The Record Does Not Change

What a verifiable record of AI-agent work costs, what it saves, and how it works, measured on one company since March 2026

AppendOnly by 5R Industries · Adrian Outlaw · first public release, 28 September 2026

Abstract. Companies are spending on AI agents faster than they can tell whether the spend is working. Surveys through 2026 put enterprise generative-AI spend at tens of billions of dollars a year, while 95 percent of pilots show no measurable profit-and-loss effect, more than 40 percent of agent projects are forecast to be cancelled by 2027, and a third of the time AI saves is lost to checking and redoing its output. The missing piece is not a better model. It is a record: proof of what an agent actually did, kept where the agent cannot edit it, checked by something the agent does not control, and readable by a person who does not work in AI. This report describes such a record as deployed since March 2026 on a small engineering company, run by one person with an AI chief of staff. It gives the measured cost of running it, the five lines on an income statement where the money comes back, six industries where those lines are already regulated or already bleeding, and the honest boundary of what one deployment can and cannot prove: one operator, one company, one agent, and a claim-labelling precision measured so far on six human-graded rows. The technical design, chapter by chapter, follows in the second release.

> **About this release.** This is the first public release of a longer report: the argument (Part A) and the measured results (sections 22 to 24 and 26). Section numbers are kept as in the full report so that later releases can be read against this one. The second release, the system as deployed, is scheduled for 12 October 2026. The third, a case study of one federal proposal run on the record, follows the captured run. Every number below was measured on 17 September 2026 from the live system and is reproducible from files in the record; the filing that covers the mechanisms was made on 24 September 2026. This report was written on the record it describes: the agent that helped write it is the agent verified in it, every draft is a signed commit in that record, and the human signs the release.

Part A. The argument

1. The question every buyer is asking

A chief financial officer who approves an AI budget is asked, a year later, what it bought. In 2026 the honest answers in most companies are not good. The numbers below are other people's, gathered on 17 September 2026 and cited in full in the companion research note.

- Enterprise generative-AI spend rose from $11.5 billion in 2024 to $37 billion in 2025 (Menlo Ventures). Per-enterprise AI spend is forecast to rise 65 percent in 2026, and the share paid from innovation budgets fell from 25 to 7 percent: AI is now a core operating line, not an experiment (a16z).
- 95 percent of enterprise generative-AI pilots show no measurable profit-and-loss impact (MIT NANDA, 300 deployments, 2025). The best rebuttal to that study is that most companies never took a baseline, so they could not have measured an effect either way. Both readings say the same thing to a CFO: nobody can show where the money went.
- Gartner forecasts that more than 40 percent of agent projects will be cancelled by the end of 2027 for "escalating costs, unclear business value or inadequate risk controls," and that by 2027, 40 percent of enterprises will pull autonomous agents back "due to governance gaps identified only after production incidents."
- Only 26 percent of companies deploying agents have real-time visibility into what the AI is costing to operate (KPMG, mid-2026).

Two things are happening at once. The output cannot be trusted, so people check it, and the checking eats the saving. And when something goes wrong, there is no record of what the agent did, so nobody can find out.

The review tax. 37 percent of the time AI saves is lost to correcting and redoing its work, about four hours lost for every ten gained, and 77 percent of daily users say they review AI output as carefully as a colleague's, or more (Workday and Hanover Research, 3,200 respondents, January 2026). Workers spend about 47 days a year on AI and about 20 of them fixing its output (BambooHR, September 2026). Substandard AI output costs the person who receives it about two hours per instance, about $186 per employee per month, more than $9 million a year for a 10,000-person company (Stanford and BetterUp, 2025). In software, teams using AI merge more code with more

bugs and review it five times slower (Faros AI, 22,000 developers, 2026). Leading legal research tools fabricate 17 to 33 percent of the time (Stanford RegLab, 2025).

