AUDITING AI AGENTS: A PRACTITIONER'S GUIDE TO FALSE ALARMS AND FALSE ASSURANCES

How to verify claims when code tests aren't enough.

Middleware order says otherwise.
Needs verification.

The system is fully secure and unauthenticated.

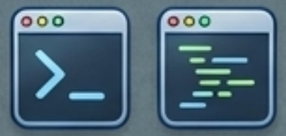
Agents are excellent witnesses to what a file says, but unreliable witnesses to what it means.

Three wrong claims in one afternoon of agent research. The agent perfectly read self-contained files but entirely misunderstood the system context. This is not a complaint about agents—it is a job description for the human.

Code has test suites. Sentences do not.

The Era of AI Code

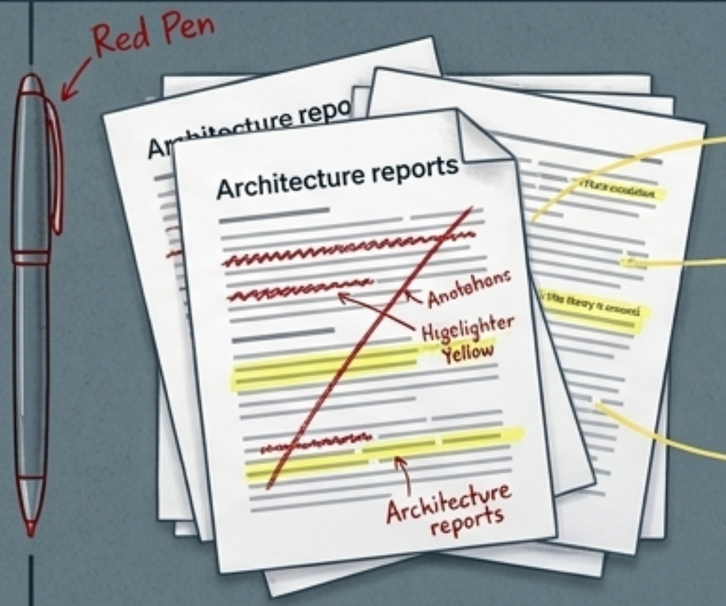
Output:



Functions, scripts, modules.

Automated Tests (Round-trip, property-based, CI/CD).

The Era of AI Claims



Output:

This is compliant.

This library is unused.

Human reading.

We solved AI code verification with tests. But increasingly, agents produce claims that end up in decisions and architecture reports. You cannot CI a paragraph. The primary verification instrument for a claim is reading the source.

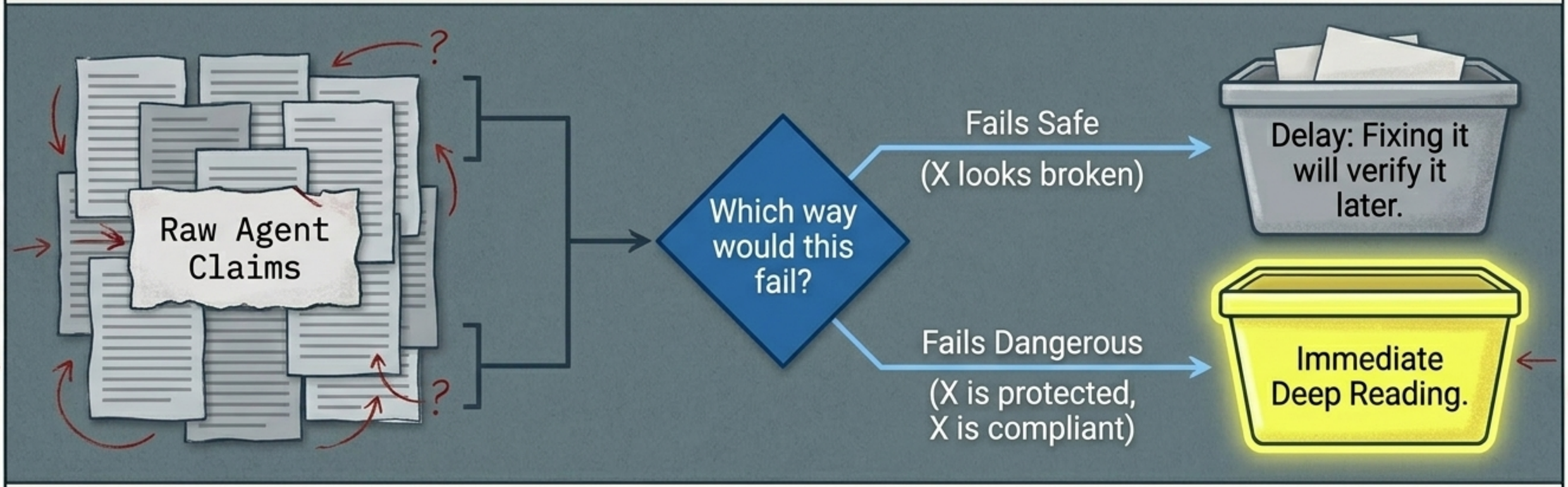
Not all hallucinations are equal.

