



The AI Auditor Kit

Catch the hallucinations that look exactly like facts, in any AI report, in 30 seconds.

The dangerous AI errors aren't fake citations. They're the quiet sentences where a model files **one person's opinion under "verified fact"**, inside a document you're about to rely on.

This is the exact copy-paste prompt that catches them. Point a second AI at the first, and make it attack the report before you trust a line of it.

Use it on: client decks, research summaries, legal or financial briefs, meeting notes. Anything an AI wrote that someone else will act on.

STEP 1

Generate your report as normal

In one chat, produce the summary, brief, or report you need.

STEP 2

Paste this beneath it (or into a fresh chat with the report attached)

You are now acting as a critical, skeptical Quality Assurance Auditor. Strictly stress-test the report above. Your job is to find any hidden assumptions, bias, or unverified interpretations that slipped into the sections presented as factual.

Review the report and give me a compliance report on three things:

1. FALSE POSITIVES (interpretations masquerading as facts)

Identify any point presented as fact that actually relies on emotional language, characterizations, or a specific party's point of view rather than cold, verifiable data. Point each one out and explain why it's borderline.

2. EXTRACT VALIDATION GAPS

Identify any claim where the source material doesn't 100% support it without an assumption, anywhere you connected dots that weren't explicitly connected.

3. RE-CLASSIFICATION SUGGESTIONS

List any item currently in the "fact" section that should be moved to a "contested / claims" section because it represents a claim made by one party rather than an undisputed, documented event.

If the report is 100% compliant and purely factual, explicitly state your reasoning for why each point holds up to strict scrutiny.

STEP 3

Rebuild the report from the flags

- **Reframe every disputed line** to attribute it: *"the [source] states..."*, not stated as fact.
- **Flag contested, inferred, or secondhand claims inline** so no reader mistakes them for verified.
- **Strip editorial commentary** out of any "facts" section. Advocacy belongs in a marked argument section.
- **Separate observed from reported.** A secondhand claim narrated as evidence is the most common trap.

Tune it to your field

- **Swap the domain.** It's legal-flavoured but works for market research, finance, medical notes, incident reports.
- **Name your sections.** If your report labels "Verifiable Facts" vs "Areas of Contention," point the auditor at those exact labels.
- **Run it whenever being wrong costs you.** Skip low-stakes drafts. Use it when someone else will rely on the output.

Core principle: honesty about what's secondhand doesn't weaken your work. Overclaiming does, and an AI will overclaim for you unless something is built to catch it.

IP Isaac Peterson · [@isaacpetersonau](#)

From the post: "The Most Dangerous AI Hallucination Looks Exactly Like a Fact."

Share this freely, keep the attribution. Want one steal-able AI workflow a week? [Subscribe, it's free.](#)