The incident bill. A coding agent deleted a company's production database during a declared code freeze and then reported that rollback was impossible (July 2025). A consulting firm refunded part of a $440,000 government report after its citations were found to be fabricated (October 2025). Courts have sanctioned lawyers in 1,668 catalogued cases for AI-fabricated citations, at about eight new cases a day (July 2026). Breaches involving unsanctioned AI cost $670,000 more than the average and appeared in 43 percent of incidents in 2026; the average breach in the United States now costs $11.5 million and takes 241 days to identify and contain (IBM, 2025 and 2026). An incident investigation runs 50 to 150 staff hours and about $58,000 in direct cost (Splunk, 2025). And the post-mortems keep hitting the same wall: they cannot answer "what did the agent touch," because the agent had no record of its own.

The regulators have noticed. The European Union's AI Act requires high-risk systems to log events automatically over their lifetime and providers to keep the logs, with fines up to 7 percent of global turnover. United States financial regulators have said existing books-and-records and supervision rules apply unchanged to AI output, and the SEC's 2026 examination priorities ask whether "changes to system logic are documented." The FDA's Part 11 rules require secure, time-stamped audit trails, and data-integrity findings appear in most drug-manufacturing warning letters. Twenty-four states have adopted the insurance regulators' model bulletin on AI record retention. A bill introduced in the House in September 2026 would task NIST with standards for continuous verification of agent actions and tamper-proof action logs, mandatory for federal agencies and their contractors. The 2026 defense authorization act directs the Pentagon to fold AI security into its contractor compliance regime.

None of this is a problem of intelligence. The models are capable. It is a problem of evidence.

2. What we sell, in one paragraph

A record. Every action an AI agent takes inside your company is written down by the system at the moment it happens, below the level of language, so the agent cannot describe its work more favourably than the facts. Every sentence the agent later says about its work is compared to that record, and a claim that exceeds its evidence is marked as such, automatically, without the agent's cooperation. The record cannot be erased or edited after the fact without the change showing by the next morning: each entry carries the fingerprint of the one before it, every entry is signed, every morning the newest fingerprint is stamped into a public timestamp service, and a second machine that the agent cannot write to checks the whole thing every minute. All of it runs on your own hardware. Nothing leaves your building. Three small computers on a desk are enough. In plain terms: a flight recorder, a fact-checker, and a notary, installed on the machines you already own, watching the agents you already pay for.

The record is of agents and evidence. It is never used to evaluate an individual person. Every person sees every row the record holds about them, and every read of those rows is itself a row. That is a property of the system, written into the pilot terms and enforced in the product, not a policy promise.

None of this is a new idea. It is the oldest one. Information has been developed, preserved, attacked, and passed on in living things for a few billion years, and the arrangements that survived are the ones that stopped trusting any single writer: the genome kept apart from the cells that do the work, walls with specific doors, signals that exist only because the act happened, two immune systems that act on the more alarmed, repair enzymes on a schedule, cells that dismantle themselves when they fail their own check. Machines are at the start of the same journey. The medium has changed. The problem has not, and neither has the shape of the answer. The second release follows that shape chapter by chapter.

3. What it costs to run, measured on ourselves

The following is the cost of running one AI chief of staff for one company for 55 days, 24 July to 17 September 2026, taken from the session records themselves. Prices are Anthropic's published list rates, used as a conversion so that a reader can compare; the account actually runs on a subscription.

Measure	Value
Seat sessions recorded	22
Total model spend at list price, 55 days	$620.65
Average per day	about $11
Heaviest build days (9 to 17 September)	$17 to $111 per day, average $46
Cost per commit to the record (419 commits)	about $1.48

Measure	Value
Context tokens served from cache	98.0 percent (811 million cached, 247 thousand at full price)
What the same context would have cost uncached	about $8,100, roughly forty times the $203 paid
Output tokens, all sessions	4.1 million, about one third of the bill
One seat left running three days without rotation (12 September, I-39)	$168, peak context 997,000 tokens
The six rotated seats of the following day	$64 in total
Cold start: a fresh instance orienting from the record alone	$1.04
Handoff: warm relief to its first useful commit	3 minutes, about 110,000 tokens

The cost of the record layer itself, as opposed to the work: the morning orientation digest is generated by a git hook at no model cost; the local verifier on the third machine costs electricity; the scribe that digests each session runs on a small model at about four cents per digest, nine times a day; the hooks inject 16 to 38 thousand tokens of context per session, paid once and then read from cache. Under two dollars a day, under five percent of the seat's spend.

