Assurance at the Boundary: The Level Below AAL-1

June Kim
Independent Researcher
june@june.kim
ORCID 0009-0005-3153-9396

July 19, 2026

Abstract

Frontier AI Auditing proposes four AI Assurance Levels in which assurance deepens as auditors gain access to non-public information, at $300,000 to several million dollars annually by its authors' own estimates. That framework takes from financial auditing, engineering, and arms control their inspection rights and accreditation powers. Here we present the level below their first, adoptable immediately at the cost of an auditor's time: assurance attached to the artifact at the boundary, against a declared standard, under a named signature. The design is not novel. All but two of the problems their paper poses have solutions already running in one of those fields, yet none has been tried on AI. The stance we take is that entitlement to a claim comes from replay, so assurance can only attach where replay is possible, which is the boundary. The practice has already run from the public side. Audits of eight publicly accessible benchmarks surfaced severe defects, each with a re-runnable receipt, and no published audit shows internal access reaching anything boundary verification could not. What the receipts need now is a reader. AVERI could hold them today, without buying any access at all.

1 Problems posed by Frontier AI Auditing

1.1 The four assurance levels

Brundage and coauthors propose four AI Assurance Levels for third-party audits of frontier AI companies (see glossary). The levels are ordered by confidence, and access is the input they name first: higher levels "tend to require greater access to non-public information relative to lower levels, larger allocations of time and talent, and more sophisticated infrastructure and analysis." AAL-2 adds "greater access to non-public information, less reliance on companies' statements, and a more holistic assessment of company-level risks." At the top, auditors "can rule out the possibility of materially significant deception by the auditee."

The paper is candid about feasibility: "The two highest assurance levels (AAL-3 and AAL-4) are not yet technically and organizationally feasible, but we outline research directions to change this." Its own estimates, offered with stated high uncertainty, put AAL-1 engagements at $300,000 to $600,000, AAL-2 at $1,000,000 or more, and AAL-3/4 at several million dollars annually.

1.2 What the framework leaves on the shelf

The framework gets four things right, and this response keeps all of them: the independence requirements, the cooling-off periods, the revenue diversification, the warning against companies shopping for favorable auditors. The paper also poses its hard problems, naming completeness, gaining confidence that "there aren't material omissions that would change the audit conclusions," along with accountability, coordination, and cost.

1.3 Each problem has a deployed precedent

Each of these problems has a precedent with decades of operating history, and the paper knows the fields that hold them. It works the financial-auditing analogy through an appendix surveying six assurance domains,

proposes a PCAOB analogue, and builds AAL-4 on the arms-control verification literature. What it takes from them is the internal controls, the inspection rights, the accreditation body with revocation power. What it leaves on the shelf is the discipline that made them work before any of that machinery existed: assurance attached to the artifact at the boundary, against a declared standard, under a named signature. Assembled, those three are the level below their first.

The contribution here is that second reading, and the epistemic frame that forces it. When entitlement to a claim comes from replay, assurance can only attach where replay is possible, which is the boundary. For all but two of the problems posed, the practice already runs, and it runs without the access the higher levels buy.

Problem they pose	Deployed precedent	What adoption looks like for AI
What should the audit object be?	Financial audit: statements at the boundary, against a declared standard; interior out of scope	Audit the records that cross: benchmark contracts, eval reports, deployment claims
Completeness ("no material omissions")	Materiality standards; auditor's interior access scoped as instrumental to the statements	Declared scope and loss profile per record; interior access spent testing the boundary map, and the opinion indexed to the declaration
Auditor accountability	Named opinion letters; PE stamps; professional licensure	Named signatures on audit records and release decisions
Coordination and recognition	Standard-setting bodies, private first, statutory backstop later	A body that reads records, remembers signatures, and standardizes what crossings must disclose
Audit cost	Economy of research: highest yield per dollar first	Exhaust free boundary checks before buying access; the cost-ordered checklist exists
"Treaty-grade" assurance (their AAL-4)	Arms-control verification: declared inventories, with adversarially credible inspections scoped to the declarations	Declared crossings, with adversary-replayable checks scoped to the declarations

Their own framework already asks for records, at its top two levels. AAL-3 wants "detailed logs and compute accounting with cryptographic provenance (e.g., 'proof of training')" and continuous drift and change detection; AAL-4 adds "tamper-evident logging across infrastructure using formally verified open-source cryptographic provenance tooling." The assurance-carrying objects at their highest levels are records; the white-box access around them is scaffolding.

