THE VERIFIER - REPORTS

REPORT NO. 01

The State of Verifiable AI

"You no longer have to trust AI - you can verify it."

CLAIM UNDER TEST

9 JULY 2026

FREE TO READ - NEVER SPONSORED

SHA-256 71a56793193502fc7ed435d160fa5d6b
0051ab83b54985a7559b0e0861465847

FINGERPRINT OF THE CANONICAL TEXT - RECIPE IN APPENDIX B

CONTENTS

PROLOGUE

How to read this report

This is a Verifier report, which means it obeys rules most industry reports do not. It tests one question, stated on the cover as a claim. Every material claim inside it appears in the ledger at Appendix A with a status - VERIFIED, TRACKING, CONTESTED or EARLY - and a source. The full text carries a SHA-256 fingerprint, printed on this page and recomputable by any reader from the published canonical text using the recipe in Appendix B; if a single character of the record changes, the fingerprint changes. A section near the end lists what would change our verdicts, because a conclusion that cannot name its own failure conditions is an opinion wearing a lab coat. Corrections, if any become necessary, will be logged on the public record at the report's page.

Two structural notes. First, this report is not sponsored. Verifier reports never are: our funding comes from fully disclosed sponsorships of our events and podcast, and none of it touches research. Second, a conflict disclosure. The Verifier's founder holds a commercial role at a company that builds agent-accountability infrastructure. To keep this report clean, that company and its products are excluded from coverage, comparison and citation throughout; the category it operates in is covered strictly at the level of open standards and named third parties. Readers deserve to know the conflict exists and how it was handled.

One reading instruction. Vendor performance claims in this field are usually true and usually incomplete: a benchmark that says "a thousand times faster" is a fact about a baseline as much as a fact about a product. Where we quote such numbers we name the baseline, and the ledger records who produced the measurement. Treat every multiplier in this report as carrying an invisible footnote that reads "against what, measured by whom" - and then check the ledger, where the footnote is visible.

THE FINGERPRINT

SHA-256 `71a56793193502fc7ed435d160fa5d6b0051ab83b54985a7559b0e0861465847`

Recompute it: take report-01-canonical.txt, decode entities, remove all whitespace, lower-case, hash. Full recipe in Appendix B.

PART ONE

The Question

CHAPTER 01

The claim under test

Somewhere in the last two years, a sentence migrated from cryptography conferences to enterprise sales decks: you no longer have to trust AI, you can verify it. This report tests that sentence against the record as of July 2026, layer by layer, at the only resolution that matters - what can actually be checked today, by whom, at what cost.

The short version is that the sentence is true in places, false in places, and sold in more places than it is true. Verification is real and in production where the thing being verified is small, recent or hardware-adjacent: a photograph's capture device, a modest model's inference, the identity of a crawler knocking on a website. It is aspirational where it is most needed: nothing today lets an outside party verify how a frontier model was trained, and no production system proves frontier-scale inference cryptographically at interactive speed. Between those poles sits a fast-moving middle - hardware attestation that verifies the environment rather than the computation, provenance standards racing their own adoption gaps, agent identity protocols moving through standards bodies at unusual speed.

THE SCOREBOARD - SIX LAYERS, ONE STANDARD

LAYER	WHAT IS VERIFIABLE TODAY	STATUS
Content provenance	That a signed device or tool produced a file, and what edits followed - where the credential survives	TRACKING
Inference - cryptographic	Full inference of small and mid-size models, with proofs verifiable in under a second	VERIFIED at that scale
Inference - frontier scale	Proof of a frontier-model inference at production speed and cost	CONTESTED
Inference - hardware	That an approved stack ran in a sealed environment, at a 2-5% tax	TRACKING
Training and data	How any given model was actually trained	EARLY - the open gap
Evaluations	That a quoted benchmark measured what it claims, uncontaminated	CONTESTED
Agents	Who an agent is - not yet what it did	EARLY

Three numbers frame everything that follows. The first is 10,000 - the rough multiple, and falling, by which proving a computation cryptographically still exceeds simply running it, down from estimates near one million a few years ago. The second is 2-5% - the throughput tax NVIDIA characterises for running transformer inference inside a hardware-sealed environment on current data-centre GPUs, an overhead heading toward zero on the newest silicon. The third is zero - the number of technical mechanisms by which a model developer can today prove to an outside party how a model was trained. The distance between those numbers is the state of verifiable AI: the cheap verification verifies the least, the deep verification costs the most, and the layer regulation cares most about has no verification at all.

> The cheap verification verifies the least. The deep verification costs the most. And the layer the law cares most about cannot yet be verified by anyone.

CHAPTER 02

The Verification Ladder

Claims about AI trustworthiness arrive in wildly different weight classes wearing the same vocabulary, so this report grades every mechanism on a five-rung ladder. The rungs are defined by one question: what must you still take on faith?

The Verification Ladder - what must you still take on faith at each rung.

