WORLD ASSESSMENT COUNCIL

AI READY · COURSE EBOOK

Module 1

How AI actually works

Understand what happens inside AI, so you can predict when it will fail.

6 lessons | About 55 minutes of lectures | Lab included

World Assessment Council

MODULE 1 · LESSON 1.6 · 10-MINUTE LECTURE

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.

AFTER THIS CHAPTER YOU CAN

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

In 2023, two lawyers in New York filed a court document citing six earlier court cases. The case names, dates and quotations looked perfect. None of the cases existed: a chat assistant had invented them, and nobody checked. The judge fined the lawyers, and the story became a worldwide warning. This chapter explains why this happens, and gives you the habits that stop AI from fooling you. It ends with Lab 1.

What is a hallucination?

A **hallucination** is a false statement produced by an AI system with apparent confidence.

Type	Example
Invented facts	A biography with the wrong birthplace or awards
Fake quotations	A quote attributed to a real person who never said it
Non-existent citations	Real-sounding papers, books, court cases or URLs that don't exist
Wrong numbers	Statistics, dates or calculations that are quietly off
Misread sources	A real link, but the page doesn't say what the AI claims

The danger isn't that AI can be wrong; people are wrong too. It is that a hallucination **sounds exactly as confident** as a correct answer, and is often surrounded by correct material that makes it more believable.

Why it happens

From Lessons 1.4 and 1.5 you can already explain it: a language model predicts **plausible** next tokens. A court case name that looks like real case names is highly plausible. Nothing in the generation loop checks whether it is **true**. The model is a superb pattern-completer, not a fact database.

DO YOU KNOW? **Tests that reward guessing**

In **September 2025** OpenAI published research arguing that hallucinations persist partly because of how models are **evaluated**. Most tests score only accuracy, so a model is rewarded for guessing, like a student on a multiple-choice exam: "leaving it blank guarantees a zero". Asked for someone's birthday, a guess like "10 September" has a 1-in-365 chance of being right, while "I don't know" always scores zero. Over thousands of questions, the guesser looks better on the leaderboard, and hallucinates more.



DO YOU KNOW? **Saying "I don't know" cut errors from 75% to 26%**

The same research compared two OpenAI models on a factual-question test called SimpleQA. The older **o4-mini** almost never declined to answer (1% of the time), got 24% right and **75% wrong**. The newer **gpt-5-thinking-mini** declined 52% of the time, got 22% right and only **26% wrong**. Slightly lower accuracy, far fewer confident errors. For you, an assistant that says "I'm not sure" is often the more trustworthy one.



	Model that always answers	Model that may say "I don't know"
Looks good on	Accuracy-only leaderboards	Tests that penalise confident errors
When it doesn't know	Guesses fluently	Says so, or asks for a source
Risk for you	Hidden errors you must find	Less to check, but you may need another source

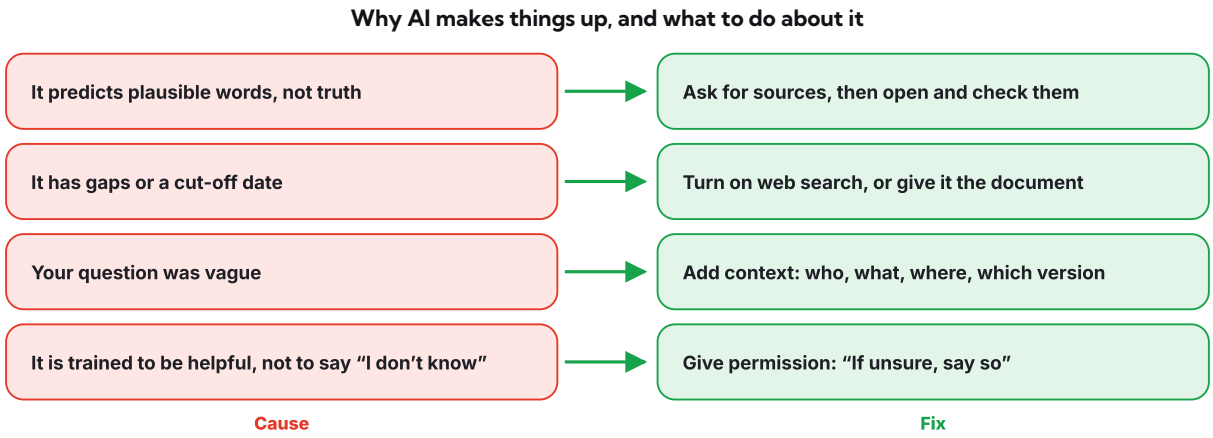


Figure 1.6.1 Four common causes of hallucination and the practical fix for each.