The disagreement is over ordering. They reach record discipline only at AAL-3 and AAL-4, after years of access-building, at several million dollars annually. Yet the same discipline is already deliverable at the public layers today, for an auditor's time.

No form of governance survives translation into another domain intact, and none of these would. The claim is about the cost of the trial rather than the completeness of the fit. Each row has a running implementation to copy from, which is what makes the attempt cheap. No one has made it.

Two problems survive the matching. Omissions from the declared boundary map get half a row, since materiality standards reach what a firm has declared and not what it has left off; harm at first crossing gets none. Both remain unsolved, and both are where the research budget belongs.

2 The boundary as audit object

2.1 What financial auditing settled

Financial auditing settled its audit-object question long ago. The audit attaches to the financial statements, the numbers that cross the boundary from firm to investors, verified against a declared public standard. Managerial accounting is out of scope. How the firm formulates its internal spreadsheets, allocates costs between divisions, or builds its dashboards is its own business. The auditor does enter the interior, sampling ledgers and confirming receivables, but that access is instrumental to verifying the boundary statements, and the opinion letter certifies the statements, never the interior. Assurance attaches to the boundary artifact.

Segregation of duties is why a firm's self-reported numbers cannot stand alone, however honest the firm. A lab holds execution, custody, recording, and authorization at once when it trains a system, evaluates it, reports the score, and decides the release.

2.2 Separation into three parties

The cure is accounting's cure, separation into parties: developers build, evaluation organizations measure, auditors verify. The monitoring logic is agency theory's oldest (Alchian and Demsetz 1972; Jensen and Meckling 1976); the party assignment is this response's. The separation fixes the boundary map. A crossing exists wherever work changes hands, so the map is induced by the division of labor rather than drawn by the audited.

Enron demonstrated the merge, and Sarbanes-Oxley codified the separation: officers certify the statements personally, an independent committee holds the auditor's engagement rather than the management being audited, and the statute bars auditors from selling consulting to their auditees, so the three interests cannot re-merge through the payment channel.

Today's frontier evaluation engagements run in the shape that statute was written to end, with the lab under evaluation commissioning the work, paying for it, holding the engagement, and often controlling what is published. The observation indicts the arrangement, not the people inside it; auditing applies segregation of duties to honest firms. And the audit layer terminates the chain. A re-runnable finding needs no fourth party to vouch for it.

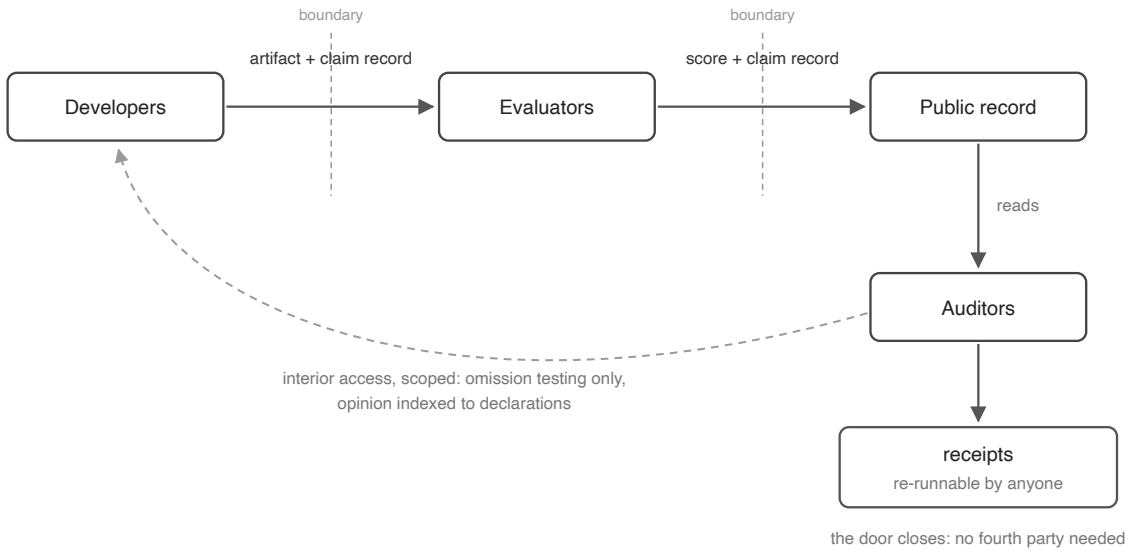

Figure 1: The three-party separation. Developers hand an artifact and claim record across a boundary to evaluators; evaluators hand a score and claim record across a second boundary to the public record. Auditors read the public record and publish receipts re-runnable by anyone, closing the chain with no fourth party. A dashed return arrow marks interior access, scoped to omission testing with the opinion indexed to declarations.

