TIER 1 - FOUNDATIONS * V1.0 -- MAY 2026

READING AI OUTPUT CRITICALLY

The five tells that the model is bullshitting you, and the verification moves that catch them before you act on bad info.

BY Alex Jahn / Agent Logic

EDITION v1.0 -- May 2026

AUDIENCE Anyone who's ever pasted an AI answer into something that mattered and felt a flash of "wait, is that actually true?"

READ TIME 15-20 minutes

COST Free. Forever.

Prepared by Agent Logic / alexanderjahn79@icloud.com / theaiguywi.com

CONTENTS

What's in here

SECTION 1

The most expensive failure mode

Confidently wrong

Of all the ways an LLM can fail you, the worst one isn't "it doesn't know." That one's easy. The model says "I don't know" or hedges and you move on.

The expensive failure is when the model writes a beautiful, fluent, utterly confident answer -- and the answer is wrong. You read it, it sounds right, you don't have any reason to doubt it, you act on it. A week later you find out the case citation was fake, or the regulation it quoted is from three versions ago, or the customer-service script you sent referenced a feature your product doesn't have. The model didn't know it was wrong. You didn't know it was wrong. The cost was real.

This is the failure mode this module is built around. Not "the AI is dangerous." Not "AI hallucinates a lot, don't trust it." Both of those framings are wrong. The right framing is more useful: **the mode**
produces fluent text whether or not it has the goods, and you are the only fact-checker in the loop.

You don't need to become paranoid. You need to become a *reader* of *A*
good editor reads a writer's draft. With trust, but with eyes.

What this module does

By the time you finish this primer you'll have:

- A short, honest answer to *why* the m
  module 1's mental model -- but more pointedly).
- The five tells. Specific patterns that signal "this answer needs verification."
- The verification moves. Quick, free, reliable ways to check load-bearing facts in under 60 seconds.
- A when-to-trust framework. Three categories of stakes, three different verification levels.
- The skeptical operator's toolkit. The five-second habits that make this automatic.

Twelve pages. No fear-mongering. No hype. Just the editorial discipline that turns AI from a risk into a multiplier.

> ***The most dangerous AI answer is the one that's 90% right. You read the right parts, trust the wrong parts, and never notice the seam.***

SECTION 2

Why it sounds so confident

A short refresher on why this even happens, then we move to tactics. If you read module 1 (*What An LLM Actually Is*) some of this will be familiar.

The model doesn't know whether what it's writing is true. It has no internal "I'm not sure about this part" signal. It generates the next chunk of text by predicting what's most plausible given the context, and the *tone of co*
Confident-sounding answers and tentative-sounding answers come out of the exact same machinery -- the model just predicts whichever one fits.

That means the model can produce four sentences in a row where the first three are correct and the fourth is invented, all in the same calm, authoritative voice. The voice is decoration. It is not signal.

A useful reframe: **treat the**
the bar telling you about something they read once. They might be right. They probably are, on familiar topics. But on anything load-bearing, you check.

Two specific quirks worth naming

These come up enough to call out:

The model agrees with you a lot. If you assert something -- even something wrong -- the model is statistically much more likely to agree than push back. This is sycophancy. It is not malice. It's prediction: in the training data, when humans confidently say something, the most common reply is *agreement*. So the model agrees. The fix: ask the model to argue *against* yo
"what's the strongest case I'm wrong?" You'll often get a completely different answer.

The model invents specifics. When generic prose isn't satisfying enough -- when you ask for a "real example" or "a specific case" or "actual numbers" -- the model often produces fluent specifics that are simply made up. Plausible-shaped names, plausible-shaped citations, plausible-shaped numbers. The pattern is real ("a real example sounds like this"). The contents are not.

We'll catch these patterns in section 3.

SECTION 3

The five tells

When the model is bullshitting, there are signs. They're not always loud, but they're consistent. After reading this section you'll start spotting them automatically.

Tell #1 -- Specifics that are too clean

Real-world facts are messy. They have weird numbers, awkward names, dates that don't round.

When the model is making something up, the made-up version is often *too clean*. Memorable-but-suspicious names. Dates that fall on quarter-end. A statistic that's exactly 30% or 50% or 75%.

Example tell: "A 2023 McKinsey study found that 75% of small-business owners reported productivity gains within 90 days of adoption."

Real studies have weird numbers -- 73% or 68% or "a majority." Real reports have specific lengths and titles. When everything in an answer is round, suspicious shape, suspicious shape, suspicious shape -- that's the tell.