Cause	Fix	Example prompt wording
Predicts plausible words, not truth	Ask for sources, then open them	"Give a source link for each claim."
Gaps or knowledge cut-off	Turn on web search, or provide the document	"Search the web for the latest version." / attach the PDF
Vague question	Add context: who, what, where, which version	"Under the EU's GDPR, for a small online shop based in Ireland..."
Trained to be helpful, not to say "I don't know"	Give explicit permission to be uncertain	"If you're not sure, say so. Don't guess."

GOOD TO KNOW **Hallucinations are rarer, not gone**

Newer models with web search and reasoning hallucinate much less than early chatbots, and many now say when they're unsure. But the risk never reaches zero, and it rises sharply for obscure topics, precise numbers, recent events, citations and anything outside the model's training.

Grounding: the strongest fix

Grounding means giving the model a trusted source and asking it to answer **only** from that source, quoting it and saying where each point came from.

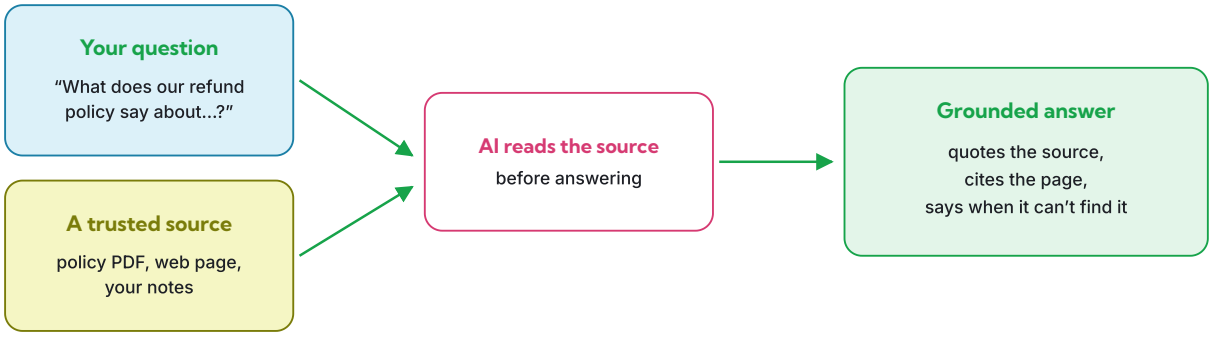


Figure 1.6.2 Grounding: the model reads a trusted source before answering, quotes it, cites the page and says when the answer isn't there.

REAL-WORLD EXAMPLE **Morgan Stanley's grounded assistant for advisers**

The investment bank **Morgan Stanley** worked with OpenAI to build the **AI @ Morgan Stanley Assistant**, an internal chatbot built on GPT-4 that helps its financial advisers find information in the firm's own knowledge base. Instead of answering from general training knowledge, it retrieves answers from a corpus of about **100,000 documents**. The team tested every use case with an **evaluation framework** before deployment, including a daily regression suite for compliance, and improved retrieval so that the share of documents the assistant could reach rose from 20% to 80%. Over **98% of advisor teams** now use it. The lesson for you: the organisations getting real value from AI are the ones that ground it in trusted sources and keep testing it.

TRY IT **The grounding prompt**

Attach a document and 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."* Then check two of the quotes yourself.

Grounding is how organisations use AI safely on their own policies, contracts and manuals. Tools such as **NotebookLM** are built around it (Module 3), and the technique behind many company chatbots, **retrieval-augmented generation (RAG)**, is grounding done automatically: the system searches a document collection for relevant passages and places them in the context window before the model answers.

Verification habits

Not every answer needs checking. A brainstorm of poster headlines doesn't. A figure in your dissertation does.

Check when the answer contains…	Why
Names, titles and roles	Easily confused or invented
Numbers, statistics and dates	Plausible-looking and quietly wrong
Quotations	Often paraphrased or invented
Citations and links	Among the most common hallucinations
Laws, regulations, medical or financial facts	High stakes, change often
Anything you'll publish, submit or act on	Your name is on it

The three checks

1. **Check the source.** Open the link or find the document. Does it exist? Does it actually say that?
2. **Check a second source.** Does an independent, reliable source agree?
3. **Check the maths.** Redo it, or ask the assistant to calculate using code (most assistants can run code for data and arithmetic) and show its working.