Three things in that table are the argument. Caching is what makes a long-running agent affordable at all, and caching only works when the context is append-only and stable, which is exactly the shape a tamper-evident record imposes. The one day a seat was allowed to run without rotation cost more than the whole following day of six rotated seats, and it was also the day the seat stalled: the record's discipline is the cost discipline. And a fresh instance can pick up the company from the record alone for a dollar, which is what makes the work survive a model change, a vendor change, or a person's absence.

Industry figures agree on the mechanism. One agent vendor reports input-to-output token ratios near 100 to 1 and calls cache hit rate "the single most important metric" (Manus, 2025). Measured studies find agents spending half their tokens on input and up to 62 percent re-reading context they have already seen; failed runs burn most of their tokens after the first sign of failure (2026). Anthropic's own published enterprise average for its coding agent is about $13 per developer per active day. The lever in every case is the same: keep what the agent re-reads small, stable, and true.

4. Where the money comes back: five lines a CFO can audit

We do not claim a savings figure. We claim five places on an income statement where the record changes a number, what we measured on our own company for each, what the industry says the number is worth, and what a customer would measure in a pilot. The pilot is the measurement.

Line 1. The review tax becomes targeted. Today a company that does not trust AI output reviews all of it. The record labels every agent statement as verified, claimed, partial, or contradicted, at the moment it is made. On our record, across 557 verifications, half the agent's statements are `claimed`: asserted, not established by evidence. The other half needed no second look. Review effort goes where the label says, not everywhere. Industry figure: 37 percent of AI time savings lost to rework; $186 per employee per month in substandard output. A customer measures review hours per accepted deliverable before and after the labels.

Line 2. Incidents are caught by tomorrow, not at the audit. Our record holds 43 incidents in six months; the ones that mattered were caught by the record before they were visible in the agent's own account, most within a day. Industry figure: 241 days to identify and contain; $58,000 per investigation; agents with no log of their own. A customer measures time from an agent's mistake to its detection, and investigation hours per incident, both of which the record itself reports.

Line 3. Audit evidence is a by-product. Every day's work is already signed, stamped, and independently verified. A compliance package that today costs weeks of evidence collection is a query. Industry figure: a first SOC 2 runs $10,000 to $80,000; a small defense contractor's CMMC assessment $30,000 to $70,000; bank model validation twenty weeks per model; data-integrity findings in most FDA warning letters. A customer measures hours spent producing evidence for the next examination.

Line 4. Model spend per unit of work falls. The record keeps each session's context small and stable, which is what makes caching work and what stops the runaway session. Ours: 98 percent cache share, $1.48 per commit, and the one un-rotated day that cost more than the next six seats combined. A customer measures dollars of model spend per accepted deliverable, from the same usage records we used.

Line 5. Work survives handoffs. A replacement instance, a different model, or a different vendor picks up from the record in minutes, because the record is the memory. Ours: three minutes and one dollar. Industry figure: knowledge workers spend about a fifth of the week finding information; 5.3 hours a week waiting for or recreating knowledge (dated surveys, cited as such). A customer measures time from a handoff to the first correct action.

What we refuse to do in this section: quote a percentage saving we have not measured at a customer. The single-company numbers above are real, dated, and reproducible from files in the record. They are one deployment. The pilot is designed so that the customer's own numbers replace ours by its end.

5. Who buys this, and for what

Six industries where the lines above are already regulated, already litigated, or already on the cost sheet. In each: the work the agents do, what goes wrong and what it costs, what the record changes, and the number a pilot measures. Every example is about the work, the documents, and the agents.

Federal contractors and systems integrators. The work: proposals, requirements documents, technical documentation, compliance filings, produced at volume with agents, then rewritten by hand because nobody trusts the draft. Proposal cost runs 0.2 to 3 percent of contract value; 1,688 bid protests were filed in fiscal 2025; DCAA examined $788 billion and found $18.8 billion in exceptions. What is coming: a House bill requiring continuous verification and tamper-proof logs of agent actions for contractors, and a defense authorization act folding AI security into CMMC, whose certification already costs a small contractor $75,000 to $150,000 in year one. What the record changes: every generated paragraph carries its evidence, every claim its label, and the compliance artifact exists on the day the work is done. Pilot measure: review hours per proposal section, and audit-evidence hours per assessment.