Frontier AI Auditing's glossary takes "reasonable assurance" and maps it to AAL-2. But AAL-2's stated object is "company-level risks." The framework takes the term of art from a regime whose audit object is the boundary statement and aims it at the interior at large. Where mature practice does certify the interior, it does so only against specified criteria and declared assertions, criteria the AAL framework does not yet name.

2.3 Claims, crossings, and pinned artifacts

Claim carries the Verifiable Knowledge sense throughout, defined with its companions in the glossary: a statement shipped with the check that would refute it, entitlement conferred by replay. A score without its

tasks, a safety statement without its test, a headline rate without its labels is an assertion, and its verdict is untrue rather than false.

The same boundary shape recurs at every layer of the AI stack, because the same failure recurs. A claim becomes uncheckable at the moment its provenance is dropped, and crossings drop it. A benchmark's contract crosses when the benchmark publishes. An eval's claim crosses when the evaluator reports the number. A deployment's claim crosses when the system reaches a user. An audit's claim crosses when the auditor publishes the finding. At each crossing, the fix financial auditing found is available. Record what crosses, against a declared standard, under a name.

And the record must pin the artifact it describes. Ristea and Mavroudis (2026) show that continuously updated AI systems shed their identifiers silently, so a finding attached to a model name rather than a pinned artifact attaches to nothing. When the developer revises the artifact daily, the interval a pin speaks for shrinks toward zero.

Accounting has the precedent. A ledger's balances change too fast to audit as states, so the record attaches upstream, to the journal entries that change them, and the statements are derived. Tamper-evident change logs, the instrument their own highest level asks for, are journal entries under another name. The receipt follows the same path when the artifact is a stream, so each decision that feeds the artifact becomes a recorded, signed crossing. For now the legible surface for those decisions is the evaluation, since a release gate cites its evals and an eval can carry a full record. How much farther upstream the record can reach is an open question, and the standard should declare where its records stop.

3 Evidence from zero-access audits

3.1 Eight audits from the public side

Audits have budgets, so the governing discipline is Peirce's economy of research (1879). Order the instruments by expected yield per dollar and spend on the highest-yield first. The economics of auditing reached the same prescription a century later. Verification is costly, so auditors buy it where the expected information per dollar is highest (Townsend 1979; Border and Sobel 1987; Mookherjee and Png 1989). Applied to AI audits, the ordering is stark, because the cheapest instrument is the only one with demonstrated yield on the public record.

Between May and July 2026 I audited seven benchmarks from the public side of the boundary, with no NDA, no lab relationship, and no access beyond what any stranger has. An eighth followed after I published a checklist distilled from the seven. Each one broke somewhere, and every severe defect surfaced before any model ran, except one exploit caught while running a benchmark as a paying stranger:

Benchmark	Finding
SWE-bench Pro	Proven floor of 15.0% of 728 tasks underdetermined; every label re-runs from a cold checkout
DeepSWE	4 of 113 answer keys fail their own verifiers; 1 of 113 after revision
Terminal-Bench	83 of 83 tasks pass after off-task destruction
τ-bench	Memorized answer passes 0% of regenerated task variants; leaked query passes 100%
SWE-bench Verified (as runner)	Test-edit exploit caught; 44 unrunnable tasks scored zero
ProgramBench	21 recall-only witnesses; 29 self-capturing goldens
MirrorCode	Headline claims autonomous builds; metric scores scoped reimplementation, 17 of 25 targets contaminated
FrontierCode (the eighth)	Public leaderboard cannot be re-derived from anything released

The checklist is ordered by cost, and its first five checks need nothing but reading. That checklist is also the auditor's second product. Verify the claim, then leave behind a check cheap enough for anyone else to run. Seven audits produced the findings; the checklist is why the eighth cost a fraction of the seven. The verdict covers one artifact; the check covers every later one.

A check specified this tightly also automates. Determinacy certifies by grep which tasks a hidden test grades against a spec that never stated the requirement, and it audited SWE-rebench, a benchmark it was not written for, down to a 14.5% pointer-checkable spine. An auditor's time is the cost of the first run, not of the hundredth.

