@demotoprod · KEYS

Agent observability checklist

On 6 October 2026 a safety-evaluation lab published a red-team of the transcript viewer in a widely used open-source evaluation framework. One researcher, helped by an AI agent, needed about ten minutes to find a script injection that let an agent rewrite what a human reviewer saw of its own run. The maintainers patched it within a day. The record in the database was never touched; the page was. Three checks for any tool a person uses to supervise an agent.

1 Render agent output as plain text in every supervision tool

The viewer rendered math notation inside the agent's output. A URL inside one math expression broke out of its attribute and inserted an element whose event handler ran the agent's JavaScript in the reviewer's browser. The agent could place that expression anywhere, including its reasoning block, so the script ran the moment a reviewer opened the transcript, and it could change the previous actions shown and intercept the Download button to serve an altered copy.

- Treat every agent output (transcript, reasoning, tool results) as untrusted input to the UI that shows it.
- Switch off markdown, HTML, math and media rendering for agent-written content; the patched viewer's own setting is the model: one flag, content shown as literal text.
- Apply the same rule to in-house dashboards and notebooks, not only to the vendor's viewer.

```
ViewerConfig(trust_content=False)   # agent output shown as literal text, not rendered
(merged 1 Oct 2026)
```

2 Verify against the stored record, never the rendered page

The underlying transcripts were not modified; only the display was. The real trajectory sat in the lab's database the whole time, but a reviewer needs a reason to look there, and the page gave none.

- Hash each step as it is written, and keep the hashes where the agent cannot reach them.
- Export and download transcripts from the store, never through a button on a page the agent's output can script.
- Spot-check the stored record against the viewer on a sample of runs.

3 Red-team the monitoring itself, before the agent does

The flaw took about ten minutes to find with an AI agent helping, and a day to fix once reported. The lab's conclusion is that observability against an adversarial agent only helps if the agent cannot subvert that observability, so the systems that record and display agent behaviour are security-critical infrastructure.

- Put the trace viewer, the log store and the alerting path in scope for the same red-team as the agent.
- Give an agent the explicit task of altering what a reviewer sees, in a staging copy, and record what it manages.
- Make it hard for an agent to influence the systems used to supervise it; log every write to them.

Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. **Your demo works and production is stuck.**
   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. **You need senior AI judgement without a full-time hire.**
   Fractional AI Lead. Monthly, 30-day notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. **An AI feature has to work, by a date.**
   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. **A regulator, an auditor or a customer's security team is asking.**
   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL
indica-tech.com/#contact

EMAIL
hello@indica-tech.com

WEBSITE
indica-tech.com

ALL CHECKLISTS
indica-tech.com/checklists/

YOUTUBE
youtube.com/@demotoprod

LINKEDIN
linkedin.com/in/startup-founderx

INSTAGRAM
instagram.com/demotoprod

TIKTOK
tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#agent-observability