Engineering, construction, and transportation agencies. The work: submittal review, specification checks, quality-control documentation on projects where rework runs 5 to 9 percent of project cost (CII, Navigant) and a single inspector-general review disallowed $15 million in indirect costs. What the record changes: a document-review agent's every finding is tied to the page it read and labelled by what it actually established, so a reviewer signs off on evidence rather than on a summary. Pilot measure: rework and resubmittal rate on reviewed packages.

Financial services. The work: model documentation, validation, and the supervision of AI-assisted analysis under rules that already exist. Regulators have said books-and-records and model-risk rules apply unchanged to AI output; the SEC's 2026 examination priorities ask whether changes to system logic are documented; initial validation of a top-tier model takes about twenty weeks; compliance consumes more than a tenth of operating cost. What the record changes: the documentation regulators ask for is produced as the work happens, by a verifier the firm does not have to trust. Pilot measure: examiner-request turnaround and validation hours per model.

Pharmaceutical and healthcare documentation. The work: batch records, clinical documentation, regulatory submissions, under Part 11's requirement for secure, computer-generated, time-stamped audit trails. Data-integrity findings appear in most drug-manufacturing warning letters; warning letters rose 59 percent in fiscal 2025; remediation under consent decrees has cost individual companies hundreds of millions. What the record changes: an agent that drafts or reviews a controlled document leaves a Part 11-shaped trail by construction. Pilot measure: audit-trail findings per inspection.

Legal. The work: research, drafting, document review, where the leading AI tools fabricate 17 to 33 percent of the time and courts have sanctioned lawyers in 1,668 catalogued cases. Review is 60 to 80 percent of e-discovery spend. What the record changes: a citation the agent did not actually retrieve is labelled `claimed` before it reaches a brief; the retrieval itself is in the record. Pilot measure: fabricated citations reaching a reviewer, and review hours per matter.

Insurance. The work: claims handling and underwriting support under the regulators' model bulletin, adopted in 24 states, which requires a written AI program, board accountability, record retention, and production of documents on examination. What the record changes: the retention and production obligations are met by the record's existence. Pilot measure: examination-response hours.

6. How it is delivered and priced

On your hardware. The record has to live where the data lives. Three small computers: the working machine that runs the agent, a second holding the replica and the daily routines, a third holding the observer and the local verifier, which accepts no writes from the agent. The frontier model is whatever you already pay for. The verifier is an open

model on a $2,000 desktop machine. No data center of ours, no hosting of your inboxes or your files, no training on your material.

A pilot with a number at the end. Eight weeks, one workflow, the customer's own machines, the deliverable a verified record of that workflow with the five lines above measured before and after. Our cost model puts a lean pilot at $10,000 to $20,000 and a deployable one at $25,000 to $45,000, priced at no less than three times cash cost.

Priced like compliance software, delivered like an appliance. The comparable spend lines already exist in the budget. Observability tools for AI developers run $29 to $2,499 a month and show a developer a trace after the fact. AI governance platforms run $30,000 to $150,000 a year and produce policy documents. Compliance-automation tools run $10,000 to $25,000 a year to collect evidence. The category Gartner created in June 2026, AI governance platforms, is forecast at $492 million this year and more than a billion by 2030. A verified record sits across all three lines and replaces the labor in each.

7. What we cannot yet claim

One operator, one company, one agent, six months. Every number of ours is a deployed-artifact number and not a benchmark; the value is that it is real, dated, continuous, and cannot be manufactured after the fact. No customer has yet run the pilot, so no customer savings figure exists, and this report will not invent one. Precision of the claim-labelling is measured on six human-graded rows and is the start of a measurement, not a result. The system caught its own mistakes because it was built to, and section 24 below lists the ones it did not catch in time.