3.2 No demonstrated yield from access

Against those eight receipts, internal access has no published demonstration of marginal yield. No audit on the public record shows added interior access producing a finding, with receipts, that boundary verification could not reach. This response does not claim access adds nothing. It observes that one mechanism is demonstrated and the other is a promissory note. The claim that access adds assurance is theirs to substantiate, and the challenge is executable. Publish one such finding.

Their own paper names the strongest candidates against this claim. It calls METR's review of Anthropic's sabotage risk report "among the first AAL-1 audits," and cites UK AISI and CAISI pre-deployment testing that "identified safety issues that developers then addressed before release." Each engagement is real, each had access no stranger has, and each found something. None shipped a receipt.

METR's red team of Anthropic's agent monitoring had "substantial access to relevant internal systems and information"; the resulting twenty-six-page report went to Anthropic, a redacted version to a subset of METR's own staff, and the attack trajectories to the vendor, so what reaches the public is that vulnerabilities existed and some were patched. The AISI evaluation publishes aggregate solve rates without the harness. The reciprocal Anthropic and OpenAI evaluations released two benchmarks and withheld the transcripts behind the comparative claims, disclaiming quantitative precision themselves.

METR's Frontier Risk Report is the closest case, because it does ship a machine-readable incident set. Twenty-four of its forty-four incidents come from public system cards and blog posts, two arrived by anonymous donation, and the eighteen drawn from METR's own privileged evaluations arrive as prose with no reproduction path. The portion a stranger can re-derive is the portion that never needed access.

Confidentiality is no shelter here, because their own framework concedes the machinery. Its reporting model sends unredacted findings to oversight bodies, and its reproducibility norm asks that another auditor with equivalent access be able to re-derive an audit. A finding's existence and the check that would refute it can be published even where its content cannot. Until then, AAL-1 is the cheapest level they define, at $300,000, and the free instruments appear nowhere in the framework.

Instrument	Cost (their estimates)	Findings with re-runnable receipts, on the public record
Boundary verification	An auditor's time	Severe defects in all 8 benchmarks examined
AAL-1 engagement	$300,000–$600,000	None identified
AAL-2 engagement	$1,000,000+	None identified
AAL-3/4 engagement	Several $M annually	"Not yet technically and organizationally feasible"

Buying access is a compensating control, the second-best substitute auditing accepts when the preferred control is missing. The primary control it stands in for is the boundary record.

3.3 The closest head-to-head

The closest thing to a head-to-head comparison ran in the wrong direction for internal access. On 8 July 2026, OpenAI audited SWE-bench Pro, estimated ~30% of tasks broken, and retracted its February recommendation of the benchmark. An unreleased pipeline flagged 27.4% of tasks; five unnamed engineers flagged 34.1%; the headline sits between. No per-task labels, no false-positive inspection, no released pipeline, so nothing they published re-derives the number. A number that cannot be re-derived is not a wrong measurement; it is not a measurement.

Same instrument, two audits, and the two numbers are compatible: a proven floor beneath a broader estimate. Accepting the ~30% requires trusting OpenAI; accepting the 15.0% requires only re-running the published checks. The warrant travels with the claim, or the claim travels on faith. OpenAI had every resource needed to ship receipts and shipped none; the zero-access audit is the one that survives deletion of its author.

Nor is the pattern one lab's. FrontierCode, a frontier coding eval from a second independent evals shop, ships a leaderboard that cannot be re-derived from anything released, graded in part by an unnamed model. Vishwarupe et al. (2026) call the shape an evidential inversion, the most consequential claims in AI safety being the least reproducible, and ask venues to make reproducibility a review requirement. Their demand fixes the norm and leaves its criterion open; the receipt closes it: replay by a stranger, from a cold start, to the same verdict.

The lesson is narrower than "access failed." Receipts remove dependence on the auditor's authority for the claims the receipt covers. They do not neutralize the auditor's selection of claims, timing, or publicity; a perfectly replayable cherry-picked audit is still a weaponized audit. A reader who doubts a receipt-backed audit re-runs it. A reader who doubts an authority-backed audit comes away empty-handed, and so does the auditor's own defense. When audits become competitive weapons between labs, and the timing of the OpenAI audit suggests they already have, this asymmetry is the difference between a dispute that resolves and a dispute that circulates.