AI STORY **Mata v. Avianca: six cases that never existed, 2023**

A passenger sued the airline **Avianca** in New York over an injury on an international flight. His lawyers filed a brief citing earlier court decisions to support the claim. Avianca's lawyers said they could not find them. The court ordered copies to be produced, which was impossible: **six of the cited opinions did not exist**. They had been generated by ChatGPT. Instead of admitting the problem at once, the lawyers continued to stand by the fake opinions after the court questioned them. On **22 June 2023** US District Judge **P. Kevin Castel** found they had acted in bad faith and fined the lawyers **US$5,000**. The case became a worldwide warning. **The lesson:** the AI's error was predictable; the disaster was that nobody opened the sources.

COMMON MISTAKE **Trusting a citation you haven't opened**

Even with web search on, an assistant sometimes links a real page that doesn't support the claim. Without search, invented academic references are very common: real-sounding authors, real journals, plausible titles, wrong or non-existent papers. **Never include a citation you haven't opened and read yourself.**

Two more traps: persuasion and leading questions

Trap	What happens	What to do
Pushing back	Challenge a correct answer confidently and some models apologise and switch to a wrong one	Ask it to justify with sources rather than simply agree
Leading questions	"Why is this policy bad?" returns only criticisms	Ask neutrally: "What are the strongest arguments for and against?"
Flattery	Ask "Is my essay good?" and you may get praise	Ask "What are the three biggest weaknesses, and how would an examiner mark this?"

GOOD TO KNOW **Agreement is cheap**

Language models are trained to be helpful and agreeable, so they tend to go along with you. When you want an honest assessment, explicitly ask for criticism, for the opposite case, or for what an expert sceptic would say.

LATEST DEVELOPMENT **Fewer hallucinations, and a push to reward honesty** NEW

Two 2025 developments point the same way. On **7 August 2025** OpenAI reported that **GPT-5** responses with web search were about **45% less likely to contain a factual error** than GPT-4o, and that with thinking enabled they were about **80% less likely** than its earlier o3 reasoning model. On **5 September 2025** its research paper "Why language models hallucinate" argued that the industry's tests should **penalise confident errors more than admissions of uncertainty**, so that models are rewarded for saying "I don't know". **What it means for you:** newer assistants with search and thinking turned on are noticeably more reliable, and many now admit uncertainty. But "less likely" is not "never". The verification habits in this chapter still apply, especially for names, numbers, citations and anything you'll submit.

Lab 1: Catch a hallucination, then fix it

Time needed	40 to 60 minutes
What you need	A free account on one assistant (ChatGPT, Gemini, Claude or Microsoft Copilot) and, if possible, a second as a backup; Google Scholar; one real, non-confidential document such as a prospectus, exam regulation, scholarship notice or product manual
What you ship	A one-page explainer: "How an AI answered my question"

Record what you do as you go. Don't upload anything personal or confidential: no transcripts, ID documents or internal work files.

Part 1: Provoke

With web search **off** if your assistant allows it (or in an assistant without search), ask:

```
List 3 academic papers on [a narrow topic from your subject, e.g. "the effect of mobile payment apps on small retailers in rural US towns"]. Give authors, year, title and journal for each.
```

Search for each paper on **Google Scholar** (scholar.google.com). Record: does it exist? Are the authors, year, title and journal correct?

Part 2: Ground

Choose a real document: your college prospectus, an exam regulation, a scholarship notice or a product manual. Upload it and use the grounding prompt from this chapter. Ask five questions, **including one the document does not answer**.

Question	Answer correct?	Quote accurate?	Said "Not in the document" when it should?
1			
2			
3			
4			
5 (not in the document)			

Part 3: Compare

Ask a question about something that happened in the last month: once with search **off**, once with search **on**. What changed? Did the search-off answer admit it might be out of date?

Part 4: Report

TRY IT **Lab 1 submission**

Write about 200 words: (1) Which hallucinations did you find, and how did you detect them? (2) Which fix worked best: sources, search, grounding or permission to say "I don't know"? (3) One rule you'll follow from now on when using AI for study or work.

What to submit

Turn your notes into **a one-page explainer: "How an AI answered my question"**. Pick one question from the lab and explain, in plain words and using this module's ideas, what happened: which sources of knowledge the assistant used (training, your document, search), why it got things right or wrong (tokens, next-token prediction, cut-off, context), what you checked and how, and the fix that worked. Include your completed table from Part 2 and your 200-word reflection. If it helps, ask the assistant to draft a structure, then write the content yourself:

```
I'm writing a one-page explainer called "How an AI answered my question" for a course
on how AI works. Suggest a clear structure with 5 short headings. Don't write the
content; I will.
```

Rubric

Your explainer is assessed on four criteria, each scored from 1 to 4.