Level 0, CLAIMED: a statement exists. A model card that says "trained on licensed data," a blog post that says "rigorously evaluated." Faith required: total. Level 1, DOCUMENTED: the statement comes with artefacts a reader can inspect - datasheets, eval tables, system cards. Faith required: that the documents describe the system that actually shipped, which nothing enforces. Level 2, AUDITED: an independent party examined the system and signed their name. Faith required: in the auditor's access, competence and incentives - and in the untestable premise that the audited artefact is the deployed artefact. Level 3, ATTESTED: hardware or infrastructure cryptographically vouches that a specific measured environment ran - a camera's secure element signing a capture, a GPU's security processor signing its firmware state, an agent signing its requests. Faith required: in the hardware vendor's root of trust and the absence of physical compromise. Level 4, PROVEN: a mathematical proof accompanies the output, checkable by anyone, resting on cryptographic assumptions alone. Faith required: in mathematics and in the correctness of the proving software - a smaller thing, though as this publication has reported elsewhere, not nothing.

The ladder does two jobs in this report. It sorts the marketing - "verifiable" gets used for everything from Level 1 upward, and the word does very different work at each rung. And it locates the frontier precisely: the story of 2024-26 is the migration of real systems from Levels 0-2, where AI has always lived, to Levels 3 and 4, which until recently were reserved for much smaller computations. Every chapter that follows states plainly which rung its subject actually occupies.

CHAPTER 03

Why 2026 forced the question

Three pressures converged to turn verification from a research aesthetic into a procurement line, and each carries a date.

The first is regulatory and lands in twenty-four days. On 2 August 2026, the EU AI Act's Article 50 transparency obligations become enforceable together with the Act's penalty regime and the fining powers over general-purpose models - the subject of this publication's standing countdown coverage. Article 50 requires, among other things, that AI-generated and manipulated content be disclosed in machine-readable form that is, in the regulation's language, effective, interoperable, robust and reliable. That sentence is quietly a technical specification: proprietary platform labels do not travel across jurisdictions and pipelines; cryptographically signed provenance does. The same date activates the high-risk obligations that vendors of verification tooling now cite in their own launch materials - a circularity worth noticing, and one reason this report grades demand-side claims as carefully as supply-side ones. California's SB 942 has been in effect since January. The direction of regulatory travel is uniform: disclosure must become machine-checkable, which is to say, the law is drafting verification infrastructure into existence.

The second pressure is economic and shows up in the gap between two adoption numbers this publication has examined at length: roughly nine in ten organisations now report using AI somewhere, while McKinsey's accounting finds about six percent capturing significant enterprise value. Part of that gap is ordinary technology-cycle physics. Part of it is a trust discount: systems whose behaviour cannot be verified get deployed in low-stakes corners, and low-stakes corners produce low-stakes value. The verification industry's own demand evidence points the same way - a vendor-commissioned reading of executive surveys has seventy-one percent of enterprise leaders saying they will not scale AI without proof of correctness. We ledger that number as what it is, a vendor-cited survey figure, but its direction matches what the deployment data shows: the bottleneck on AI value has migrated from capability to accountability.

The third pressure is the arrival of machines on both sides of the trust equation. Automated agents now generate more than ten billion weekly requests across one major network alone, and that network's chief executive predicts bot traffic will pass human traffic within a year - a prediction we log rather than endorse. Agents transact, browse, book and buy, and none of the web's legacy identity mechanisms - user-agent strings, IP lists, robots.txt - can survive contact with software that can trivially forge all three. Meanwhile the machinery of proof itself got fast: the base-layer proving milestones this publication tracks elsewhere took real-time verification of a major blockchain from sixteen minutes to sixteen seconds in a single year, and the same industrial improvement in proof systems is what makes proving neural-network inference thinkable at all. Demand met supply on a schedule neither controlled, and 2026 is the year they collided.

PART TWO

The Stack, Layer By Layer

CHAPTER 04

Content: provenance at capture

The most deployed verification layer in AI is not about AI models at all - it is about everything they threaten. Content provenance answers a narrower question than "is this real": it answers "what signed this file, and what has happened to it since." The open standard is C2PA, whose Content Credentials attach a cryptographically signed, tamper-evident manifest to a file - capture device or generating tool, edit history, AI-involvement declarations - verifiable offline by any conforming validator, because the certificates travel inside the manifest.

On the Verification Ladder this is Level 3, attestation, and 2025-26 is the period it became infrastructure. The hardware root of trust arrived in consumer hands: one flagship smartphone line now signs every photograph by default with hardware-backed keys and became the first phone certified under the standard's formal Conformance Program at its highest defined assurance level. Professional bodies from five camera makers sign at capture, with more announced through the autumn, including the largest phone platform of all. On the generation side, the major commercial image and video models now label their output by default, combining signed manifests with invisible watermarking - the layered approach one leading lab formalised in May, and which one search-and-browser company is rolling out as verification surfaces across its assistant, search results and browser. More than twenty billion images carry that company's invisible watermark by its own count. Newsrooms from the international wire services to the public broadcasters sign what they publish, and one press-workflow staple used across photojournalism confirmed in February that credential preservation is coming to the ingest desk itself.