4 Records, readers, signatures

Becker (1968) prices a deviation at the probability of being caught times the consequence that follows, and inspection games (Avenhaus, von Stengel, and Zamir 2002) make the catching strategic. Review is costly, so its probability is chosen rather than given. Splitting the caught-probability into review and catch-given-review names the components:

Deterrence factor	Component	Deployed precedent
Probability of being caught, given review	Records: replayable boundary provenance	Financial statements, PE calculations, claim records
Probability of review	Readers: a body that reads and attributes	Standard-setters, professional associations
Consequence	Signatures: a named person staked	Opinion letters, PE stamps, SOX certification

4.1 Records raise catchability

DeAngelo (1981) decomposes audit quality into the probability that a breach is discovered and the probability that a discovered breach is reported. A replayable boundary record drives the discovery term toward one for whatever the record covers; the reporting term is what signatures and independence are for.

Receipt cost and access cost also scale differently, which is what the AAL cost curve shows.

Cost grows with	Replayable receipt	Purchased access
Stakes	No. The receipt re-runs the same way whether the leaderboard number steers a hiring decision or a treaty	Yes. Vetting, secure facilities, and cleared personnel scale with what is at stake
Supply-chain length	No. A replayable check bypasses every link, re-run locally	Yes. Trust degrades across every link

What qualifies as a record? A record that always vindicates its keeper is a press release. Only a record that could convict is informative when it acquits.

4.2 Readers supply the probability of review

A record nobody reads deters nothing, and this is the current failure. SWE-bench Verified's contamination was common knowledge while it stayed the field's reported number for two years, records public, reading absent, consequence none. The market cannot correct what it cannot attribute.

Review probability is an institutional variable. Standard-setters and professional associations in accounting and engineering supply it: a body that reads records, remembers signatures, and makes findings legible across firms. That reader function is a better fit for AVERI than pursuing access levels its own paper marks infeasible. Reading is feasible today; the records already exist wherever a boundary is public.

Where the value of a measurement is diffuse, no single reader has an incentive to fund it. That is the standard public-goods failure (Samuelson 1954; Olson 1965), and Mengesha (2026) diagnoses the same structure behind frontier AI safety's underinvestment in coordination. A recognition body exists to concentrate that diffuse demand into directed reading. Concentration requires a funding mechanism, member demand, procurement leverage, or philanthropy; it does not happen by existing.

Two failure modes bound the institution's design. A body that stamps without reading is an issuer printing unbacked notes, and once issuing gets cheaper than checking, the stamps inflate the way any currency does. Lizzeri (1999) shows that an unconstrained certification intermediary profits most by revealing almost

nothing. And a body that reads everything submitted inherits a congestion problem, since submissions are cheap to generate and expensive to verify. Its intake therefore needs the discipline of a mature disclosure queue: triage on evidence rather than presentation, and provenance on its own dismissals.

Both bounds point to the same starting design, and land registration has run it for a century and a half. A registry adjudicates nothing at intake beyond form; it timestamps, indexes, and attributes what is deposited, so there is no stamp to inflate and no congestion at the door. And the archive it accumulates is the one thing later institutions cannot retroactively create.

The practice even comes in two grades, cheapest first: a deeds registry records instruments without vouching for their validity, and a title registry guarantees what it registers. A title is itself a chain of instruments back to a root grant, which is the claim record under a different sovereign. The body should open as a deeds registry for provenance claims and grow toward reading, and eventually title-grade vouching, as demand concentrates.

4.3 Signatures attach the consequence

A named signature gives the penalty an address, the way an engagement partner's name goes on an audit opinion and an engineer's stamp goes on a drawing. The verdict travels with the record and survives deletion of the signer; the name confers accountability, never authority.

When vouching was local, every voucher held a stake in the outcome, and that stake disciplined the vouch. Across the compartment walls of a modern supply chain the voucher is now a stranger who can profit from a false vouch and lose nothing. A named signature puts that party back at the boundary. Klein and Leffler (1981) state the condition exactly. Performance is self-enforcing when the rents lost by cheating exceed the one-time gain. Shapiro (1983) prices the premium such a reputation earns.

For now the signer stakes reputation rather than personal liability; that is a claim about sequence. Liability requires a forum, a plaintiff, and a causal chain from decision to harm, and all three are missing in AI. Dye (1993) shows that liability, auditing standards, and auditor wealth are jointly determined once a forum exists, so the forum comes first. Reputation requires none of them, and in a small, dense, reference-driven field it binds. Reputation erodes rather than revokes, so consequence depends on whether the community reads and remembers, which is the reader function again. And an industry sets its own reputations, while liability at least routes through a party outside it.

