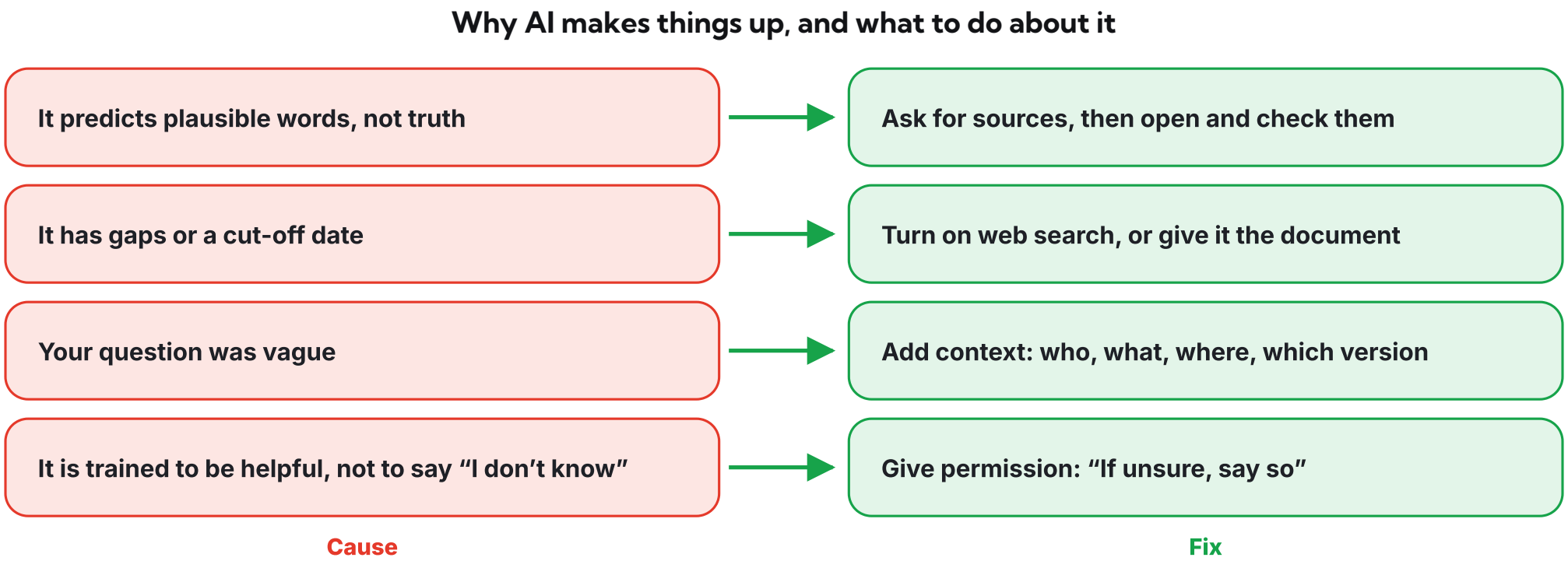
Tests that only count right answers reward guessing. In one test, a model that rarely said "I don't know" got 75% of answers wrong.

A model that declined half the time cut errors to 26%. Source: OpenAI, September 2025



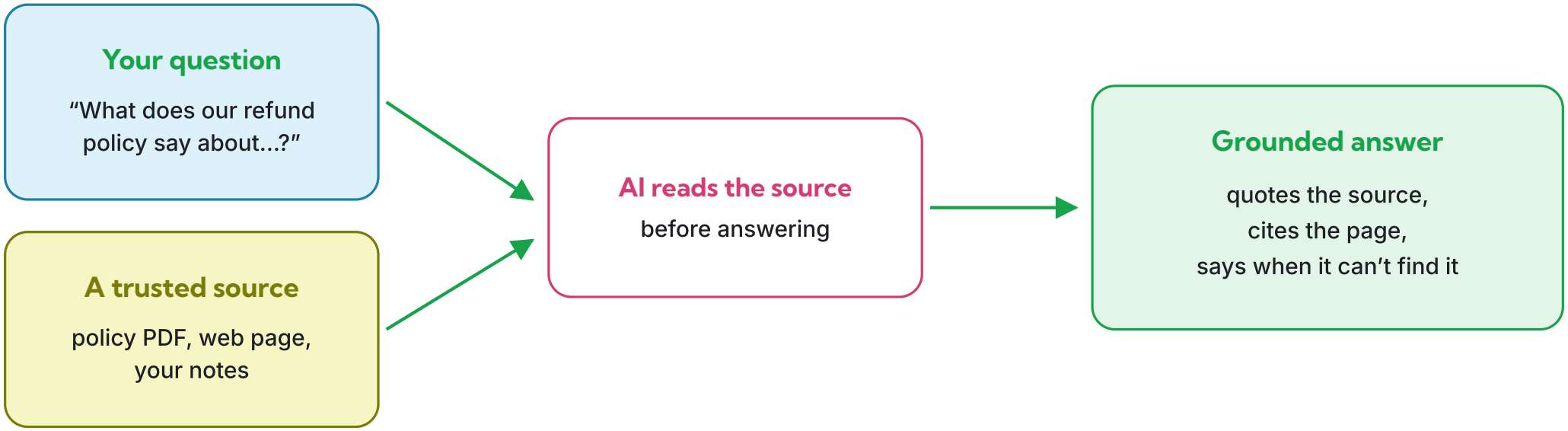
FOUR CAUSES, FOUR FIXES

Why AI makes things up, and what to do



THE STRONGEST FIX

Grounding: answer from a trusted source



REUSABLE PROMPT

The grounding prompt

Attach the document, then use:

```
Answer using only the document I've attached. For each point, quote the exact sentence and give the page or section. If the document doesn't answer the question, say "Not in the document". Don't use outside knowledge.
```

VERIFICATION HABITS

The three checks

1. **Check the source**: open it; does it say that?
2. **Check a second source**: does an independent one agree?
3. **Check the maths**: redo it, or ask for code
4. Always for names, numbers, dates, **quotes, citations**

CITATIONS

Never cite anything you haven't opened yourself.