The honest state of the layer, however, lives in three gaps this publication has reported before and re-verifies here. The first is claims versus conformance: membership of the adoption coalition passed several thousand organisations, but the formal Conforming Products List - the registry of products independently verified against the specification - remains far shorter, and the distance between "supports the standard" in a press release and "certified conformant" in the registry is where trust quietly leaks. The second is the last hop: most distribution intermediaries still strip metadata in ordinary transcoding, so a credential signed at capture frequently arrives at the viewer's screen dead - the reason the ecosystem's current work centres on durable credentials that pair the manifest with watermarking and fingerprinting so a stripped credential can be recovered. Platform behaviour is improving unevenly: one professional network preserves credential chains end to end; others read credentials, keep server-side records and show their own labels; messaging apps and email strip everything; the screenshot remains unsolved. The third gap is governance under stress. The trust layer was rebuilt in mid-2025 around a formal Trust List, with the interim list frozen on the first of January; certificates cost real money from a short list of authorities, there is no free issuer, and the standard's most instructive incident remains the camera maker that shipped signing, discovered a vulnerability, and revoked every certificate its cameras had issued - a revocation that invalidated every credential those bodies had produced and, as of this writing, has not been restored. That episode is the layer working as designed and hurting as designed, simultaneously.

The layer also carries a tension its advocates discuss less than its architecture: provenance is identity infrastructure, and identity cuts both ways. A signed capture can disclose the photographer's name, the device's serial and the exact coordinates of the shutter press - which is protective metadata in a newsroom and target-painting metadata in the hands of a source, an activist or a witness. The specification's own answer is instructive for this report's wider thesis: recent versions added redactable assertions and, notably, zero-knowledge identity proofs - the ability to demonstrate that a credential chains to a trusted signer without revealing which one. The provenance layer, in other words, is

importing machinery from the proof layer to solve the problem its own success created, and anonymous-but-verified capture workflows are the plausible 2026-27 deliverable. Institutional adoption meanwhile crossed from industry into state practice during the period: the US cybersecurity agency's advisory formally recommended content credentials for government and critical-infrastructure use, the Library of Congress opened a community of practice on provenance in archival workflows, and the wire services' editorial rules increasingly treat an unsigned image of a major news event as a rejection, not a review item. When archives, agencies and wires converge on the same primitive, a standard has stopped being a proposal.

Two boundary conditions deserve permanent ink. Provenance proves that a claim was made by a key, not that the world was as the caption says - a camera can be pointed at a staged scene, and no manifest will object. And the specification itself is a moving target whose versions are reported inconsistently even by careful secondary sources this year; the ledger records the discrepancy rather than adjudicating it, which is what a ledger is for.

> **VERDICT TRACKING** Provenance at capture is real, shipping and regulation-backed - and its chain still breaks at the last hop more often than it holds. The infrastructure exists; the pipeline discipline does not, yet. Watch the Conforming Products List, not the membership count.

CHAPTER 05

Inference, the cryptographic route

The maximal promise - Level 4, a mathematical proof that this model, on this input, produced this output - stopped being theoretical during the period this report covers. The field is zkML: compile a neural network into an arithmetic circuit, run inference, and emit a succinct proof that anyone can check in milliseconds without seeing the weights or the input. The idea predates the practicality by years; the overhead was the whole story. Proving a computation cryptographically cost on the order of a million times the computation itself when the field's current generation of builders started, an obstacle one practitioner memorably described as swimming through concrete.

The record of 2025-26 is a staircase of that overhead falling. A compiler released in March 2025 proved inference of a 138-million-parameter vision model in 2.2 seconds. In July 2025 came the result the field treats as its Hello World moment: the first production-oriented system to prove a full large-language-model inference end to end - a 124-million-parameter transformer, the smallest ancestor of the models that matter, but a transformer nonetheless, with the attention mechanisms, non-linear activations and autoregressive token loop that had made the architecture cryptographically hostile. The proving system behind that result reached public release in February 2026 with benchmarks of 54 to 158 times faster proof generation than the previous open-source baseline and several-hundred-fold faster verification, bringing proof checks under a second; by its June open-sourcing the project claimed more than twelve million inference proofs generated. Those multipliers are vendor-measured against a named baseline and we ledger them as such - but the baseline is the field's own previous best, the direction is corroborated by independent work reporting similar-order speedups, and the trajectory (a million times overhead, then a hundred thousand, then ten thousand) is attested across multiple teams. What "production" concretely means at this rung is worth spelling out, because the phrase gets borrowed by systems three sizes larger. The proofs being generated by the million are for compact models doing disputable work: risk scores that gate transactions, classifiers whose verdicts trigger money movement, match engines whose fairness someone might sue over, on-chain logic that cannot call a model without evidence the call was honest. The common shape is a small model, a consequential decision, and a

verifier who is structurally adversarial - conditions under which a thousands-fold compute premium on a millisecond inference is an invisible line item, and a sub-second, portable, anyone-can-check proof is the entire product. Silicon vendors noticed: the two largest names in AI hardware now run strategic collaborations with proving teams, which is the sincerest form of due diligence the hardware industry offers - the announced collaborations aim the proving stack at the vendors' own attestation and identity programmes, an early sign that the hardware rung and the proof rung expect to compose rather than compete.

