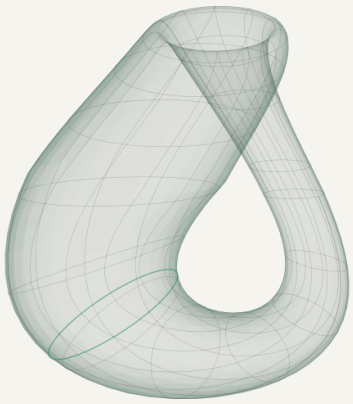

How to read an AI claim.

A buyer's guide: eight checks you can run on any AI claim before you act on it.

BEFORE YOU START

Every vendor pitching you an AI product will show you a number. A headline accuracy rate, a case study result, a benchmark score. Some of those numbers are true and useless. Some are true and misleading. A few are simply wrong. Almost none of them arrive with the information you would need to tell which is which.

This guide is eight checks. Each one is a question you can ask about any AI claim, illustrated with a real case, so you know what a good answer actually looks like when you hear one. None of it requires technical background. All of it requires reading past the headline.

Two of the cases are dated: a specific paper, a specific month, a specific pair of numbers. We are keeping the date attached on purpose. The lesson is not "this model is bad." It is "this is what happens to a number between the moment it's measured and the moment it's repeated," and that lesson does not expire when the model does.

Is the headline number the actual result, or a number one step over?

THE QUESTION

When a claim leads with one striking figure, ask what that figure was measured against, and whether a different, more representative number sits one column over in the source's own data.

THE CASE, IN THREE SENTENCES

A*-Thought-V2 (arXiv, September 2026) headlined a 2.29x efficiency gain. That figure is 0.80 divided by 0.35, the new method against the untouched original model; against a fairer, same-data comparison (0.80 divided by 0.66, against the already fine-tuned intermediate version), the real gain was 1.21x, both numbers arithmetically exact from the paper's own cells, one of them just measured against a harder target. Topological Necessities (arXiv, September 2026) did the same thing three days later: its abstract led with a 36.0-point gain that belonged to a single task inside a four-task table, next to an overall gain of 10.8 points, a third the size, once every task was counted. That second paper's own prose and its own results table did not even agree with each other about a smaller number in the same table, and the two figures do not reconcile under any pairing of the table's rows, so the honest response was to report the contradiction rather than pick a side.

WHAT A GOOD ANSWER LOOKS LIKE

The vendor points you to the full table, not just the headline sentence, and the headline number and the overall number are the same number, or the difference is explained plainly, in the vendor's own words, before you have to go find it yourself.

Has anyone actually tried to make this check fail?

THE QUESTION

A gate, a safety check, or a quality control that has never triggered looks identical to one that no longer works. Ask when it last caught something real, on the live system, not in a design review.

THE SHAPE, IN THREE SENTENCES

A gate that has been quiet for a long time can mean the bad case genuinely hasn't happened, or it can mean the system moved underneath the gate and the gate is now checking nothing. Both look the same from the outside: zero recent firings. The only way to tell them apart is to manufacture the exact bad case the gate exists to catch and run it against the version that actually executes, not a copy in a docs folder.

WHAT A GOOD ANSWER LOOKS LIKE

The vendor can tell you the date they last proved the check could fail, and what they did to prove it. "It's never triggered" is not an answer to "does it still work." It's the question restated.

Does "it finished" actually mean the work is done?

THE QUESTION

Ask how the vendor's system tells the difference between an agent that completed a task and one that stopped early, believing it was finished.

THE CASE, IN THREE SENTENCES

A 2026 benchmark (Long-Horizon-Terminal-Bench, arXiv 2607.08964v2, July 2026; the paper's abstract page still carries version 1's count of 15 models while the version 2 body says 17, which is its own small example of why the body, not the summary, is the source) ran 17 frontier models through 46 real multi-step terminal tasks. Decomposing the unresolved runs: 79% were still working when time ran out, but 19% were early exits, and inside those the study found 14 runs, a false finish, where the agent stopped at 75% or more complete believing the job was done; on one task seven different models stopped between 80% and 87%. Binary pass/fail grading cannot see this at all, and 62.8% of all runs in the study earned real partial credit a pass/fail grader would have scored as a flat zero.

