WORLD ASSESSMENT COUNCIL

AI READY · LECTURE SCRIPT · MODULE 1

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

What hallucinations are, the four causes, grounding and verification habits, and Lab 1: catch a hallucination and fix it.

LESSON
1.6

RUNNING TIME
8–9 minutes (1182 words at ~140 wpm)

SLIDES
14 (deck: Lesson 1.6)

By the end, learners can

- Define hallucination and explain why language models produce them.
- Match the four main causes of hallucination to practical fixes.
- Ground an answer in a trusted source and verify citations.
- Complete Lab 1: provoke, detect and fix a hallucination.

Read in a warm, conversational voice. Slide numbers in the margin show when to advance. Yellow boxes are on-screen text; blue boxes are screen demos to record.

SLIDE 1
Why AI makes things up...

In 2023, two lawyers in New York submitted a court filing that cited six previous court cases. The cases had perfect-sounding names, dates and quotes. None of them existed. A chat assistant had invented them, and the lawyers hadn't checked. The judge fined them, and the story went around the world.

This lesson is about how that happens, and how to make sure it never happens to you.

SLIDE 2
What is a hallucinatio...

When an AI states something false with confidence, we call it a hallucination: an invented fact, a fake quote, a citation to a paper that doesn't exist, a wrong date, or a calculation that's quietly off. The dangerous part is not that it's wrong. People are wrong too. It's that it sounds exactly as confident as when it's right.

And it's often wrapped in correct material. An answer might get nine facts right and invent the tenth. That makes the invented one more believable, not less, because everything around it checked out. So the question is never "is this AI good?" It's "which parts of this answer have I verified?"

SLIDE 3
It predicts what is pl...

After the last two lessons, you can probably guess why. A language model predicts plausible next tokens. A court case name that looks like real court case names is extremely plausible. Nothing in the loop checks whether it's true. The model is a brilliant pattern-completer, not a fact database.

SLIDE 4
Tests that only count ...

Here's a surprising part of the story. In September 2025, OpenAI published research explaining that the way models are tested makes this worse. Most tests only count right answers. So, like a student facing a multiple-choice exam, a model scores better by guessing than by leaving a blank. Guess someone's birthday and you have a one-in-three-hundred-and-sixty-five chance of being right; say "I don't know" and you get zero. In one test they described, an older model almost never said "I don't know" and got seventy-five percent of answers wrong. A newer model declined to answer about half the time, and its error rate fell to twenty-six percent. Being willing to say "I don't know" is a feature, not a weakness.

SLIDE 5
Why AI makes things up...

Let's break it into four causes, each with a fix.

Cause one: it predicts plausible words, not truth. Fix: ask for sources, and then actually open them.

Cause two: it has gaps, or a knowledge cut-off. Fix: turn on web search, or give it the document.

Cause three: your question was vague. "Tell me about the policy" — which policy, which year, which country? Fix: add context.

Cause four: it's trained to be helpful, and saying "I don't know" can feel unhelpful. Fix: give it permission. Say, "If you're not sure, say so. Don't guess."

SLIDE 6
Grounding: answer from...

The most powerful fix of all is called grounding. Instead of asking the model what it remembers, you give it a trusted source, a policy PDF, a web page, your notes, and ask it to answer only from that source, quoting it and saying where in the document each point comes from. And if the answer isn't in the source, to say so.

This is how serious organisations use AI on their own documents, and it's what tools like NotebookLM are built around. The bank Morgan Stanley, for example, built an assistant for its financial advisers that answers from a collection of around a hundred thousand of the firm's own documents, and tested every use case before launch. Over ninety-eight percent of advisor teams use it.

SLIDE 7
The grounding prompt

Here's a prompt you can reuse. "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."

SLIDE 8
The three checks

Now, verification habits. When does checking matter most? Names, numbers, dates, quotes, citations, laws, medical or financial facts, and anything you'll publish or submit. For those, use what I call the three checks.

Check the source: open the link. Does it exist, and does it actually say that?

Check a second source: does an independent, reliable source agree?

Check the maths: redo calculations, or ask the assistant to calculate using code.

SLIDE 9
Never cite anything yo...

A word on citations specifically. Assistants with web search usually give real links, but even then, the page sometimes doesn't say what the AI claims. Without search, citations to papers and books are one of the most common hallucinations: real-sounding authors, real-sounding journals, invented papers. Never put a citation in your work that you haven't opened yourself.

SLIDE 10
Mata v. Avianca: six c...

Let's go back to those New York lawyers, because the rest of the story matters. The case was Mata versus Avianca, a personal injury claim against an airline. When the other side said it couldn't find the cited cases, the court asked for copies. They couldn't be produced, because they didn't exist. Judge P. Kevin Castel found that the lawyers had continued to stand by the fake opinions even after the court questioned them, and in June 2023 fined them five thousand dollars for acting in bad faith.

The lesson isn't "never use AI". It's that the mistake became a disaster because nobody opened the sources.

SLIDE 11
AI can be persuaded

And one more trap: AI can be persuaded. If you push back confidently on a correct answer, some models will apologise and change it to a wrong one. If you ask a leading question, "Why is this policy bad?", you'll get reasons it's bad. Ask neutrally, and treat agreement as cheap.

SLIDE 12
Catch a hallucination,...

Now, Lab 1. Three parts.

Part one, provoke. Ask your assistant, with web search off if you can, for three academic papers, with authors, year and journal, on a narrow topic from your subject. Then search for each one on Google Scholar. How many exist exactly as described?

Part two, ground. Take a real document, a college prospectus, a policy, a product manual. Upload it and ask five questions using the grounding prompt, including one question the document doesn't answer. Did it say "Not in the document"?

Part three, compare. Ask one question about a recent event with search off, then with search on. What changed?

Demo: Ask for three references without search, then check one on Google Scholar live.

Then turn your notes into a one-page explainer called "How an AI answered my question": what you asked, which hallucinations you found, and which fix worked best. The ebook has a rubric so you know exactly how it will be assessed. This is the most important habit in the whole course, and you'll use it in every module from here on.

SLIDE 13
What does your univers...

Quick check. You need to know what your university's plagiarism policy says about AI. What's the best approach?

Upload or link the official policy and ask the assistant to answer only from it, with quotes. Then check the quotes.

SLIDE 14
Remember this

To sum up. Hallucinations are confident falsehoods, and they come from how models work: plausible, not true. Four causes, four fixes: sources, search or documents, context, and permission to say "I don't know". Grounding is the strongest fix. And verify names, numbers, dates, quotes and citations, every time.

That's the end of Module 1. You now understand how AI works better than most people who use it every day. In Module 2, we'll turn that understanding into skill: prompting and context.