The costs that remain are two, and they are honest ones. Circuits speak finite-field arithmetic while models speak floating point, so weights are quantised on the way in - a fidelity tax typically measured at half a percent to two percent of benchmark accuracy, meaning the proven model is a very close sibling of the deployed model rather than its identical twin, a nuance that matters exactly in the high-stakes settings where proofs matter. And scale remains the wall: what is proven in production today are perceptrons, convolutional networks and small transformers. A frontier model is three orders of magnitude larger than the largest publicly proven inference, and no team has demonstrated frontier-scale proof at anything resembling interactive latency or tolerable cost. The gap is closing at a rate that embarrasses two-year-old predictions, but as of July 2026 it is a gap, and vendors' own materials quietly agree - the production deployments they cite are compact models doing consequential work, not large ones doing conversational work.

> **VERDICT SPLIT** For small and mid-size models, cryptographic inference verification is VERIFIED - production systems exist, proofs check in under a second, and the overhead curve has fallen a hundred-fold in three years. At frontier scale it is CONTESTED - claimed as imminent, demonstrated by no one. This is the single most misquoted distinction in the field.

CHAPTER 06

Inference, the hardware route

While cryptography climbs toward large models from below, hardware attestation arrived from above and quietly took the enterprise market. The proposition is different in kind: not a proof that a computation was performed correctly, but cryptographic evidence that a specific, measured environment performed it - Level 3 on the ladder, attestation, doing pragmatic work that Level 4 cannot yet afford.

The mechanics matured fast. The current data-centre GPU generation shipped the first accelerator-native trusted execution environments: a dedicated security engine on the die, memory encrypted before it leaves the package, keys generated inside the security processor and never exported. Remote attestation closes the loop - the GPU produces a vendor-signed report of its firmware measurements and workload state, checkable before any sensitive data or weights are released to it, and composite attestation schemes now bind the CPU enclave and GPU evidence into one verdict, with a chip-industry trust authority verifying both sides on the newest server platforms. The operational numbers explain the adoption: the vendor characterises the throughput tax for transformer inference at two to five percent on current parts, attestation costs one to three seconds once at provisioning rather than per request, and on the newest architecture - the first with trusted I/O paths - matrix arithmetic inside the sealed mode benchmarks at effectively native speed, with the residual losses traced by recent engineering work not to the silicon but to a serialised bridge between the confidential VM and the GPU that better-designed loaders largely recover, in one published case cutting a two-hundred-and-eighty-seven-second confidential model load to eight.

What the buyer must still believe is the part the sales material compresses. The root of trust is the hardware vendor: the attestation is a signature by the manufacturer over measurements the manufacturer's silicon took of the manufacturer's firmware. That is a real and useful trust reduction - the cloud operator, the platform administrator and the co-tenant all exit the trust base - but it removes no landlord; it changes which one holds the keys. Side channels are the standing caveat; the vendor ships mitigations for known classes and the honest formulation from practitioners is that risk is reduced by orders of magnitude, not eliminated. And attestation's verdict is about the environment, not the computation: a sealed enclave will faithfully attest to running whatever approved stack it was given, including one with a bug in it. The distinction the market keeps eliding fits in four words: attested is not proven.

> **VERDICT TRACKING** Hardware attestation is the workhorse of verifiable AI in 2026 - near-zero cost, production-grade, regulation-friendly. It verifies where a computation ran and what software ran it, at the price of trusting the silicon vendor. As the cheap rung it is winning; as the top rung it is being oversold.

CHAPTER 07

The training gap

Descend to the layer that regulation, litigation and safety all care about most, and verification runs out entirely. There is today no technical mechanism by which a model developer can prove to an outside party how a model was trained - which data went in, what procedure ran, whether the artefact that was evaluated is the artefact that shipped. The clearest statement of the gap comes from the AI-safety research community itself, whose survey of the problem is blunt: a provider could present a false or partial account of training and there would be no way to know. This is the ladder's basement occupied by the stack's foundation.

