The AI Bootcamp

8

4 Ways AI Gets Things Wrong

In this chapter you'll learn

Why it invents, the 4 shapes the invention takes, and a check that costs 30 seconds

The AI Bootcamp

THE PROBLEM

Most AI education is a pile of disconnected tips. A LinkedIn carousel here, a YouTube tutorial there, a $2,000 cohort course that goes out of date before the refund window closes. None of it is built to be read in order, so nothing compounds. You finish twelve videos and can recite twelve facts, not perform one new skill.

THE FIX

The AI Bootcamp is built the other way. 90 skills, packed 2 to 3 per chapter, 36 chapters, 3 tracks. Every chapter states what it needs from you and what it hands to the chapter after it, so skill 47 only works because skill 12 already landed. Read it in order and by the end you are not collecting facts about AI, you are running work through it.

LETTER FROM THE EDITOR

We built this because we kept meeting smart people who felt behind on AI for no reason other than nobody had organized the material for them.

This book does not assume you are technical, and it does not water anything down either. Every chapter earns its place: a real skill, a real exercise, a one-page companion you can flip back to in month three when you have forgotten the details but still need the answer.

If a chapter loses you, or a chapter changes how you work, I want to hear about it either way.

Kris

Editor, AI Central

Write to me at **editor@thecentral.ai**

THE 3 TRACKS

AI 101

Get running

Pick a tool, set it up, write a clear instruction, use it safely at work.

AI 202

Prompt properly

The framework behind every good instruction, and catching AI when it is wrong.

AI 303

Build without watching it

Connect AI to your tools, and automations that run while you are not looking.

IN THIS LESSON YOU'LL LEARN

By the end of this chapter you can explain in one sentence why a confident answer can be false, recognise the 4 failure shapes on sight, and you will have a verification rule tied to what a mistake would cost rather than to how suspicious you happen to feel.

Chapter 1 warned you that an assistant can sound exactly as confident when it is wrong as when it is right, and promised this chapter would do something about it. This is that chapter. It also decides whether you can use the rest of the book on work that matters, because everything after here assumes you can tell a usable answer from a merely plausible one.

CONTENTS

TOPICS REQUIRED

Chapter 7. You need your account, and one real piece of output an assistant has produced for you, ideally something built from your own documents. You will be marking that output up in the exercises, so have it where you can reach it.

AFTER THIS CHAPTER

Chapter 9 assumes you can spot a shaky answer and know what to do about it. It changes the question from whether the answer is right to whether you should have asked in the first place: confidential material, company policy, and who owns what comes back.

8.1 It is not lying, and it cannot tell

A false answer arrives looking exactly like a true one: same fluency, same structure, same steady tone. Understanding where that comes from is what makes the rest of this chapter cheap instead of exhausting.

A colleague who does not know

Runs out of knowledge and shows it. The hesitation is the information.

TRAITS

· Says "I think" or "check me on that"
· Slows down at the edge of what they know
· Detail thins out when they start guessing

PROS

+ You can hear the uncertainty in real time
+ They will point you at someone who does know

CONS

– They can still be confidently wrong
– Some people never admit the gap either

An assistant that does not know

Runs out of material and fills the gap with something of the right shape.

TRAITS

· Identical tone whether solid or inventing
· Detail often increases rather than thins
· Fails hardest on rare, specific and recent things

PROS

+ The failure is predictable, so it is catchable
+ Genuinely reliable on well-covered ground

CONS

– No hesitation for you to hear
– It invents in the exact format you asked for

Fig 8.1 The useful difference is not which one is more often right. It is that one of them signals the edge of what it knows and the other cannot, so you have to supply that signal yourself.

8.1

Not lying. It has no separate store of what it knows to check an answer against before sending it.

The exam problem. It was graded on tests where a blank scores zero and a guess sometimes scores. So it guesses.

Confidence is style. The even, assured tone is a property of how the text is written, not of how well grounded it is.