Accounting, engineering, and advertising all ran this sequence: voluntary practice and voluntary seals first, statute decades later. Engineering societies organized from the 1850s and the first licensure law arrived in 1907; financial audits predate the securities laws that made them mandatory; advertisers founded the Audit Bureau of Circulations in 1914, and broadcasters sold against notarized affidavits of performance. Receipts precede regulation because enforcement needs a record to act on, and regulation keeps building on the record long afterward. Seventy years after audits became mandatory, Sarbanes-Oxley added personal certification by named officers.

None of this physically constrains anyone. The law does not physically protect anybody; neither does a security camera, which points at the door rather than sweeping the interior, records a timestamped projection rather than the whole scene, and deters exactly those who expect the footage to be reviewed and expect to still be around when it is. The mechanism here is evidentiary in the same sense. It deters where evidence is likely to trigger consequences; it fails against actors with no reachable reputation and no continuation. Accounting, aviation, and engineering all rely on detection regimes shaped this way, and their attribution norm travels too. All three treat disclosed failure differently from concealed failure.

5 Omitted crossings and first-crossing harm

The boundary map can lie by omission. A company can comply immaculately at cheap boundaries, publishing exemplary benchmark and eval records, while the material crossings never appear on the declared map. Internal deployment is the live instance, models doing consequential work with no outside witness, and Charnock et al. (2026) already catalogue what developers should disclose about it. The three-party separation narrows this attack. When the map is induced by which party did the work, an omitted crossing leaves a signature, a reported result with no second party's record behind it. What survives the separation is work that never changes hands at all, which is harm at first crossing.

A second variant floods the record until effective review is uneconomic, complying in volume. Cheaper reading alone answers neither. But a used record carries the traces of execution that narration composed

after the fact does not: the failing test run, the reverted diff, the job log. And the standard can demand them.

The answer to omission is the standard: canonical record schemas, mandatory mapping from claim to evidence, materiality rules, declarations committed before outcomes are known, and adversarial challenge rights. That is what an accounting standard is, and metrology's standing conference of weights and measures has done the same job in public and on the record since its first meeting in 1889. Writing the equivalent for AI crossings, which evals must be disclosed, in what form, at what materiality threshold, is the standard-setting work that remains. It is smaller than it sounds. A declared projection reduces to a short, unglamorous schema whose fields the claim record has already enumerated. The disclosure costs producers little, since MirrorCode's per-task, per-model outcome grid was already a figure in its paper.

Whoever writes the standard can tilt it, and a check only the incumbent can afford is replayable in syntax and gatekept in practice, so mandated checks must stay cheap enough that a rival can run a competing one. Standards also freeze. Advertising's audit perimeter stalled at what its 1960s accreditation body could verify, delivery audited forever and outcomes on faith. So write the first standard expecting it to be the durable one.

The standard is also where interior access earns its place, in the same instrumental role the financial auditor's access has: sampling the interior to test whether the declared boundary map omits material crossings, with the opinion indexed to the declaration.

Harm at first crossing. The central case in frontier AI safety is the artifact whose first boundary crossing is the harm, and evidence gathered afterward arrives too late. Pre-crossing records require the auditor before the release, since after it the perturbable surface is whatever the lab chose to leave public. Insurers, regulators, and procurement all buy on anticipation rather than on history, so demand for records before the crossing already exists.

The developer's incentive already points that way, and it does not wait on a mandate. A defect caught before release is a fix inside the development loop; the same defect caught after release is a retraction, and OpenAI's withdrawal of its own February recommendation is what the late path costs. This is the same asymmetry the attribution norm runs on, applied one step earlier, before there is anything to conceal.

Cheap and private describes the arrangement segregation of duties exists to prevent, so the record requirement is the guard. A confidential pre-release disclosure is a crossing, and the record follows it. Developers get the early loop and keep the contents; the crossing still leaves a trace that a later reader can attribute. And the contents are not the whole of what the audit produces. An auditor who finds a defect privately can publish the check that found it, which is the part that generalizes to the next artifact.

A pre-release audit can still miss the thing, and for an artifact whose first crossing is the harm there is no second run. That is the hard limit of everything argued here and the strongest case for the access research program their paper outlines. The two conclusions are compatible: spend the deployed precedents where they demonstrably work, which is every layer whose records are public, and spend the access research where nothing else reaches.

6 Recommendations

Each recommendation names the actor able to execute it within existing authority, without waiting on anyone else. None asks the companies for access that does not already flow; each asks that a record follow it.