The reasons are structural rather than negligent. Training runs are not reproducible in the cryptographic sense - parallelism, hardware non-determinism and sheer scale mean the same inputs do not yield bit-identical weights, so the elegant trick that works elsewhere (recompute and compare) has nothing to grip. Weights are opaque in a way source code is not: properties an auditor certifies on Tuesday cannot be mechanically bound to the checkpoint serving traffic on Thursday, and behaviours can be trained to appear only outside evaluation conditions, which makes the audit rung structurally weaker here than anywhere else on the ladder. What exists instead is a perimeter being built around the void, and 2026's version of it is worth crediting precisely. Software supply-chain attestation has been extended to model artefacts, so the provenance of a checkpoint file - who built it, from what declared inputs, signed how - can now travel with the file even though the truth of the declaration cannot be proven. Data-provenance practice has professionalised sharply under regulatory pressure: the field-level lineage audits now expected of training corpora, the academic auditing of dataset licensing at scale, and the transparency indices that score foundation-model developers have together turned Level 1 documentation from a courtesy into a compliance artefact, with the EU's general-purpose code of practice supplying the template. Research prototypes go further - hardware-attested training steps with on-ledger lineage, proof-of-learning schemes that commit to checkpoints during the run - and the honest description of all of them is promising, partial and unadopted by any frontier lab.

The consequence deserves stating without cushioning, because every downstream verification inherits it. A perfectly proven inference of a model is a proof about a set of weights whose origins remain a matter of testimony. The stack's deepest verification currently rests on its least verifiable layer, and

every roadmap in this report should be read with that inversion in mind.

> VERDICT EARLY
>
> Training verification does not exist; training documentation is professionalising fast under law. The gap between those sentences is the largest open problem in verifiable AI, and no one credible claims otherwise. What would move the verdict: a frontier-scale proof-of-training demonstration, or hardware-attested training adopted for a flagship model.

CHAPTER 08

Evaluations: who grades the graders

Between the model and the world sits the layer where "verified" is claimed most often per unit of actual verification: evaluation. A benchmark score is a measurement, and measurements are exactly as trustworthy as their conditions - which, in 2026, are contested on three fronts this publication tracks as a standing beat.

The first is contamination. Public benchmarks leak into training corpora by the ordinary physics of web-scale scraping, after which a score measures recall as much as capability; the field's response - held-out private sets, periodically refreshed - improves the measurement at the price of making it unreproducible by outsiders, a trade that swaps one verification problem for another. The second is the judging loop. As frontier evaluation outgrew human graders, models began grading models, and this publication's own reporting this month covered a controlled experiment in which the graders' verdicts diverged systematically from ground truth in ways that favoured their own lineages - self-preference, measured. An evaluation stack in which the instrument is drawn from the population being measured has a calibration problem that no amount of scale dissolves. The third front is incentive: the parties producing the most-quoted numbers - labs marketing models, vendors marketing platforms, consultancies marketing transformation - are rewarded for the largest defensible figure, and the reader's only structural defence is the one this desk's benchmark manual prescribes: check the task, the denominator, and who paid.

What verification would look like here is beginning to take shape at the edges. Attestation objects - verifiable records of who evaluated what, under which protocol, with which data - are the framework the research community now argues should span the AI supply chain, and evaluation is the plausible entry point precisely because the field has momentum and money. Cryptographic tools have a niche opening too: proofs that a specific reported score resulted from running a specific committed model on a specific committed test set would convert "trust our eval" into "check our eval," and the machinery from the inference chapter is, at small model scale, already capable of it. Nothing production-grade does this today. The gap between the volume of numbers quoted and the volume of numbers checkable remains the widest in the stack relative to how much decision-weight the numbers carry.

> VERDICT CONTESTED
>
> Evaluation is where AI's biggest claims meet its weakest verification. Scores are Level 1 artefacts routinely cited with Level 4 confidence. The fix - attested, reproducible, third-party evaluation with committed models and data - exists as architecture, not yet as practice.

CHAPTER 09

Agents: identity before accountability

The newest layer is the one moving fastest, because it has to. Software agents now act - browse, book, purchase, negotiate - at a scale measured in tens of billions of requests weekly across a single network, against a web whose identity primitives were designed for a politer century. User-agent strings are trivially forged, IP allowlists rot, robots.txt is a request. In 2026 the response arrived with unusual institutional speed, and it is worth recording precisely because its precision is the point.

The mechanism is cryptographic request signing. An agent operator generates a keypair, publishes the public key in a directory at a well-known path on a domain it controls, and signs every outbound request using the HTTP message-signature standard; any receiving site verifies the signature and knows, at Level 3 assurance, which operator's agent is knocking. The draft architecture - authored jointly by engineers at the network company that proposed it and the search company that joined it - reached its fifth revision on 2 March 2026, and the internet's standards body has chartered a working group with published milestones: the core specifications to the steering group by April, an operational best-practices document by August. Deployment ran ahead of ratification. The proposing network activated verification at its edge in March and by June had launched a formal verified-agent category with nineteen named agents at debut, spanning the flagship assistant browsers; the largest cloud's WAF, major CDNs and commerce platforms shipped support; a public registry format for agent key discovery, co-announced with that cloud's agent platform in February, defines a signature card carrying operator identity, expected behaviour and keys; and the two global card networks adopted the protocol as the authentication foundation of their agent-payment programmes. In parallel, the decentralised-identity world contributed the complementary layer: an agent-identity specification for the dominant agent-tooling protocol was donated to the standards foundation in March, grafting verifiable-credential machinery - capabilities, authorisations, provenance - onto agents themselves.