The verification move: If the answer hinges on a stat or citation, ask the model: "*What's the source for that number? Give me the title of the study, the publication year, and a direct link.*" If it produces a confident answer, paste the title into Google. If it doesn't show up -- the source is invented.

Tell #2 -- Citations that look right but feel off

This is the number one most-cited example of AI hallucination, and it's worth its own tell because it shows up everywhere -- legal, academic, journalistic, business research.

The model produces references that have the *shape* of r abbreviations, page numbers, journal names. Sometimes they're real. Sometimes they're fabricated. Sometimes the case is real but the page or the quote is invented.

Example tell: "*See Smith v. Western District Construction*, 412 F.3d 287 (7th Cir. 2009)."

Plausible. Confident. Possibly real. Possibly not.

The verification move: If a citation is load-bearing -- meaning your decision changes if the citation is fake -- paste the case name, statute, or paper title into Google or the official source database (PACER for federal cases, the relevant agency's site for regulations, etc.). Five seconds. If it doesn't appear, it's invented.

Tell #3 -- The "as of [date]" tell

If you ask the model about something time-sensitive -- pricing, current events, who runs a company, what version of a product exists -- and you get a confident answer with no caveats, that's a tell.

The model has a knowledge cutoff. It doesn't know what happened after a certain date, and depending on the model that date might be six months ago or two years ago. Recent prices, current product versions, who-acquired-whom -- these are the categories where the model is most likely to confidently state something that *was* true a

Example tell: "Twilio's standard SMS rate is $0.0079 per message in the US."

Maybe true today. Was true at training time. Might be off by 20% by the time you read this.

The verification move: For any time-sensitive fact, the prompt itself should include the language: "*If this answer depends on information that may have changed since your training cutoff, say so explicitly.*" Then check anyway. Pricing pages, vendor websites, official sources -- 30 seconds.

Tell #4 -- Smoothing over uncertainty

This is the most subtle tell. When the model isn't sure, instead of saying "I'm not sure" it tends to *smooth over the gap* with confident-but-vague language. Watch for these patterns:

- "*A common approach is...*" (whose approach? where?)
- "*Most experts agree that...*" (which experts? agree based on what?)
- "*Studies have shown...*" (which studies? cite them.)
- "*It's generally understood that...*" (by whom? where is this written?)

These phrases sound authoritative. They're often the model's tell that it's spinning a coherent-sounding answer without actually having the goods. The pattern: when the model gets vague *just* at the

The verification move: Ask follow-up: "*Cite the specific source for that claim -- author, title, date, where I can find it.*" If the model can't, the original claim was generated, not retrieved.

Tell #5 -- Drift across the answer

Long answers are particularly dangerous because the model's confidence in early parts can carry over to later parts that are weaker. The first three points are solid because they're well-trodden territory. Point four is where the model started inventing.

The verification move: When reading a long AI answer, treat each major claim as independent. Don't grant point four credibility because point one was right. The model didn't earn cumulative trust -- it generated each chunk independently. You evaluate each chunk independently.

The five tells, summarized:

1. Specifics that are too clean (round numbers, suspicious shapes)
2. Citations that have the shape but might not exist
3. Time-sensitive facts stated without caveats
4. Vague-confident language hiding uncertainty
5. Long-answer drift -- later sections weaker than earlier

If you spot any of these, the right response isn't "the model is bad." It's "this specific claim needs verification before I act on it."

SECTION 4

The verification moves

You don't need to verify everything. You need to verify the load-bearing claims -- the parts where being wrong has a real cost.

Three quick moves cover 95% of cases.

Move 1: The "source it" follow-up

You don't have to leave the chat window. The first move is asking the model itself:

> *"For each load-bearing claim in your last answer, give me the specific source -- author, title, year, and where I can verify it. If you don't have a specific source, say so explicitly."*

This works surprisingly well. The model will often distinguish -- *"This first [actual citation]; this second claim is my generalization based on common patterns; this third claim I cannot source specifically."* Half the time, the model itself flags what it generated versus what it pulled from solid training data.

Move 2: The 30-second Google check

For any specific fact (a number, a citation, a date, a person, a quote), one targeted search:

- Paste the specific claim into Google or your search engine of choice.

- Look at the first three results. Are they from credible sources? Do they confirm the claim?
- If the claim doesn't appear anywhere on the first page of results -- it's probably invented.

