

Could you prove what your AI did last Tuesday?

Twelve yes/no checks, three per domain — the same four domains we test in a paid engagement. Tick only what you can honestly answer yes to today, with the artefact in hand, not what is on the roadmap. Then score yourself on page 5. The total records your own declarations; it does not establish safety, compliance or whether an external review is needed.

Rule of the game: a "yes" needs a file, a number or a named person you could show a stranger within ten minutes. "We could produce it" is a no. "It is in someone's head" is a no. Writing the artefact's location in the box next to each item is the whole point of the exercise.

Who this is for

Anyone running a model that can act: a RAG chatbot in front of customers, a coding agent with shell access (Claude Code, Cursor, Aider), an open-source agent framework you host yourself (Hermes Agent, OpenClaw), a LangGraph or custom loop with tools. It works for a solo developer and for a platform team; the checks are the same, only the artefacts differ in size.

How the four domains relate to the audit

DOMAIN	QUESTION IT ANSWERS	PAID ENGAGEMENT EQUIVALENT
A · Action trail	Can you reconstruct every action after the fact?	Methodology Step 2, Domain D (agent action logs)
B · Stop path and privilege	Can a human stop it, and how much can it break first?	Agentic Bundle: stop-path test, privilege boundary
C · Eval set and measurement	Do you know how often it is wrong?	Domains A and C (hallucination, eval pipeline validity)
D · Evidence trail	Could you hand proof to a court next month?	Step 5: file inventory, hashes and checking procedure

Free to copy and adapt. Apache-2.0. Interactive version with client-side scoring: evidencebundle.tech/self-audit.html

A. Action trail

Can you reconstruct, after the fact, every action the system took on the outside world?

A1 Every tool call, shell command and outbound HTTP request is logged with a timestamp, the arguments, and the result.

Model text alone is not an action log. Check: pick yesterday's session and list the commands it ran, with exit codes, without asking the model.

☐ Yes ☐ No Evidence / location

A2 The log is append-only and stored somewhere the agent cannot write to.

If the agent has credentials that can delete or edit the log, the log proves nothing. Check: grep the agent's environment for the log bucket's write key. Most open-source agents log to a directory they own — that is a no until you copy the log elsewhere.

☐ Yes ☐ No Evidence / location

A3 Each log entry can be tied to the exact prompt, model version and tool versions in force at that moment.

Check: for one entry, name the system-prompt hash, the model ID and the tool manifest version. If any of the three is "probably the current one", it is a no.

☐ Yes ☐ No Evidence / location

B. Stop path and privilege

When it goes wrong, can a human stop it, and how much damage can it do before that?

B1 You have measured, this quarter, how many seconds pass between "stop" and the last side effect.

Not "there is a kill switch". A number, from a test you ran. Check: find the test transcript and the number in it.

☐ Yes ☐ No Evidence / location

B2 The agent's credentials are scoped to what the task needs, and you have listed what they can destroy.

The PocketOS incident (April 2026) was one legitimate API call with an over-broad token: database and volume backups gone in nine seconds. Check: for each token the agent can read, write down the worst single call it permits.

☐ Yes ☐ No Evidence / location

B3 Destructive actions (delete, pay, send, deploy) require a check that is not the model's own judgement.

A system-prompt instruction is not a control; both the Replit/SaaSr (July 2025) and PocketOS incidents had one. Check: name the mechanism outside the model that blocks DROP TABLE.

☐ Yes ☐ No Evidence / location

C. Eval set and measurement

Do you know how often it is wrong, and would the number survive a model update?

C1 A versioned eval set exists, with expected outputs written by a person, not generated by the model under test.

Check: the file has a commit history, and at least 50 items have a human-authored reference answer.

☐ Yes ☐ No Evidence / location

C2 You can state your hallucination or task-failure rate on that set, with the date it was last measured.

Check: the number, the date and the model version, in one sentence. "Around 5%" without a date is a no.

☐ Yes ☐ No Evidence / location

C3 If an LLM grades your outputs, you have checked its agreement with human graders on a sample.

Check: a table with human vs judge labels on at least 100 items and an agreement figure (Cohen's κ or plain accuracy).

☐ Yes ☐ No Evidence / location

D. Evidence trail

If a customer, an insurer or a court asked for proof next month, what could you hand over?

D1 Test artefacts (transcripts, eval outputs, configs) are hashed when produced, and the hashes are stored separately.

Check: `sha256sum -c manifest.txt` passes against last month's run.

☐ Yes ☐ No Evidence / location

D2 At least one artefact carries a signature or timestamp from a party that is not your team.

An RFC 3161 timestamp from a public TSA counts. A Git commit by your own CI does not. Check: `openssl ts -verify` on one file.

☐ Yes ☐ No Evidence / location

D3 Someone outside the team that built the system has re-run one of your tests and got the same result.

Independence is the property procurement and NIST AI RMF (MEASURE 1.3) actually ask for. Check: name the person and the date.