The stack is layering rather than converging, and the layers are worth naming because they will be conflated in every pitch deck this year. Request-level identity - the signing protocol above - answers "who is at the door" during discovery and browsing. Payment-time verification - the card networks' agent programmes that adopted the same signatures as their foundation - answers "is this agent good for the money" at checkout. Agent-to-agent message formats answer how two pieces of software negotiate, and the credential layer donated to the identity foundation answers what an agent is authorised to claim about its capabilities and its principal. A single agent identity flowing through all four is the obvious operational pattern, and the registry format's signature card - operator, expected behaviour, keys, in one JSON object discoverable like a robots file - is the first artefact designed to be that identity's passport. None of it yet touches the hard layer beneath: the mandate. Whether the human actually authorised this purchase, at this price, from this merchant, is a question the current stack answers with a login cookie, which is to say, at Level 0.

The precision that matters: all of this verifies who, none of it verifies what. A signature attests that a request came from a named operator's agent; it says nothing about whether the agent was authorised by its principal to make it, and nothing about whether what the agent did matches what it claims it did. Identity, authorisation and action-verification are three different problems, and 2026 has industrialised exactly one of them - the first, easiest and most commercially urgent. The remaining two are where the training-gap and inference chapters of this report re-enter: verifying an agent's actions is verifying computation, with everything that entails. One more entry for the record: at least one secondary source this quarter reported the signing standard as "finalised in May" by the wrong standards body; the primary record shows an internet-standards working group with milestones still ahead as this report went to press. The ledger holds the discrepancy.

VERDICT
EARLY

Agent identity went from proposal to production inside eighteen months - the fastest standards adoption in this report. Agent accountability has not begun: knowing who an agent is precedes, and does not approach, knowing what it did. The gap between the two is where the next report in this category will be written.

PART THREE

The Economics

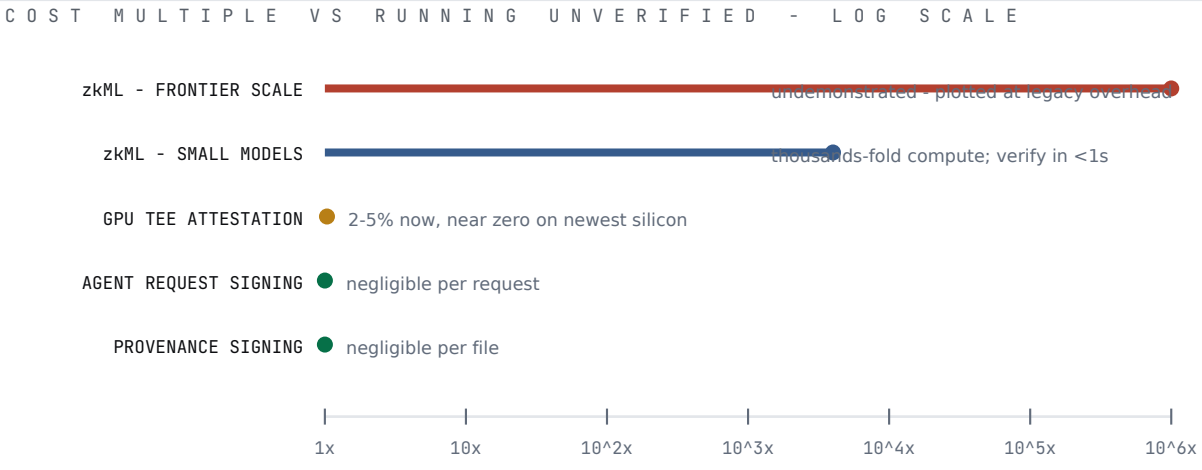
CHAPTER 10

The price of proof

Verification is an engineering economy, and the market's confusion about it usually reduces to a category error: comparing techniques as if they buy the same thing. They do not. The table below is this report's attempt at the comparison the sales decks skip - what each mechanism verifies, what it costs, and what you must still take on faith.

WHAT A UNIT OF TRUST COSTS - JULY 2026

MECHANISM	VERIFIES	OVERHEAD	RESIDUAL TRUST	MATURITY
Signed provenance (C2PA)	Which key produced or edited a file	Negligible per file	Signer's key custody; the pipeline preserving the manifest	Shipping; last-hop fragile
Request signing (agents)	Which operator's agent sent a request	Negligible per request	Operator's key custody; registry governance	Shipping at the edge
GPU TEE attestation	That a measured stack ran sealed	2-5% throughput now; near zero on newest silicon; 1-3s once	Silicon vendor's root; side-channel margin	Production, enterprise
zkML proof (small models)	The computation itself, input to output	Thousands-fold compute vs raw inference; sub-second to verify	Proving-software correctness; quantisation fidelity	Production, narrow
zkML proof (frontier models)	The computation itself, at the scale that matters	Undemonstrated	Not applicable	Research
Training verification	How the model came to be	Not applicable	Everything	Does not exist



Orders of magnitude - the verification tax by mechanism, log scale. The two production rungs differ by three orders of magnitude in cost and in what they prove.