	False Alarms (No authentication)		False Assurances (Fully RFC compliant)
(Reality)	The system is safer than reported.	(Reality)	The system is worse than reported.
(Cost)	1 hour of investigation.	(Cost)	A security incident, an audit finding, a decision built on sand.
(Default Visibility)	Highly visible (investigating broken things is natural).	(Default Visibility)	Invisible by default (nobody investigates good news).
(Outcome)	Mildly embarrassing, but cheap.	(Outcome)	Catastrophic failure.

HAZARD

The failure modes of AI claims are strictly asymmetric. False alarms survive. False assurances are the silent killers of AI-augmented development.

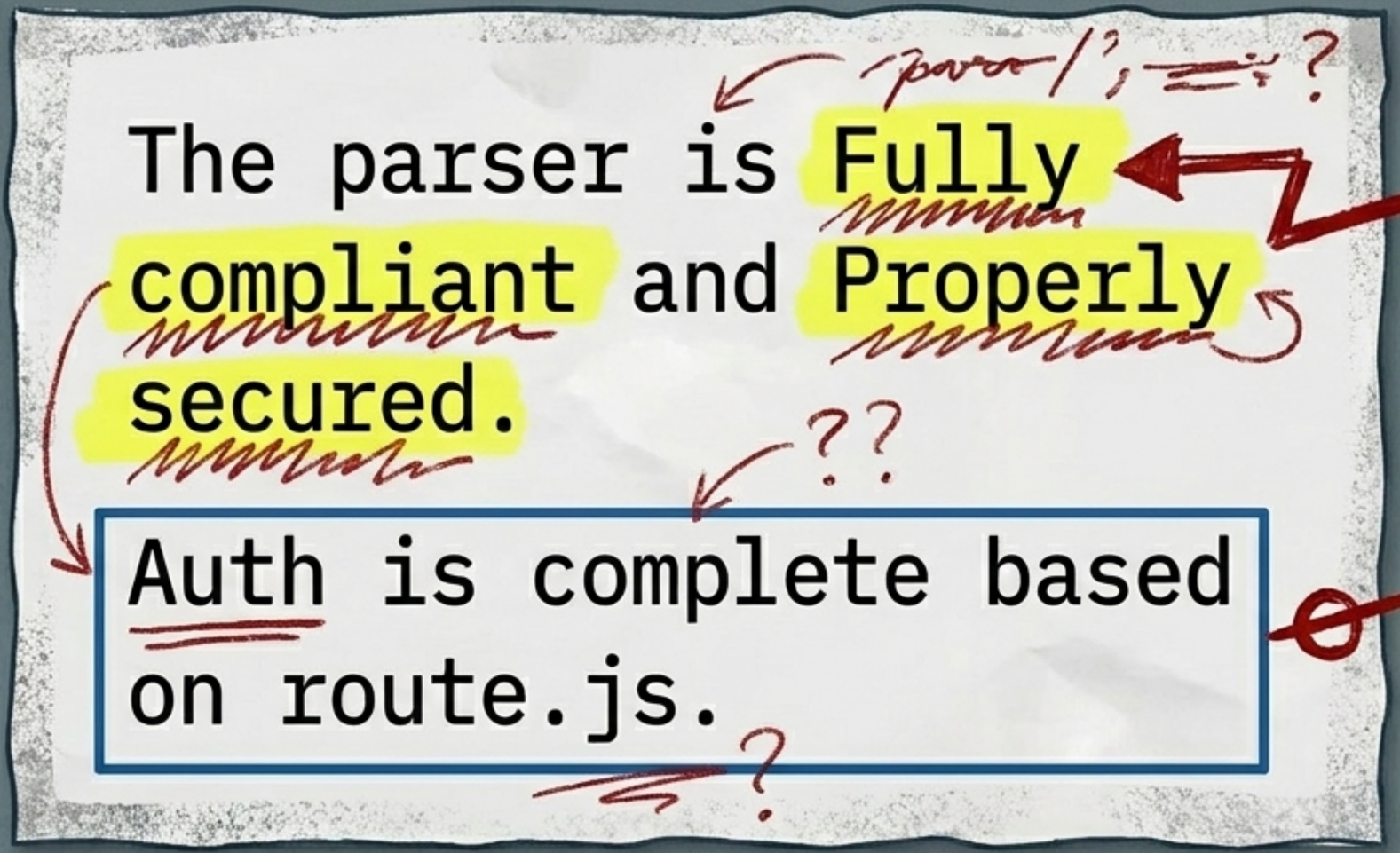
Sort an agent's claims by the direction of failure.