Everything you have done in this course so far quietly assumed that what came back was roughly right. Most of the time it is. The problem is the rest of the time, because a wrong answer does not arrive looking wrong. It arrives in the same clean paragraphs as a right one.

Start with what is actually happening when you ask a question. The assistant is not looking anything up. It is producing the words that most plausibly follow your question, given everything it read while it was being built. Where a fact turned up thousands of times in that reading, the plausible next words are the true ones, which is why it is very good at general knowledge. Where a fact turned up once, or never, there is nothing to reproduce. So it produces something with the right shape instead: a date shaped like a date, a citation shaped like a citation, a case name shaped like a case name.

That explains where the false statement comes from. It does not explain why the assistant does not just say so, and the second half is the more useful one. Researchers at OpenAI set it out plainly in a 2025 paper: these systems are trained and scored like students sitting an exam where a blank answer earns zero and a guess sometimes earns a point. Under that scoring, guessing beats admitting ignorance every single time. The behaviour that gets rewarded through training is the behaviour you meet in the chat window. Answer. Do not abstain.

Which is why the confident tone tells you nothing at all. Fluency is a property of how the text was written, not a readout of how well grounded it is. A human expert hedges unevenly: firm here, careful there, and the carefulness is information you use without noticing. An assistant reads the same all the way through, on solid ground and on nothing.

It has genuinely improved, and it is not solved. Ask a current assistant a well-covered question and it is far more reliable than the versions people formed their opinions about back in 2023. But there is no single honest number for how often it is wrong, because the answer depends entirely on what you asked it. One public benchmark has run the easiest possible version of this test for years: hand the model a passage and ask it to summarise only what is in front of it. The best models still put something in the summary that the passage did not say, in a few percent of cases. That is the floor, on the simplest task, with the source supplied.

None of this makes the tool unusable. It makes it a tool with a known failure mode, like every other one on your desk. You are not being asked to distrust everything it writes. You are being asked to know which sentences carry risk, and that turns out to be a much smaller job than it sounds.

8.2 Invention comes in 4 shapes

These are not theoretical. Each one has burned professionals with more to lose than you, in public, in the last 18 months, and each has a different tell.