Read as an economy, the stack resolves the way real deployments in 2026 have in fact resolved. Attestation wins wherever verification must ride along with production traffic, because a low-single-digit tax on inference is a rounding error against GPU pricing, and what enterprises are buying - assurance that the operator, the cloud and the neighbours cannot see or substitute the workload - is exactly what attestation sells. Cryptographic proof wins where the proof must outlive and out-travel its context: where

the verifier is an adversary, a regulator, a court or a blockchain, none of whom will accept a silicon vendor's signature as final, and where the computation is small enough to afford it - which is why its production niches today are compact models making consequential, disputable decisions. Provenance wins by default at the content layer because its per-file cost is effectively zero and regulation is about to make its absence expensive. And the pattern that emerges is not a race with a winner but a settling into rungs: cheap attestation everywhere, expensive proof where portability of trust justifies it, and a research frontier grinding the price of the top rung down by an order of magnitude every couple of years. The strategic question for any buyer reduces to one sentence: who has to believe you, and how much is their belief worth? When the believer is yourself or your regulator, attest. When the believer is your adversary, prove - if you can afford to, at the size you need, which for the largest models nobody yet can.

> Who has to believe you, and what is their belief worth? Everything else in verification economics is arithmetic on that sentence.

CHAPTER 11

The verifier’s toolkit

A report about verification that cannot itself be checked would be a genre error, so this chapter is the practical inversion of everything above: what a reader with an ordinary laptop can verify today, layer by layer, in minutes. Content: take any image from a signing camera, phone or AI generator and inspect it at the public credential viewer or with the open-source inspection tool; the manifest, the edit chain and any AI-involvement flags render in seconds, and a stripped credential renders as exactly the absence this report describes - run the same file through a messaging app first and watch the last-hop problem happen on your own screen. Agents: the signing protocol's key directories are public by design at a well-known path on each operator's domain, and a site owner behind a supporting edge can watch verified-agent classifications arrive in ordinary analytics; the difference between a signed assistant and a spoofed user-agent string is now a dashboard column. Inference: the open-source proving systems publish verifier binaries precisely so that checking does not require believing - a published proof of a model's inference verifies on commodity hardware in under a second, which the reader is invited to time. Attestation: the GPU vendor's attestation verification is a service call whose inputs and outputs are documented; what the reader can check without a data-centre badge is the shape of the evidence - what is measured, who signs, what a relying party actually receives - which is enough to test any vendor's "verified" claim against the ladder in Chapter 2. And this document: the fingerprint on the integrity page recomputes from the published canonical text with one line of standard tooling, by the recipe in Appendix B. A reader who performs even two of these checks will have done more verification than most of the market commentary this report competes with - which is, in one sentence, the state of the field.

PART FOUR

The Scoreboard

CHAPTER 12

Two years on the record

Verifiable AI's claims arrive breathless; its record, assembled with dates attached, is more interesting than the breathlessness. The ledger of load-bearing events, 2023 through this week:

THE RECORD, DATED

DATE	EVENT	WHY IT CARRIES WEIGHT
Oct 2023	First consumer camera ships with in-body credential signing	Provenance leaves the specification and enters hardware
2024	First data-centre GPUs ship accelerator-native trusted execution	Attestation reaches the silicon AI actually runs on
Mar 2025	Compiler proves a 138M-parameter vision model in 2.2 seconds	zkML crosses from demos to engineering
Mid 2025	Provenance standard launches formal conformance testing and rebuilt trust list	Claims and certification formally separate
Jul 2025	First full LLM inference cryptographically proven end to end	The transformer architecture falls to proof systems
Aug-Sep 2025	Camera-maker revokes all issued credentials after signing flaw; flagship phone ships default signing with hardware keys	The trust layer shows both failure handling and consumer scale in one season
Jan 2026	Interim provenance trust list frozen; state AI-disclosure law takes effect in California	Governance and regulation both harden
Feb 2026	Proving system released with 54-158x generation speedup over prior open baseline; agent key-registry format co-announced with major cloud	Both proof and identity layers publish their production interfaces
Mar 2026	Agent-signing architecture reaches fifth draft; standards working group chartered; edge verification activated in production; agent-identity spec donated to identity foundation	Agent identity moves through standards machinery at unusual speed
May-Jun 2026	Layered provenance-plus-watermark verification announced across major AI platforms and browsers; verified-agent category launches with nineteen named agents; proving system open-sourced citing twelve million proofs	The verification stack becomes a consumer-visible product
2 Aug 2026	EU AI Act transparency and penalty regime becomes enforceable	The deadline the entire supply side has been building toward

CHAPTER 13

Twelve months of checkable markers

A report that grades others' claims owes the reader claims of its own that can be graded. The following are the markers this publication will score, in public, on the dates given - each phrased so that success and failure are both recognisable.