You cannot read everything the agent read—that erases the speedup you hired the AI for. The actual skill is deciding which claims to spend your reading time on. Verify the assurances first.

HUMAN REVIEW REQUIRED

The linguistic signatures of a hallucination.



Adjectives are where fiction hides. Facts come with file paths; adjectives come from the vibe of the code. Open the standard.

Distrust claims about context. Truths live in system boundaries (middleware ordering, load order, specs). Agents trust isolated files.

HUMAN REVIEW REQUIRED

Demand citations, then spot-check them.

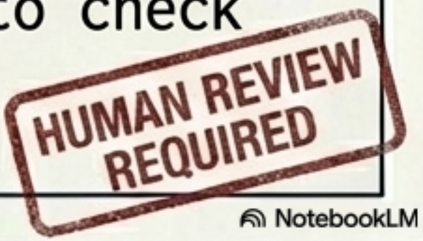
BEFORE (Uncited)

The unused legacy module is safely isolated.

AFTER (Cited)

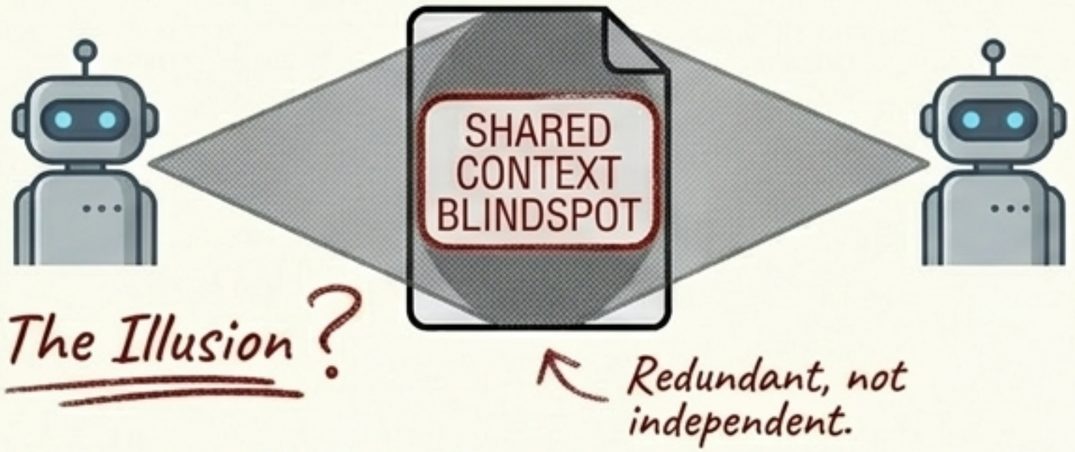
The legacy module is isolated [config/routes.js:42]

Make agents report file and line numbers for every load-bearing claim. Citations don’t magically make claims true. Instead, uncited claims cannot be cheaply audited, and a citation that fails to check out is an early-warning system for the rest of the report.