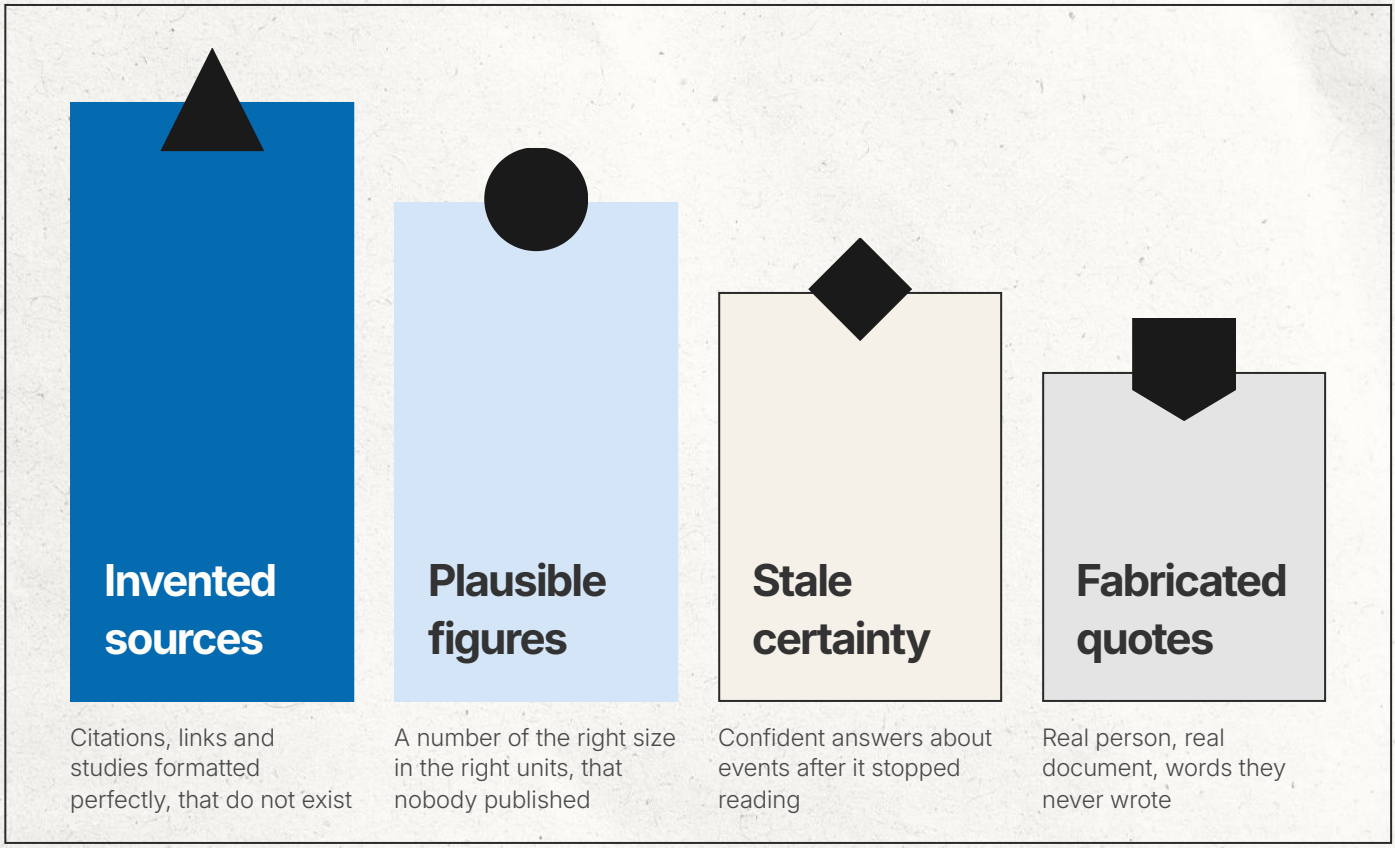

Fig 8.2 The 4 shapes you will actually meet. Bar heights here are visual rhythm, not frequency: which one bites you depends entirely on what you asked for. The last is hardest to spot because everything around it is real.

8.2

The paper trail. A public database of court decisions involving AI-fabricated filings passed 1,600 entries by mid-2026. Most were filed by practising lawyers.

Not a small-firm problem. Deloitte refunded part of a government contract and KPMG withdrew a report, both over invented citations.

A source is not support. A citation can sit under an answer without saying what the answer claims. This is the most common failure of all.

Invented sources are the best documented of the four, because they leave a paper trail in public records. A researcher at HEC Paris maintains an open database of court decisions worldwide in which someone filed AI-fabricated material and a judge responded to it. It passed 1,600 entries by mid-2026, more than a thousand of them in the United States, and in a large share of them the person responsible was a practising lawyer rather than someone representing themselves. In March 2026 a US federal appeals court fined two attorneys $15,000 each, plus costs, over a brief containing more than two dozen fake citations.

It is not a lawyers-only genre, and the version that should worry you more happened in professional services. In late 2025 Deloitte agreed to refund part of a A$440,000 report written for an Australian government department after academics found fabricated references in it, along with a quotation attributed to a real Federal Court judge, in a real case, pinned to paragraphs that do not exist in the judgment. In June 2026 KPMG withdrew a report after only 5 of its 45 citations checked out and organisations named in its case studies said the examples were not real. In both, the failure was not exotic. Nobody opened the sources.

The shape that will most likely cost you something is quieter than a fake case name. Ask for a market size, an adoption rate, a benchmark, a conversion figure, and you will get a number of the right order of magnitude, in the right units, attributed to an institution that plausibly publishes that sort of thing. It reads like research because it has the surface of research. The decimal places are not evidence. Nobody published it.