30 seconds. Free. The single highest-leverage habit in this entire module.

Move 3: The cross-model check

For high-stakes claims, ask the same question to a different model. ChatGPT and Claude were trained on overlapping but not identical data. If both confidently produce the same answer with the same details, that's stronger evidence the answer is grounded. If they disagree -- at minimum you know there's a real ambiguity to investigate.

This isn't a perfect check (both models can be wrong in the same way if they were trained on the same wrong source). But for most business decisions, the cross-model check is a quick sanity test.

60

Seconds.

That's the upper bound on running all three verification moves on a load-bearing claim. Anyone telling you they don't have time to verify AI output is telling you they don't have time to be careful with their business.

SECTION 5

When to trust, when to verify, when to throw ou

Not every AI answer needs verification. Knowing the difference saves you hours.

Three stake levels

Low stakes:
Drafting an email to your sister. Brainstorming names for a project. Casual research for personal curiosity. Outline-shaped first drafts of internal docs.

Verification: None required. Use the model's output, move on.

Medium stakes:
misquoted internal policy. Customer-facing emails. Pricing-adjacent calculations. Policy summaries. Anything that goes to a person with money or a relationship at stake.

Verification: One of the three moves above. The "source it" follow-up is usually enough.

High stakes:
could sue you, fire you, or stop trusting you. Legal references. Medical claims. Regulatory compliance. Anything that goes into a contract, a court filing, a board meeting, or a credentialed deliverable.

Verification: All three moves, plus a credentialed human in the loop.
work, never your authority.

The two questions to ask before you act

Before you copy any AI answer into something that matters, ask yourself two questions:

1. **If this is wrong, what's the cost?** Cheap mistake -> ship. Expensive mistake -> verify.
2. **What part of this answer is load-bearing?** The specific number? The citation? The recommendation? Verify

These two questions are the entire heuristic. Internalize them and you'll spend almost no extra time, but you'll catch the failures that matter.

SECTION 6

The skeptical operator's toolkit

The five habits that turn this module's lessons into reflex. None take more than a few seconds.

1. Read the answer once before acting on it. The number of people who paste AI output into a customer email

2. Identify the load-bearing claims. What's the one or two facts this answer hinges on? Underline them mentally.

3. Apply the five tells. Do any of the load-bearing claims look too clean, too specific, too time-sensitive, too vague-confident, or too late in a long answer to trust at face value?

4. If yes, run a verification move. "Source it" follow-up, 30-second Google check, or cross-model -- pick one. Most cases need only one.

5. If the answer touches anything in the high-stakes category, get a human. A real lawyer for legal questions. A real accountant for tax questions. A real licensed pro for licensed-pro work. The model is an assistant, never an authority, on regulated matters.

That's the toolkit. Five habits. Maybe 30 extra seconds per AI interaction. Saves you the one bad outcome that would have cost you a customer or a license.

The mindset shift

The single most useful mental shift this module is trying to install:

> ***The AI's job is to draft. Your job is to edit. Anyone who flips that order -- letting the AI ship and editing nothing -- is one bad answer away from a problem they didn't see coming.***

SECTION 7

Where to go from here

You've now completed the first three Tier 1 modules:

1. **What an LLM actually is** -- the mental model. Prediction machine, not thinking machine.
2. **The 3-question prompt framework** -- the operating procedure. Want, context, constraint.
3. **Reading AI output critically** *(this one)* -- the verification habit. Five tells. Three stake levels.

Together that's the foundation. Three more Tier 1 modules to go (When NOT to use AI, AI as a tutor, AI for life admin) and then you're into Tier 2 -- applying this to your actual job. The next big leap.

Get the next module the day it drops: theaiguywi.com/training

One short email per release. No drip. No spam. Opt out anytime.

If you want this same editorial discipline trained into your team -- the habits, the verification reflex, the toolkit applied to your actual work products -- that's the consulting offer. We install it the same way I run it in my own carpentry business.

Reach out: alexanderjahn79@icloud.com

A short call. Honest scope. We figure out together if it's a fit.

Closing -- the lock-in line

If you remember nothing else:

5

Five tells. Three moves. Two questions. Two stakes-levels that demand verification.

The model drafts. You edit. Anyone who flips that order is one bad answer away from a problem they didn't see coming.

You have the foundation. Tier 2 is where you start building.

Agent Logic -- *theaiguyw*
Fond du Lac, WI. This is module 3 of 6 in Tier 1 (Personal).