WHAT A GOOD ANSWER LOOKS LIKE

Completion is a number a script computes against a rubric written before the work started, not a claim the system makes about itself. "Resolved" means a defined score threshold, not "the agent said it was done."

Can the system actually say "not yet," or only "yes" and "no"?

THE QUESTION

Ask whether a status field that reports "unknown" or "pending" is telling you the system is still computing, or is quietly standing in for a real problem the interface has no way to display separately.

THE SHAPE, IN THREE SENTENCES

A merge check has three honest answers: yes, no, and not yet. A dashboard with only two displayed states will misreport whichever one is rarer, and if "not yet" is the common, transient state, people learn to read every ambiguous status as "probably just pending" until the one time it genuinely isn't, and by then the habit is already formed. The fix isn't a longer timeout; a longer wait only delays the same collapse to a later moment.

WHAT A GOOD ANSWER LOOKS LIKE

In-progress, resolved-and-blocked, and resolved-and-clear are three different displayed states, each in its own words, so a stuck process never reads the same as a healthy one still running.

If the AI made the work faster, what happened to the backlog?

THE QUESTION

Ask what happened to the queue after the AI went in, not what the AI itself can do. If the backlog didn't shrink, ask where the bottleneck actually moved.

THE SHAPE, IN THREE SENTENCES

Gartner forecasts more than 40% of agentic AI projects will be canceled by the end of 2027, and its own named reasons are escalating costs, unclear business value, and inadequate risk controls, none of which are about the model's capability; this is a forecast, not a measurement, and worth reading as one either way. Removing the cost of producing work does not remove the work; it reaches the next constraint, which turns out to be how much oversight, review, and accountability the people around the system can sustain. Adding agent capacity is close to free; adding the oversight to check that capacity is not, and that channel does not get ten times wider because the agent fleet got ten times bigger.

WHAT A GOOD ANSWER LOOKS LIKE

The vendor can describe the operational layer around the model, explicit state, defined gates, a verifier separate from the producer, not just the model's own throughput. If the pitch is entirely about how much faster the agent works, ask who checks it, and how that scales.

What does it cost you to verify this is actually right, and who is checking it?

THE QUESTION

Ask how you would confirm a given output is correct, and how long that takes. Then ask whether the person doing that checking is the same system that produced the output, or someone independent of it.

THE CASE, IN THREE SENTENCES

Raising an accuracy number cuts the number of wrong answers, but not the number of answers you still have to check, because if you can't tell in advance which ones are wrong, you have to verify nearly all of them; a higher score does not buy you a shorter checklist. In a study of physicians reading x-rays (Radiology, November 2024), reviewer accuracy was 92.8% when AI advice was correct, using the format reviewers trusted more quickly, and collapsed to 23.6% when the advice was incorrect under that same format; the effect held, smaller, under the study's other format too (85.3% correct, 26.1% incorrect), evidence that a human next to AI output is not the same thing as a human checking it. Chat-style answers and citation trails feel like a fix, but a citation is a pointer to work the reader now has to do, open the source, confirm it says what's claimed, which is often as much effort as answering the question yourself.

WHAT A GOOD ANSWER LOOKS LIKE

The vendor can show you a receipt, a result verified against your own task by a check independent of the system that produced it, not just a demo, and the checkpoint is designed with visible confidence and a real, logged override, not a rubber stamp with an extra click.

Would you run this if you knew it could execute code with your permissions?

THE QUESTION

Before installing any AI skill, plugin, or extension, ask whether you are treating it like documentation you skimmed, or like a program you are about to run with your own access.