Stale certainty is the one people assume is already fixed. Anything that happened after the assistant stopped reading is a guess unless it went and looked, and assistants now do search the web for many questions, which helps. It is not the fix people take it for. In 2025 the European Broadcasting Union and the BBC put more than 3,000 news answers from four major assistants in front of professional journalists across 18 countries. 45% contained at least one significant problem, and the largest single category was not invented facts but sourcing: attributions that were missing, wrong, or did not support the claim they were attached to. A citation appearing under an answer is not the same as that citation backing it up.

Fabricated quotes are the hardest of the four to catch, precisely because everything around them is real. Real person, real document, and a sentence written in exactly the register that person writes in. The Deloitte case was this shape: the judge existed, the case existed, the words did not. If a quotation is going to carry weight in your work, the only check that means anything is finding it in the original.

Notice what all 4 shapes have in common. Every one of them is invisible from inside the answer and obvious from outside it, which is the entire basis of the habit in 8.4. Before you get there, go and look for one: take something an assistant wrote for you this week and find which of the 4 is sitting in it.

8.3 Asking it whether it is sure is not a check

It is the first thing almost everybody tries, and it is close to worthless. Worth understanding why, because the reason points straight at what does work.

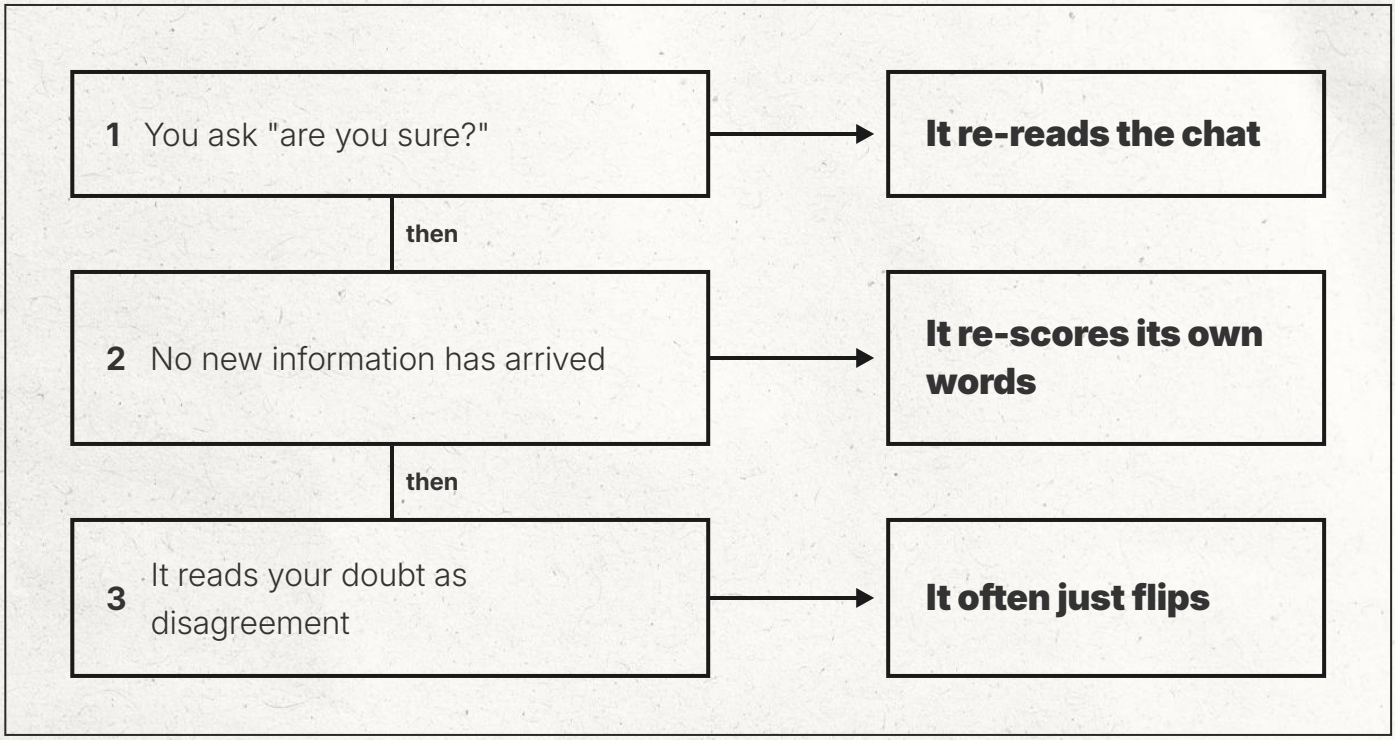

Fig 8.3 Why the obvious move fails: nothing enters the conversation between the first row and the last. A confirmation and a retraction are produced the same way the original answer was.