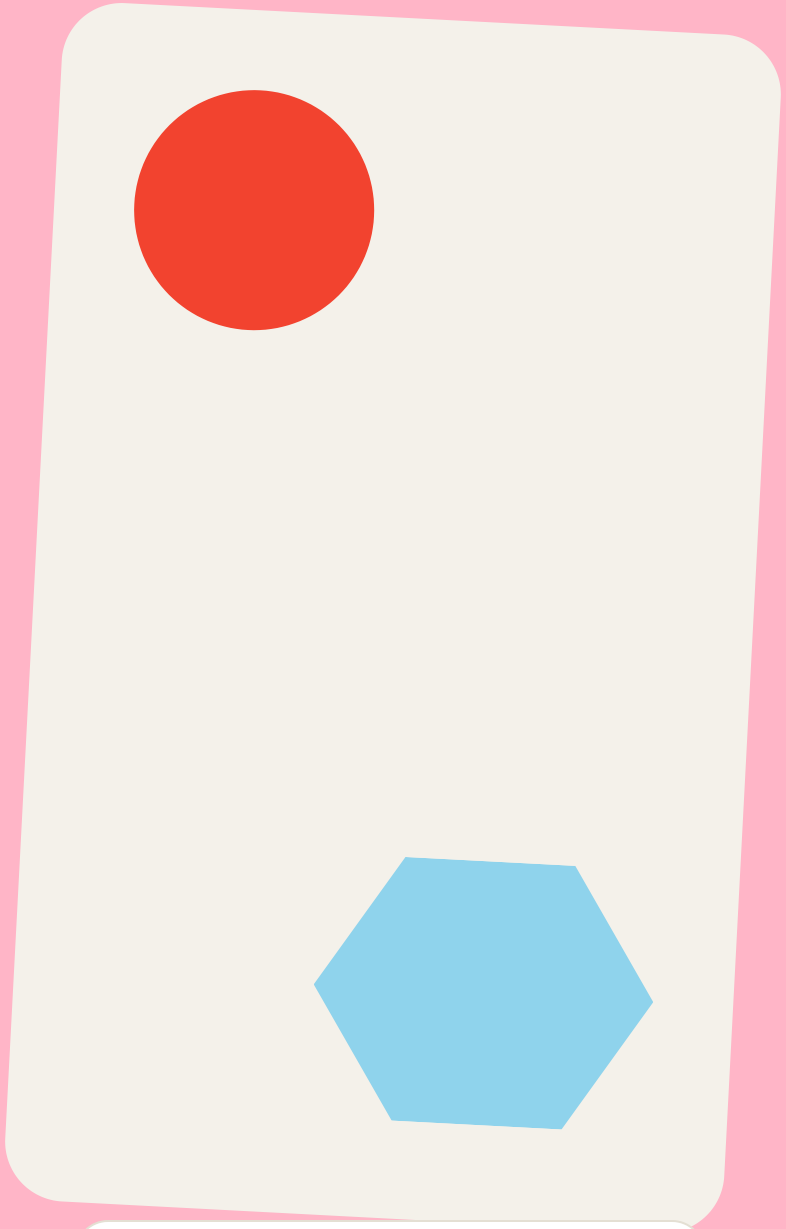
Invented papers with real-sounding authors and journals are among the most common hallucinations.

AI STORY · 2023

Mata v. Avianca: six cases that never existed

The court asked for copies of the cited cases. They couldn't be produced. The lawyers stood by them anyway, and in June 2023 were fined $5,000 for bad faith.

The error became a disaster because nobody opened the sources.



WORLD ASSESSMENT COUNCIL

ANOTHER TRAP

AI can be persuaded

Pushing back

- Confidently challenge a right answer
- Some models **apologise and switch**
- Agreement is **cheap**

Leading questions

- "Why is this policy **bad**?" gets only negatives
- Ask **neutrally**: "What are the pros and cons?"
- Ask it to **argue the other side**

LAB 1

Catch a hallucination, then fix it

1. **Provoke**: ask for 3 papers (search off); check each on Google Scholar
2. **Ground**: upload a real document; ask 5 questions, 1 it can't answer
3. **Compare**: a recent-event question, search off vs on
4. **Report**: what you found and which fix worked best

QUICK CHECK

What does your university's policy say about AI use? Best approach?

A Ask the assistant from memory

B Ask a friend's chatbot

C Give it the official policy, ask it to answer only from it with quotes, then check

D Search social media

KEY TAKEAWAYS

Remember this

- Hallucination = **confident falsehood**
- Plausible ≠ true
- Fixes: **sources, search/documents, context, permission to say "I don't know"**
- **Grounding** is the strongest fix
- Verify **names, numbers, dates, quotes, citations**