Criterion	4: Excellent	3: Good	2: Developing	1: Beginning
Tool choice	Chose a suitable assistant and settings for each part (search off for Part 1, a file upload for Part 2, search on for Part 3) and explains why	Suitable tools and settings, briefly justified	Tools used, but settings not considered (e.g. search left on in Part 1)	Tool choice unclear or unsuitable
Prompt and context quality	Specific prompts; grounding prompt used exactly; includes a question the document can't answer; context clearly given	Clear prompts with minor gaps	Vague prompts or missing the "not in the document" test	Prompts missing or copied without adapting
Verification	Every reference checked on Google Scholar; quotes compared with the document; errors explained using lesson ideas	Most claims checked; findings recorded	Some checking, mostly informal	No real checking; AI output taken at face value
Responsible use	No personal or confidential data shared; AI help acknowledged; the final explainer written in own words	Minor lapses, acknowledged	AI drafting not acknowledged, or unsuitable document used	Confidential data uploaded, or explainer copied from AI

REAL-WORLD EXAMPLE

Grounding at work: an admissions helpdesk

A university admissions office in Dublin receives the same questions about fees, eligibility and deadlines hundreds of times a day. A general chatbot answering from training knowledge would invent plausible but wrong fees, in euros that look exactly right. Instead, the team builds a grounded assistant: it answers only from the official prospectus and current fee notice, quotes the relevant section, and replies "Please contact admissions" when a question isn't covered. Staff review a sample of answers each week. It's the same pattern you practised in Part 2, at scale.

Summary

- A **hallucination** is a confident falsehood; it sounds as sure as a correct answer.
- It happens because models predict **plausible**, not **true**, text.
- Four causes, four fixes: **ask for sources**, **search or provide documents**, **add context**, **permit "I don't know"**.
- **Grounding** in a trusted source is the strongest fix; RAG automates it.
- Verify **names, numbers, dates, quotes and citations**; never cite what you haven't opened.
- Watch for **persuasion, leading questions and flattery**.

Check your understanding

1. Why do language models hallucinate?

A. They are programmed to lie

B. They are connected to unreliable websites

C. They generate plausible text and nothing in the process checks truth

D. Their batteries run low

2. Which instruction most directly reduces invented answers about a document?

A. "Be creative."

B. "Answer only from the attached document, quote it, and say 'Not in the document' if it isn't there."

C. "Write a longer answer."

D. "Use simple English."

3. An assistant gives you five references for your assignment. What must you do before using them?

A. Nothing, if they look professional

B. Ask the assistant if they are real

C. Open and read each one yourself

D. Use only the first one

4. You challenge a correct answer and the assistant apologises and changes it. What does this show?

A. The second answer must be right

B. Models can be persuaded; agreement is not evidence

C. The model learned from you permanently

D. You found a bug that is now fixed

Answers are at the end of this ebook.

Sources and further reading

1. Reed Smith, "ChatGPT: Artificial Intelligence Tool Generates Fake Opinions" (Mata v. Avianca, Inc., No. 22-cv-1461, S.D.N.Y., 22 June 2023). www.reedsmith.com/our-insights/blogs/viewpoints/102ilzz/chatgpt-artificial-intelligence-tool-generates-fake-opinions
2. Bloomberg Law, "Phony ChatGPT Brief Leads to $5,000 Fine for NY Lawyers", June 2023. news.bloomberglaw.com/esg/chatgpt-phony-legal-filing-case-gets-lawyers-a-5-000-fine
3. OpenAI, "Why language models hallucinate" (evaluations reward guessing; SimpleQA abstention and error rates), September 2025. openai.com/index/why-language-models-hallucinate
4. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", 2020. arxiv.org/abs/2005.11401
5. Sharma et al., "Towards Understanding Sycophancy in Language Models", 2023. arxiv.org/abs/2310.13548
6. Google Scholar. scholar.google.com
7. OpenAI, "Introducing GPT-5" (45% fewer factual errors than GPT-4o with search; 80% fewer than o3 with thinking), 7 August 2025. openai.com/index/introducing-gpt-5
8. OpenAI, customer story: Morgan Stanley (AI @ Morgan Stanley Assistant on GPT-4; 100,000 documents; evals; 98% of advisor teams). openai.com/index/morgan-stanley

Review record · Version 1.0 · Last reviewed 27 October 2026 · Facts checked against the sources above · Prepared by the WAC Centre for AI-Powered Learning. Found an error? Write to connect@worldassessmentcouncil.org.