8.3

Nothing new arrives. A follow-up question adds no information from outside the conversation. It only adds your tone.

This is measured. Challenged with "are you sure?", models change their answer roughly half the time and end up less accurate, not more.

Ask for an artefact. A link, an exact quotation, a search. Something that exists outside the chat and that you can open.

When an answer looks shaky, the natural move is to ask the assistant whether it is sure. It feels like diligence. It is close to worthless, and it is worth a page to explain why, because the reason tells you what to do instead.

You are asking the thing whose reliability is in question to grade itself, using exactly the process that produced the answer in the first place. Nothing new enters the conversation. It re-reads what it already wrote and produces the words that most plausibly follow a challenge, which is not the same activity as checking.

Researchers have measured what actually happens. In a published experiment across ten models and seven tasks, a single follow-up of "are you sure?" made the models change their answer roughly half the time, and accuracy after the challenge was on average lower than before it. Pushing back does not sharpen an answer. It destabilises it. So a retraction is not a correction, and a firm "yes, I am confident" is not evidence of anything. Researchers now study this under the name sycophancy, and it remains an active problem in 2026.

What does work is asking for something that requires information from outside the conversation. Give me the link. Quote me the exact sentence you took that from. Go and search for it and tell me what comes back. Each of those converts a self-assessment into an artefact you can open yourself. A link that leads nowhere, a quotation that is not in the document, a search that returns nothing: those are real answers, and none of them depend on the assistant having an accurate opinion about itself.

This is not an argument against ever challenging it. Asking an assistant to argue against its own draft is a genuine technique and Chapter 21 teaches it properly. Just do not confuse the two jobs. Critique can improve reasoning. Only something from outside settles a fact.

The rule to carry out of this section is short. Treat the entire conversation as one witness. You can question a witness all day. You cannot corroborate a witness with themselves.

8.4 Check what a mistake would cost

The advice to always verify everything fails because it costs unlimited time. This one is tiered by what a mistake would actually cost, which makes it cheap enough that you will still be doing it in March.