1. Record every crossing. Frontier AI companies should publish a claim record at every public crossing: each reported evaluation score shipped with its tasks, per-task outcomes, harness configuration, and grader identity, sufficient for a stranger to re-run it. The cost is low where the artifacts already exist and can lawfully be shared. Where publication itself would be unsafe, as with dangerous-capability details, an authorized independent party takes on the reproduction; the record's form does not change.
2. Name the signer. Frontier AI companies should name a signer on every published record and every release decision. The signature adds an address for consequence, though the verdict must still survive deletion of the signer.
3. Spend cheapest first. Auditors and evaluation organizations should spend audit budgets from the cheapest demonstrated instruments upward, running the boundary checks, five of which cost nothing beyond reading, before purchasing engagement-priced access.

4. Publish receipts. Auditors should publish receipts rather than headline rates. An audit whose number cannot be re-derived adds authority without adding evidence, and leaves even the auditor's defenders unable to answer a challenge.
5. Registry before reader, reader before access. AVERI and any prospective recognition body should build the registry function before the reader function, and the reader function before the access function: first record, timestamp, and attribute deposited provenance claims, vouching for nothing at intake; then read published records, remember signatures, and attribute findings across firms. Registration is feasible today at deeds-registry cost; reading follows as demand concentrates; the highest access levels are, by their authors' own account, not yet feasible at all. Timestamping is the part that cannot be deferred. A check is clean only against models whose training cutoff precedes its publication, so without a trustworthy date there is no way to separate a model that solved the task from one that read it, and the result means nothing either way. Registration is what settles that date, which makes it a precondition for evaluation validity rather than a matter of good order.
6. Write the disclosure standard. Standard-setters should write the disclosure standard for crossings: record schemas drawn from the claim record, materiality thresholds, declarations committed before outcomes are known, and checks kept cheap enough that a rival can run a competing one. Draft it expecting the first version to prove durable.
7. Aim access at omissions. The access research program should direct interior access toward omission detection, testing whether declared boundary maps omit material crossings and covering harm at first crossing. It should also substantiate its premise by publishing one access-dependent finding, with receipts, that boundary verification could not reach. Even one would move the public record off zero.

7 Conclusion

Every component is adoptable immediately by actors who need no one's permission. Nothing here is a terminal architecture, yet every stronger successor regime, including theirs, consumes exactly the records this practice starts preserving now.

The frontier companies have asked for the strong version themselves. OpenAI's chief executive told the United States Senate in 2023 that "regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models"; Anthropic has published the case for targeted regulation, urging governments to act within eighteen months.

Regulation is the one form of accountability that requires nothing of its requester today: no record, no signature, no reader. Each requester could instead publish, now, the records a future regulator would need, several of which its own papers already typeset. A mandate arriving before those records exist would have nothing to grip. The observation indicts the arrangement, not the sincerity of the request.

Step one is on every path. Timestamps do not decay while institutions deliberate, and yesterday's record cannot be manufactured tomorrow. The receipts are already accumulating. AVERI need only hold them.

8 Glossary

Auditing terms follow their established use in financial auditing and in Brundage et al.; epistemic terms follow Verifiable Knowledge and What Cannot Be False Cannot Be True.

AI Assurance Levels (AAL-1 to AAL-4). Brundage et al.'s four levels of audit rigor. AAL-1: multi-week engagements relying substantially on company representations, scoped to particular systems. AAL-2: deeper access to non-public information, company-level scope. AAL-3: multiyear engagement with white-box access and continuous monitoring. AAL-4: continuous verification designed to detect active deception, aspiring to "treaty-grade" confidence.

Reasonable assurance. Financial auditing's term of art for a higher degree of confidence than "limited" assurance. Their glossary maps it to AAL-2.

Compensating control. An alternative control accepted when the preferred control is unavailable, treated by auditors as second-best rather than equivalent.

Segregation of duties. The accounting principle that execution, custody, recording, and authorization must not rest with one party, because a party holding all four can misstate without any record catching it.

Materiality. The threshold above which an omission or misstatement would change a reasonable reader's conclusion. Completeness in auditing is always relative to a claim and a materiality threshold, never absolute.

Boundary, crossing. As used here: the line between an organization and anyone who relies on its claims. An artifact crosses when it reaches a user, a customer, a leaderboard, or the public record; a claim crosses when someone outside is expected to act on it. A crossing need not be public: a confidential disclosure to an auditor crosses, and the record requirement follows it. Under the three-party separation (developers build, evaluation organizations measure, auditors verify), a crossing exists wherever work changes hands between parties.

Claim. A statement presented together with the check that would refute it, so that anyone re-runs the check to the same verdict. Entitlement to a claim comes from the replay, never from the author's word or credential. A claim record carries what the replay needs: the provenance edges down to its roots, the check procedure, the kill condition, and the attesting signature.