THE CASE, IN THREE SENTENCES

An agent skill is not a document; the vendor's own engineering documentation says skills can include code the agent executes at its discretion, and a 2026 research paper (arXiv 2604.03081, April 2026) showed that malicious logic hidden inside a skill's own code examples achieved an 11.6% to 33.5% bypass rate against defenses that stopped an obvious, explicit attack 100% of the time. A February 2026 scan of 3,984 skills across two marketplaces found 36.82% carried at least one security flaw and 13.4% a critical one, and among the skills confirmed outright malicious, 91% combined prompt injection with traditional malicious code; on ClawHub, one of the two marketplaces scanned, 17.7% of skills fetch untrusted third-party content, turning even a well-meaning skill into an indirect injection path. The attack hides in the part everyone assumes is safe: the examples the agent was always going to reuse, not an obvious malicious instruction.

WHAT A GOOD ANSWER LOOKS LIKE

Before a skill touches your systems, someone who isn't the agent that wants it confirms who authored it, what it can touch, whether its documentation carries hidden instructions, and whether it reaches an external endpoint, then pins the version and logs the decision. "It's just a markdown file" is the reason the attack works, not a reason it's safe.

Is this cheaper than the alternative, or just cheaper than the number you were shown?

THE QUESTION

When an AI tool is compared to hiring, or to any other alternative, ask what's actually being loaded into each side of the comparison, and whether the number quoted for the human side is the full cost or just the visible one.

THE SHAPE, IN THREE SENTENCES

Base pay is not the full cost of a hire; benefits, recruiting, and turnover all sit on top of it and are frequently left out of a side-by-side comparison. The honest version of this comparison loads both sides fully: the true, all-in cost of the role you're comparing against, not just an offer letter, against the true, all-in cost of the AI tool, not just its sticker price. Then it asks the harder question underneath the arithmetic: what share of the role is genuinely repeatable, since an agent replaces the repeatable slice of work, not the judgment, relationships, or accountability that come with the rest of it.

WHAT A GOOD ANSWER LOOKS LIKE

The vendor shows you how to build the comparison with your own numbers, your own salary bands, your own share of repeatable work, rather than handing you a single benchmark ratio and calling it the answer. A number that only works as an industry-wide average is not a number about your decision.

ONE MORE THING

Verification itself can fail quietly, and it is worth knowing the shapes that takes, because you will meet all five checking a vendor's claims, not just ours. A score that comes back identical across nearly everything is not a measurement, it's a broken instrument. A source can exist in layers, and the layer in front of you may be a record about the thing, not the thing itself. An automated check can be blind to the exact values it's searching for, so a zero result reads as absence when it's actually a blind spot. A pattern-matching rule can be off by one small detail, a heading level, a character, and report something present as missing. And the sharpest one: a claim can get checked against the wrong source entirely, marked false when it was simply never attributed to the source you checked it against.

We also ran check 1 of this guide, the headline-versus-one-step-over check, on our own published Field Notes while building it, and it caught two of ours. One post reported a skills-marketplace figure that belonged to a narrower slice of the data than the sentence implied. Another reported a study's result under the condition that produced the largest gap, without naming the condition, when the effect held under both conditions the study tested, just by a smaller margin. Neither was intentional, and that is exactly the point: this is the shape checklists like this one exist to catch. It caught us, and both are corrected in public with the correction dated on the page. That is the only credibility this guide can claim, and the only kind worth having: not that we are careful, but that we check by a second route, run by someone who did not produce the first answer.

WHO WROTE THIS

How EP works.

Ena Pragma puts AI to work on the manual jobs that run mid-market operations: we connect the systems a business already uses, put AI on the repetitive work, and keep it running with controls and a record of every action. Every check in this guide is one we hold ourselves to before we hold a client's work to it. We do not grade our own homework: a result our own systems produce gets checked by something independent of the thing that produced it, before it ships. We publish what we can verify and say plainly what we can't yet. Claims we can't source, we don't make. That is not a marketing position. It is the actual discipline behind the work, and it is why this guide exists: the questions above are the same ones we ask about our own claims before we ask you to trust them.