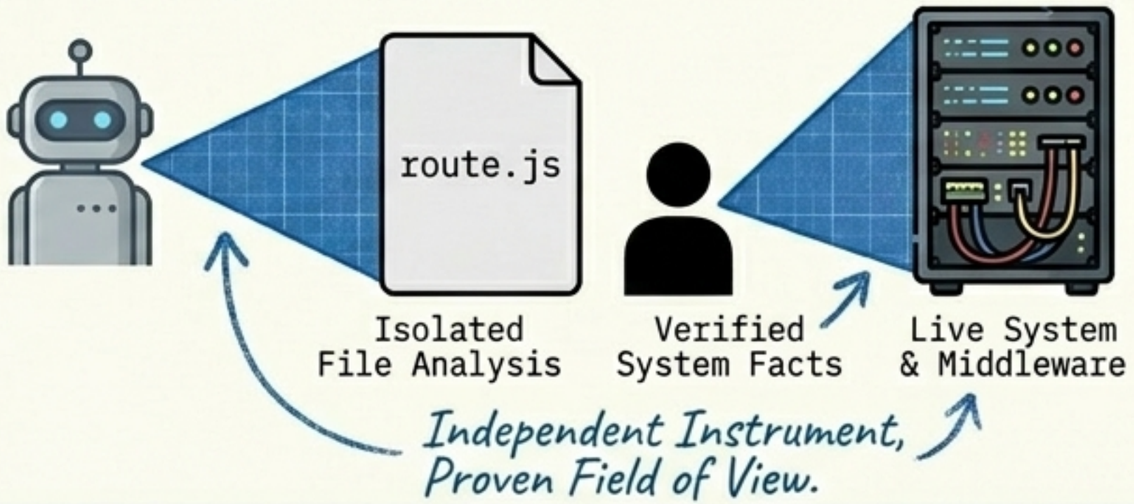


“Are you sure?” is not verification.

SCENARIO A (THE ILLUSION)



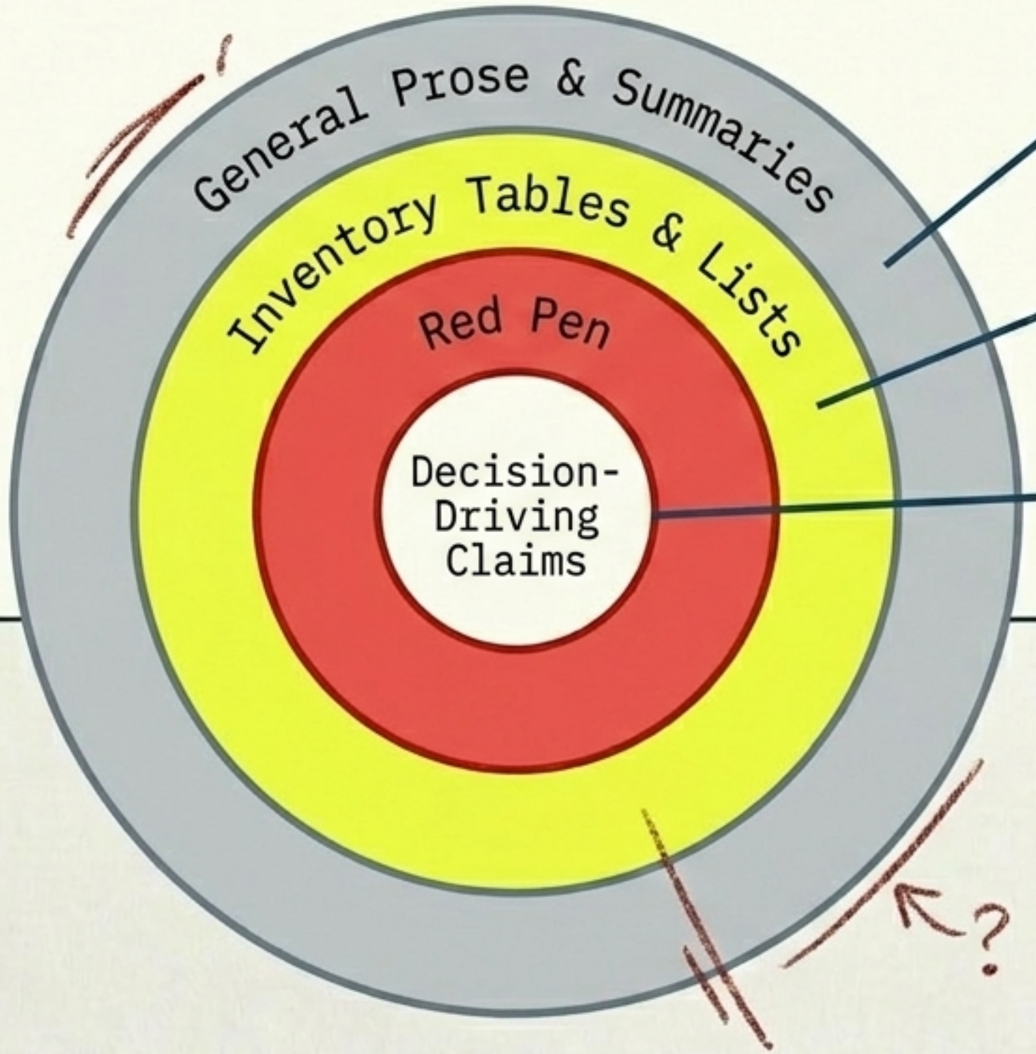
SCENARIO B (TRUE DILIGENCE)