Check, kill condition. The check is the claim's test run as a program; the kill condition is the observable outcome whose firing refutes the claim. A check whose test can never fail is a mocked pass, not a check.

True, false, untrue. True: the check ran and passed. False: the check ran and broke. Untrue: no check returned a verdict, because none was shipped or none completes. True and false are both halting states; untrue is the absence of a verdict, and a false outranks an untrue because a check that broke says where, while a claim with no check says nothing even when it happens to be right. Accountable failure outranks unaccountable assertion.

Receipt. The committed artifact that lets a stranger re-run a check from a cold start to the same verdict: the case data, the script, the pinned environment, the expected output.

Attestation. The signed check log: I ran this, here is the receipt. Attestation identifies who stands behind a record; it confers no entitlement. An attestation without a replayable check beneath it is authority, not evidence.

Underdetermined. Of a benchmark task: the public statement does not pin the behavior the hidden grader scores, so passing measures recovery of an unstated choice rather than solution of the stated problem.

Disclosures

Funding. This work received no funding. The author has no financial relationship with any organization whose benchmarks, products, or frameworks are assessed here, and no organization reviewed this response before publication.

Model use. This response was drafted with LLM assistance, under the author's authorship, with the author answerable for every claim in it. LLMs also served as instruments inside the audits cited above: as coding agents executing published protocols, and as blind independent coders in the closure measurement behind the FrontierCode audit. Each such use resolves to a committed transcript or receipt in the linked repositories, on the same replay standard this response asks of others.

References

Works cited by author and year. Institutional reports and the author's audit posts are linked where they appear.

- Alchian, A. & Demsetz, H. (1972). "Production, Information Costs, and Economic Organization." American Economic Review 62 (5).
- Avenhaus, R., von Stengel, B. & Zamir, S. (2002). "Inspection Games." Handbook of Game Theory with Economic Applications, vol. 3. Elsevier.
- Becker, G. S. (1968). "Crime and Punishment: An Economic Approach." Journal of Political Economy 76 (2).
- Border, K. C. & Sobel, J. (1987). "Samurai Accountant: A Theory of Auditing and Plunder." Review of Economic Studies 54 (4).
- Brundage, M., et al. (2026). "Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies." arXiv:2601.11699.
- Charnock, J., Mehta Moreno, R., Miller, J. & Anderson, W. L. (2026). "What Should Frontier AI Developers Disclose About Internal Deployments?" arXiv:2604.23065.

- DeAngelo, L. E. (1981). "Auditor Size and Audit Quality." Journal of Accounting and Economics 3 (3).
- Dye, R. A. (1993). "Auditing Standards, Legal Liability, and Auditor Wealth." Journal of Political Economy 101 (5).
- Jensen, M. C. & Meckling, W. H. (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." Journal of Financial Economics 3 (4).
- Klein, B. & Leffler, K. B. (1981). "The Role of Market Forces in Assuring Contractual Performance." Journal of Political Economy 89 (4).
- Lizzeri, A. (1999). "Information Revelation and Certification Intermediaries." RAND Journal of Economics 30 (2).
- Mengesha, I. (2026). "The Coordination Gap in Frontier AI Safety Policies." arXiv:2603.10015.
- Mookherjee, D. & Png, I. (1989). "Optimal Auditing, Insurance, and Redistribution." Quarterly Journal of Economics 104 (2).
- Olson, M. (1965). The Logic of Collective Action. Harvard University Press.
- Peirce, C. S. (1879). "Note on the Theory of the Economy of Research." Report of the Superintendent of the United States Coast Survey; reprinted in Operations Research 15 (4), 1967.
- Ristea, D. & Mavroudis, V. (2026). "Referential Security as a New Paradigm for AI Evaluations." arXiv:2605.25673.
- Samuelson, P. A. (1954). "The Pure Theory of Public Expenditure." Review of Economics and Statistics 36 (4).
- Shapiro, C. (1983). "Premiums for High Quality Products as Returns to Reputations." Quarterly Journal of Economics 98 (4).
- Townsend, R. M. (1979). "Optimal Contracts and Competitive Markets with Costly State Verification." Journal of Economic Theory 21 (2).
- Vishwarupe, V., Shadbolt, N., Jirotka, M. & Flechais, I. (2026). "NeurIPS Should Require Reproducibility Standards for Frontier AI Safety Claims." arXiv:2605.08192.