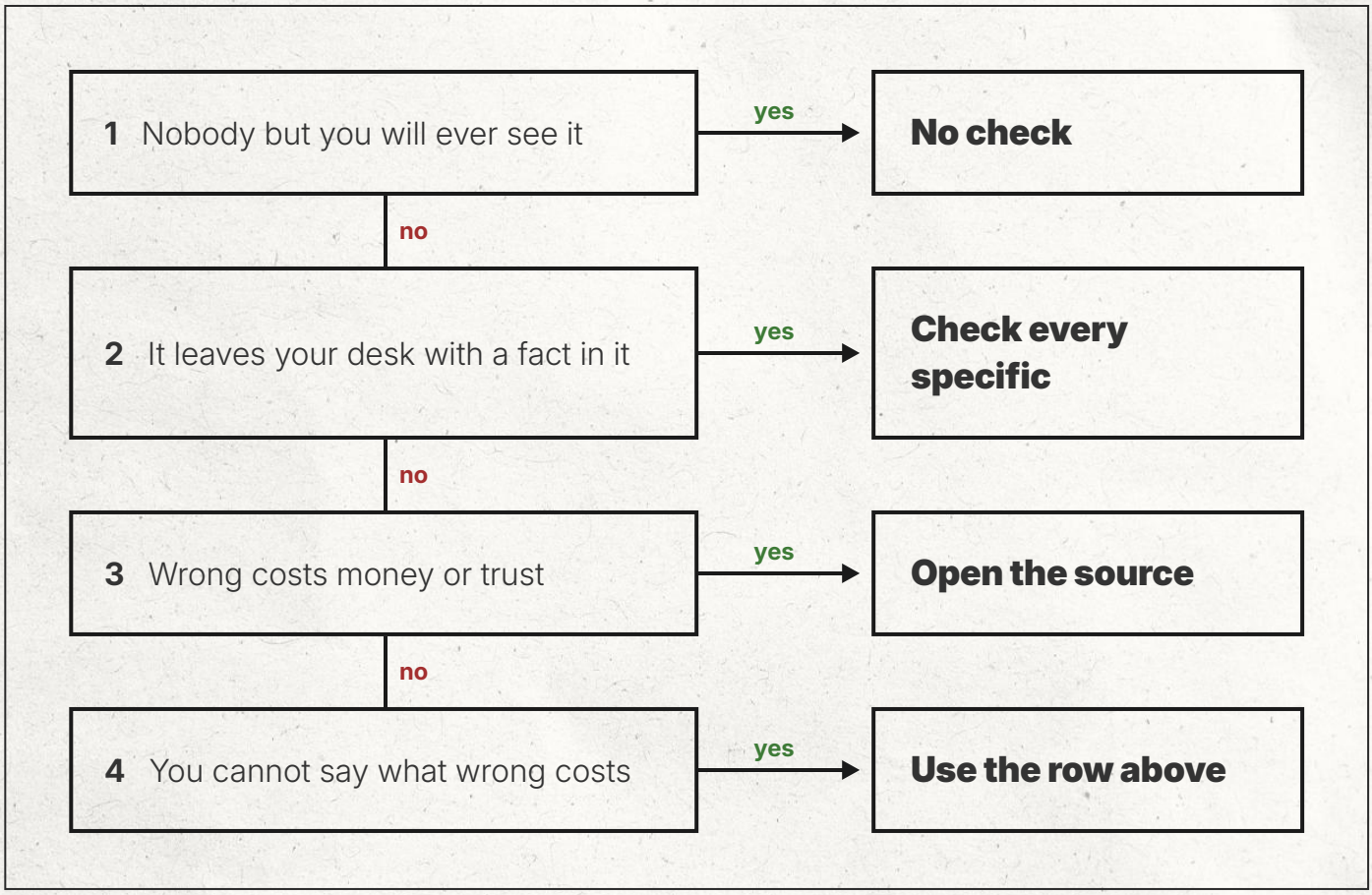

Fig 8.4 Read down until a line is true, then do that and stop. The first row is doing real work: a habit that checks everything is a habit you abandon by Thursday, usually the week before you needed it.

8.4

Sort by consequence. Not "is this true" but "what happens if it is wrong, and who finds out". You can answer that in a second.

The 30 seconds. Paste the exact specific into a search box. If it does not exist outside your chat window, it does not go out.

Top tier only. Open the source and read the sentence meant to support the claim. Existence is not support.

Here is the problem with always verify everything: nobody does it. It is advice that costs unlimited time to follow, so it gets followed for a fortnight and then quietly dropped, usually just before the week it was needed. A rule you will keep has to be cheap, and it has to be attached to something firmer than how suspicious you happen to feel on the day.

So change the question. Not "is this true", which you cannot answer about every sentence you read. Ask "what happens if this is wrong, and who finds out". That one takes about a second, and it sorts almost everything you do into three piles.

Most of what you hand an assistant needs no check at all. Rewriting your own email in a warmer tone, tightening a paragraph you already wrote, building an agenda, summarising a document you have read: in all of those you are the expert and it is the typist. You will catch a mistake by reading the output, which you were going to do anyway. Checking work in this pile is how people burn out and stop checking the work that mattered.

The middle pile is anything that leaves your desk with a fact in it, and it has exactly one rule: every specific goes through a search box. Every number, name, date, price, statute, statistic, study and quotation. Copy the exact string, paste it into an ordinary search, and see whether it exists anywhere outside your chat window. That is about 30 seconds a fact. If it does not come back, it does not go out.