Asking a second agent feels like diligence. It isn’t. Two agents share blind spots the way two readers of the same misleading file do. Independent verification requires a different instrument—the source, the spec, the running system.

HUMAN REVIEW REQUIRED

Sample like an auditor, not a proofreader.



Action: Full source verification.

Action: Random spot checks.

Action: Skim to catch hidden claims.

You cannot verify everything, and uniform skimming verifies nothing. Auditors solved this long ago: sample strictly by impact.

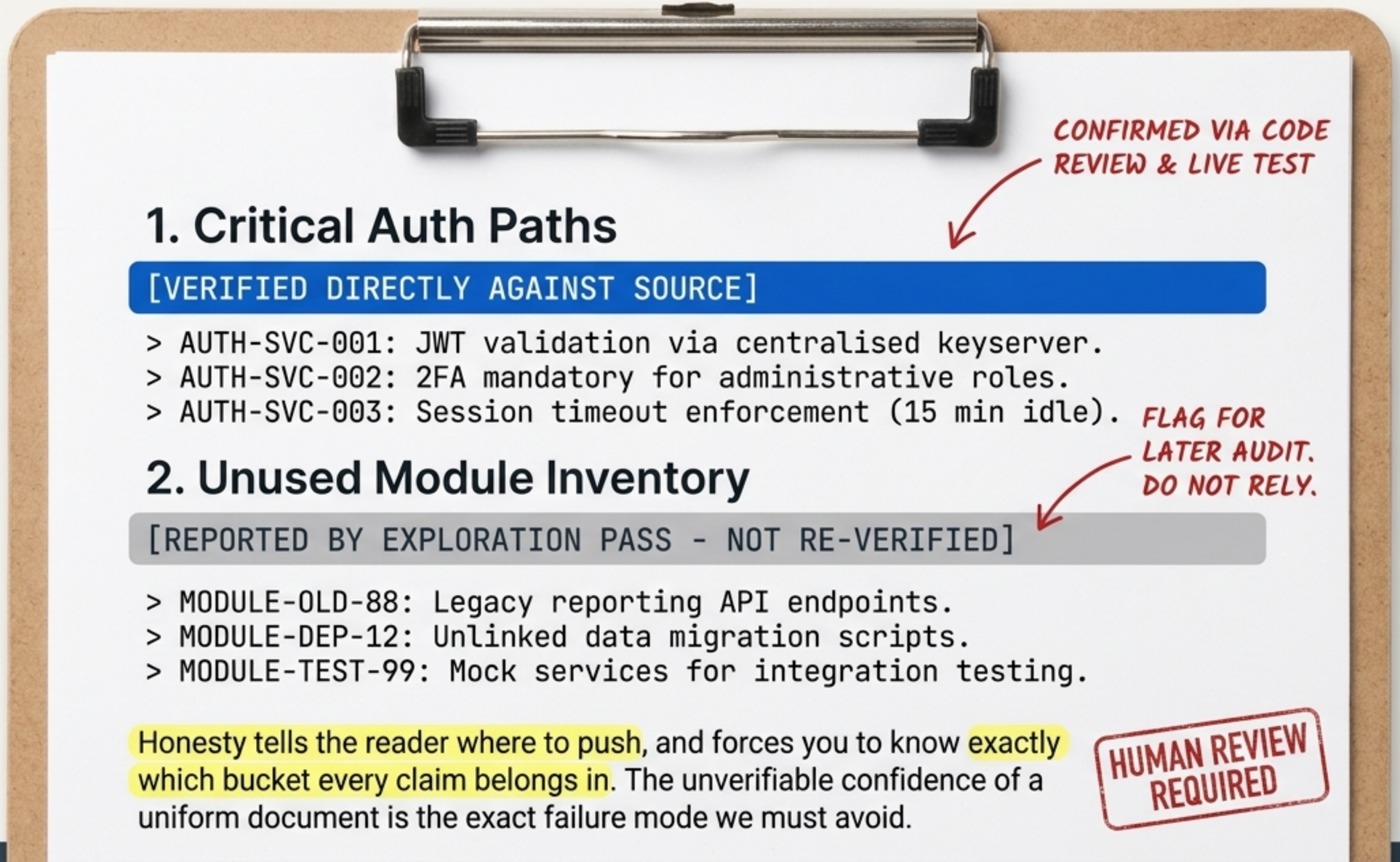
HUMAN REVIEW REQUIRED

Wrong claims are signal, not just noise.