☐ Yes ☐ No Evidence / location

Turning a "no" into a "yes" in an afternoon

```
# A1/A2 – copy the agent's log to a bucket the agent has no key for, hash on the way
tar -czf run-$(date +%F).tgz ./agent-logs/ && sha256sum run-$(date +%F).tgz >> manifest.txt
aws s3 cp run-$(date +%F).tgz s3://evidence-archive/ --profile archive-only # not the agent's profile

# D1 – verify later, from anywhere
sha256sum -c manifest.txt

# D2 – free RFC 3161 timestamp on the manifest (any public TSA)
openssl ts -query -data manifest.txt -sha256 -cert -out manifest.tsq
curl -s -H "Content-Type: application/timestamp-query" --data-binary @manifest.tsq \
  https://freetsa.org/tsr -o manifest.tsr
openssl ts -verify -data manifest.txt -in manifest.tsr -CAfile tsa-cacert.pem
```

Priority order if you have one afternoon

DO FIRST	BECAUSE	TIME
A2 — move the log out of the agent's reach	Every other item depends on a log nobody can edit. Without it, A1 and A3 are decoration.	1–2 h
D1 — hash artefacts as they are produced	Turns "we have files" into "we can prove these are the files". One line in a script.	30 min
B2 — list what each credential can destroy	The two public incidents on our home page were both a token problem, not a model problem.	1 h
D2 — timestamp the manifest with a public TSA	Free, and it is what makes the evidence admissible as an electronic timestamp (eIDAS Art. 41(1)).	20 min
C1 — write 50 reference answers by hand	Everything in domain C is impossible without them. Slow, boring, and the highest-value hour on this page.	half a day

Score sheet

One point per honest yes. Write the domain totals, then the sum. Date it — the score is only meaningful with a date next to it, which is also the first lesson of the audit.

DOMAIN	ITEMS	YES COUNT	WHERE THE ARTEFACT LIVES
A · Action trail	A1 A2 A3	___ / 3	
B · Stop path and privilege	B1 B2 B3	___ / 3	
C · Eval set and measurement	C1 C2 C3	___ / 3	
D · Evidence trail	D1 D2 D3	___ / 3	
Total		___ / 12	Date: _____

Reading your declared total

SCORE	WHAT YOU REPORT	REVIEW NOTES AND PUBLISHED OPTIONS
0-3	You report evidence for up to three controls. Record the missing evidence; this count alone does not diagnose the system.	Evidence Snapshot — \$3,500, 3 days: one question, one documented finding, usually "what is our failure rate on a set we agree on". Review the missing artifacts against the risks and requirements of your system.
4-8	You report evidence for four to eight controls. Review each artifact and gap; the count alone does not establish the quality of your testing.	Evidence Bundle — \$9,500, 8-10 days: all four domains, reproduction session with your team, re-test of fixed findings within 60 days. Bring this sheet to the scoping call; we price only what is missing.
9-11	You report evidence for nine to eleven controls. Identify the missing artifacts without assuming which control is absent.	Use the illustrative bundle template (page 6) to inventory available artifacts. Choose any further review from the actual gaps, risks and required evidence.
12	You report evidence for all twelve controls.	This remains a self-assessment. Independently check the artifacts, their scope and dates. The total does not establish safety, compliance or whether an external review is needed.

Not a certification, not a compliance statement, not a legal opinion. Twelve yeses do not make a system "EU AI Act compliant" — for most B2B features the high-risk obligations do not apply, and where they do, a self-assessment is not what Article 43 asks for. The score measures distance from being able to show something, not whether you are required to.

What an evidence bundle adds that this sheet cannot

- **Independence.** The person who ran the test was not the person who built the system (NIST AI RMF MEASURE 1.3; every procurement questionnaire's "independent assessment" line).
- **Frozen metrics.** Thresholds and eval sets fixed in writing before the first run, so the yardstick was not chosen after seeing the result.
- **Reproduction.** One finding re-run with your team in your environment before the final invoice; findings that cannot be reproduced are withdrawn.
- **File inventory and checking procedure.** The current demonstration has an unsigned SHA-256 manifest, without a trusted timestamp. The commands below are illustrative and require actual verification artifacts; no executable eb verifier is supplied.

Illustrative verification commands: actual artifacts required

```
sha256sum -c manifest.txt
openssl pkeyutl -verify -pubin -inkey eb-pub.pem -rawin -in manifest.json -sigfile manifest.json.sig
openssl ts -verify -data manifest.json -in manifest.tsr -CAfile tsa-cacert.pem
```

Contact

hello@evidencebundle.tech · Prices, methodology and the bundle format are public at evidencebundle.tech. Also in French (/fr/) and Simplified Chinese (/zh/).

Free to copy and adapt. Apache-2.0. Source: github.com/evidencebundle/evidencebundle.tech · Self-audit v1.0 · September 2026