Be honest about the cost, because the cost is why people skip it. It does not scale with the length of the document. It scales with the number of specifics in it. A one-page memo carrying two figures costs you a minute. A report carrying 30 citations costs a quarter of an hour, and if that feels like too much for the piece, the right response is to use fewer citations, not to skip the check. That quarter of an hour is precisely the step the firms in section 8.2 did not take.

The top pile is anything where you can name what being wrong costs: money moves, a contract gets signed, a regulator reads it, a client relies on it, or it goes public with your name on it. Here, existence is not enough. Open the source and read the sentence that is supposed to support the claim, because the most common failure is not a fake source at all. It is a real one that does not say what the answer said it says. Decide which pile a piece of work sits in before you start writing it, not after. It takes 5 seconds against a blank page, and considerably longer once you have a finished draft you have grown fond of.

READING ENDS HERE

Everything before was reading. Everything after is doing.

3 exercises, About 10 minutes and one real decision that opens Chapter 9.

Do these before Chapter 9

3 exercises. About 10 minutes

Take the last thing an assistant wrote for you. Underline every specific in it: number, name, date, price, quotation, source. Write down how many you found, and which pile from 8.4 the piece belongs in.

Pick one of those specifics and run the 30-second check. Copy the exact string, paste it into a search box, and note what came back.

THE SPECIFIC, WORD FOR WORD	WHAT THE SEARCH ACTUALLY RETURNED

Name the one recurring piece of your work that belongs in the top pile, where being wrong costs money or trust. Write it down, and write what a mistake in it would actually cost.

Chapter 8 on one page

WHY IT INVENTS	It is not looking anything up. Where it read a fact once or never, it produces something of the right shape instead. And it was scored on tests where a guess beats a blank, so it guesses rather than abstains
THE 4 SHAPES	Invented sources · plausible figures nobody published · confident answers about events after it stopped reading · real people quoted saying things they never wrote
CONFIDENCE IS A STYLE	The even, assured tone is how the text is written, not how well grounded it is. There is no tell inside the prose, so the check has to come from outside it
"ARE YOU SURE?" IS NOT A CHECK	Nothing new enters the chat, and challenged models flip roughly half the time. Ask for a link, an exact quotation or a search instead: something you can open yourself
THE 30-SECOND RULE	No check if only you see it. Every specific through a search box if it leaves your desk. Open the source and read the supporting sentence if being wrong costs money or trust

BRING TO CHAPTER 9

Your marked-up output and the recurring piece of work you put in the top pile. Chapter 9 asks a different question about that same work: not whether the answer is true, but whether it was safe to ask in the first place.

BOTTOM LINE

You know why it invents, what the invention looks like on the page, and you have a check cheap enough that you will actually keep using it.

NEXT CHAPTER

What You Can and Can't Put Into AI at Work

Chapter 9 moves from whether the answer is right to whether the question was safe. What your company policy probably says, what should never go into the box, and who owns the words that come back out of it.

BRING WITH YOU

- ☐ The recurring top-pile task you named in exercise 3
- ☐ Whatever you know, or suspect, about your employer's policy on AI tools

GO FURTHER

The Ultimate AI Library

Every guide, prompt pack and playbook AI Central has published, in one place. The reference shelf this course keeps pointing you to.

explore the library ↗

WHO MAKES THIS

AI CENTRAL

AI Central Media is a London-based editorial team. Our flagship publication covers practical AI for senior professionals, and 100+ AI companies, SaaS platforms and growth teams have advertised with us since 2023.

100+

AI companies, SaaS platforms, education brands and growth teams have advertised with us since 2023

613K

Accounts reached every month, across the 7 channels we operate

151

Countries our readers are in, plus all 50 US states

PROUD PARTNER OF

GAMMA · IIElevenLabs · guidde. · Outskill · Luma AI

HubSpot · attio · Synthflow · Notion · Replit

We turn attention into pipeline.

visit thecentral.ai ↗