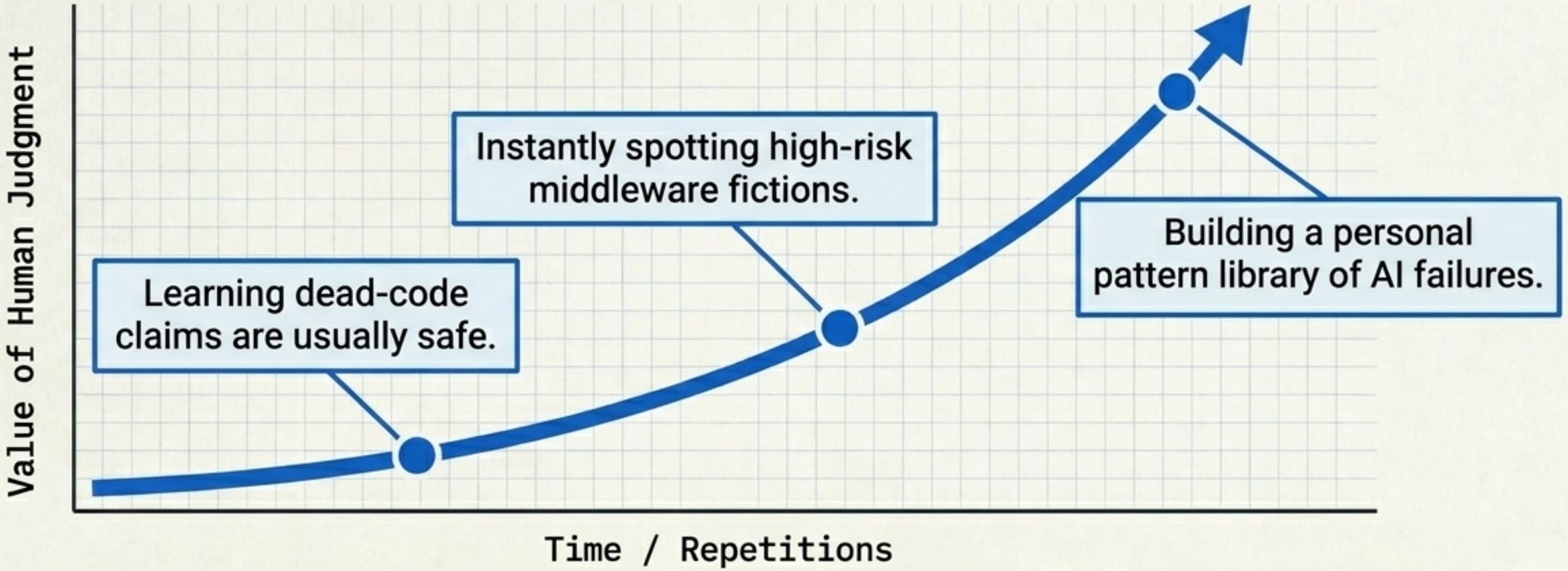
When you strike a false alarm, don’t just delete it. Downgrade it to its true, useful version. The agent’s error almost always points at something real, it just mislabels the severity. Verification is refinement, not just subtraction.

HUMAN REVIEW REQUIRED

Write the verification status directly into the deliverable.



The Compounding Developer.



This isn't a checklist; it's a practitioner skill that gets cheaper with use. Knowing which claims, from which analyses, fail in which direction is pure judgment. It is the part of your work that gets strictly more valuable as the agents get faster.

HUMAN REVIEW REQUIRED