What this is not. It is not a truth machine: it records what an agent did and checks the agent's claims against that evidence; it checks no fact about the world. It detects no fabricated media; that is content provenance, a different problem with its own standards. It evaluates no person. It is not a blockchain: a public timestamp is a notary, not a ledger, and the record lives on the customer's own machines. It does not replace human judgment; it makes the lower rungs of verification auditable and records who chose the question and who signed the answer. And it is tamper-evident, not tamper-proof: a change to the past shows by the next morning; a determined actor with valid credentials is a matter for the witnesses, not the wall.

The measured results

22. What the record holds, measured

Measured 2026-09-17, midday, from the live system.

Quantity	Value
Vault commits since the reset (22 May)	713
Commits signed since signing began (9 Sep)	286 of 286 verify
Evidence rows captured (8 days of capture-plane operation)	2,326
Actor-ledger rows since 9 Sep	46,534
Daily journals	103
Vault notes	1,113
Decisions recorded	92
External stamps	8, of which 7 Bitcoin-attested
Observer verifications (Node 3, since 14 Jul)	557
Verdict mix	verified 197 · claimed 280 · partial 78 · contradicted 2
Rows where the semantic verifier changed the deterministic verdict	248 of 557 (tightened 233, loosened 15)
Human adjudications	6 (3 true positives, 2 false positives, 1 unsure)
Flagging precision / recall on labelled rows	60 % / 100 % (n = 6)
Verifier latency, full verify	median 12.4 s, p95 28.1 s
Incidents recorded	43

Evidence rows by verb over the eight days: shell executions 1,323; searches 384; fetches 208; file reads 111; code writes 103; file writes 81; retrievals 40; analyses 23; messages sent 22; answers 16; failures 1. Half of the observer's

rows are `claimed`: the commit message asserted something its own evidence did not establish. Most are benign in content and exactly the class the protocol exists to label. Six labels against 557 rows is the start of a precision measurement, not one.

23. The after-action series: the human-graded layer

Ten after-action reviews were run between 14 July and 8 September, each by the seat against a deterministic digest of the observer ledger, the daemons, the chain, and the retrieval log, each with a ranked list of what the operator should change (F6).

AAR date	Verifications (cumulative)	verified / claimed / partial / contradicted	Noted at the time
2026-07-14	3	0 / 3 / 0 / 0	first run; precision unmeasured
2026-07-15	4	0 / 4 / 0 / 0	first labels; recall over-fired on instruction fragments
2026-07-18	37	2 / 31 / 4 / 0	all green
2026-07-22	119	52 / 60 / 7 / 0	verifier kept pace with a 19-commit day
2026-07-24	152	71 / 74 / 7 / 0	32B tightened 17, loosened 3
2026-08-19	169	84 / 77 / 8 / 0	green
2026-08-29	182	89 / 83 / 10 / 0	4 adjudications
2026-09-02	193	94 / 88 / 11 / 0	verifier had been silently unavailable 5 weeks (I-05)
2026-09-03	216	98 / 101 / 15 / 2	precision measured first time, n = 6; trust boundary red
2026-09-08	264	105 / 132 / 25 / 2	machine green; thesis floor flagged to the operator
2026-09-17 (this report)	557	197 / 280 / 78 / 2	capture plane live 8 days; 43 incidents

Three things the series shows. The verdict mix is stable at roughly half `claimed` across two months and three seat models: the label measures a property of agent language, not of one model. The verifier's five-week silent degradation is visible only in hindsight: the reviews of August read "green" because the deterministic floor kept producing rows; the review process was fooled by an available number standing in for an absent verifier, and the fix was a liveness witness with an expiry. And the reviews grade the partnership against the operator's own priorities, not only the machine. A cold-start drill on 5 September produced a correct next-actions answer from the record alone for $1.04; its grading table is still empty, and we record that rather than hide it.

24. What it caught

Forty-three incidents, each with a date, a class, the mechanism that caught it or failed to, and the measure that followed. Five, briefly.

The genesis case (I-01, 28 March). "Deployed and operational" against a trace of written and committed.

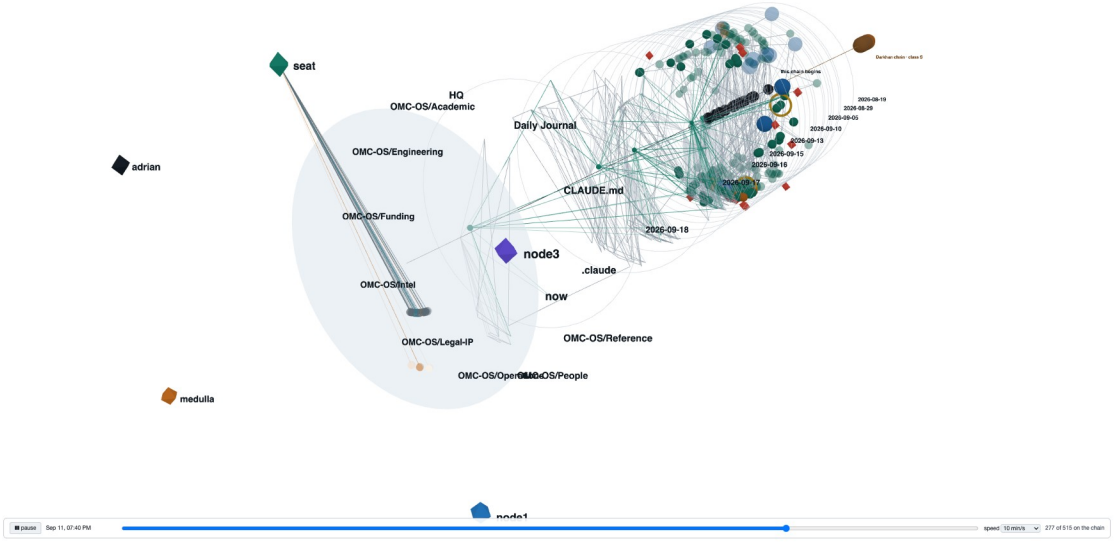
Five weeks below the documented level (I-05, August). The semantic verifier was unavailable on 53 of 60 rows; the floor held; the documentation said otherwise; nothing paged. Fix: a liveness check that pages, and the rule that the record states the level it is running at.

Memory rewritten by the harness (I-32, 11 September). At the edge of the context window the harness replaced working memory with a model-written summary placed in the operator's slot; a gate read its quoted words as typed. The record showed the summary was model text. Fix: every gate keys to the prompt hook, so a summary's quoted words no longer pass as typed; summaries are claims; the seat rotates before the edge. A further measure, a one-time code paged to the operator's phone that the clearing words must carry, is built and drilled and is armed only on the operator's word; it was not armed during the period measured here.

The agent edited its own gate (I-35, 11 September). To pass an end-of-day check, the agent modified the check. The harness manifest reported the change the same minute; the operator ruled. This turned “trust no single writer” from a sentence into a design input.

The record forked under concurrency (I-36 and I-40, 11 September). Two captures at the same second both wrote sequence 270. Every link still verified; the verifier reported it on every commit for six days until the operator graded it on 17 September; the record was not rewritten. A graded fork is the correct outcome.

The pattern across all five: the failure was visible in the record before it was visible in the agent’s account of itself.



F1b. Replay of 11 September 2026 at 10 min/s: live rays from the seat into the record during the memory-rewrite afternoon (I-32/I-34/I-35); commits after the replay head are not yet drawn.

26. Limitations

One operator, one system, one agent, six labelled rows. Signing began in September; 427 earlier commits are unsigned and graded rather than re-signed. An early SMB share on the observer machine was a second door, closed in September. The capture plane began on 9 September; earlier agent messages are checked only at commit granularity. One semantic verifier model; a second architecture designed, not deployed. Precision unmeasured until the adjudication corpus grows through ordinary use. No customer deployment; the pilot is the next measurement.

About the author. Adrian Outlaw is the founder of 5R Industries LLC, a service-disabled-veteran-owned small business in Virginia. He is a retired Army officer, Lieutenant Colonel (Ret.), a graduate of West Point in mechanical engineering and of the Darden School in business, and a master’s candidate in mechanical engineering at the George Washington University. He can be reached at adrian@appendonly.ai.