By 2 August 2026: Article 50 enforcement begins. Marker: whether machine-readable disclosure in the field is dominated by the open provenance standard or fragments into per-platform labels - measurable by what the major platforms actually emit. By August 2026: the agent-signing working group's operational best-practices milestone falls due. Marker: delivered on schedule or slipped, and how many of the launch-cohort verified agents remain signing in production. By autumn 2026: the largest phone platform's announced credential support ships or slips - the single largest possible expansion of signed capture, checkable on release day. Through Q4 2026: the provenance Conforming Products List either grows past a hundred certified products or the claims-to-conformance gap persists as this report found it; we will publish the count. Within twelve months: a cryptographic proof of an inference from a model at or above ten billion parameters, from any team, at any latency - the demonstration that would move the frontier-scale verdict from CONTESTED, or its continued absence, which would confirm it. Within twelve months: the restoration, or continued suspension, of the revoked camera-maker's signing service - the ecosystem's live test of whether trust-layer failures heal. And within twelve months: any frontier lab adopting hardware-attested training or checkpoint commitment for a flagship model - the first plausible crack in the training gap. Where a marker lands, the next edition of this report says so, next to what we predicted, in the same table.

CHAPTER 14

What would change our verdict

Falsifiability is the house discipline, so: the overall TRACKING verdict on the cover claim would move to VERIFIED if, within the coming cycle, frontier-scale inference proof were demonstrated at production-tolerable cost, or hardware-attested training shipped on a flagship model - either would extend checkable verification to a layer where testimony currently rules. It would move toward CONTESTED if the provenance last-hop problem persists at scale after Article 50 enforcement - signalling that legal mandate plus shipped infrastructure still cannot hold a chain of custody across the real internet - or if the agent-identity cohort fragments into incompatible proprietary registries, or if a structural break is demonstrated in the deployed attestation roots or proving systems this stack now leans on. Our per-layer verdicts carry their own conditions in their own chapters. Readers who believe we have weighted the evidence wrongly are invited to say so on the record; the corrections log is part of the report.

CHAPTER 15

The verdicts

The claim under test was: you no longer have to trust AI - you can verify it. The finding, layer by layer: content provenance TRACKING, real and regulation-backed, fragile at the last hop; cryptographic inference verification VERIFIED at small scale and CONTESTED at the frontier - the field's most misquoted distinction; hardware attestation TRACKING, the economy's workhorse, verifying the environment and sold as more; training EARLY, the open gap beneath everything; evaluations CONTESTED, maximal decision-weight on minimal verification; agent identity EARLY and accelerating, with agent accountability not yet begun.

The composite verdict is TRACKING, and the sentence-length version of this report is this: in 2026 you can verify where an AI ran, who signed its outputs, and - for small models - exactly what it computed; you cannot yet verify what any frontier model is or how it came to be, and every honest roadmap in the field is a plan for closing exactly that distance. The claim on the cover is no longer false. It is not yet true. It is, for the first time, checkable - and this report has tried to be the checklist.

APPENDIX A

The claims ledger

Every material claim in this report, with the status The Verifier assigns it and the source that carries it. Statuses: VERIFIED (checked against the primary record), TRACKING (credible, single-sourced or vendor-reported, monitored), CONTESTED (sources genuinely conflict, or the claim is an absence we assert), EARLY (a category-state judgement).

VERIFIED	EU AI Act Article 50 transparency obligations, the penalty regime and GPAI fining powers become enforceable on 2 August 2026; the obligations require machine-readable marking of AI-generated content	Regulation (EU) 2024/1689, Arts. 50, 99, 113; The Verifier, "Twenty-Five Days to August" (Jul 2026)
VERIFIED	California SB 942 (AI content disclosure) took effect January 2026	California SB 942 (2024), operative date; compliance coverage (Mar 2026)
VERIFIED	A production zkML system proved full inference of GPT-2 (124M parameters) in July 2025 - the first end-to-end LLM inference proof	Lagrange Labs, DeepProve-1 announcement (Jul-Aug 2025)
VERIFIED	The February 2026 DeepProve release benchmarked 54-158x faster proof generation and several-hundred-fold faster verification than EZKL, the prior open-source baseline; figures are vendor-measured against that named baseline	Lagrange Labs release materials (Feb 2026); independent commentary (Apr 2026)
TRACKING	The same vendor's June 2026 open-sourcing cited more than twelve million inference proofs generated - a vendor-reported cumulative figure we have not independently confirmed	Lagrange open-source announcement via press distribution (Jun 2026)
VERIFIED	zkPyTorch (March 2025) proved VGG-16 (138M-parameter) inference in 2.2 seconds	zkPyTorch release documentation; ICME Labs zkML review (2025)
VERIFIED	zkML quantisation typically costs 0.5-2% benchmark accuracy, because circuits require fixed-point arithmetic	zkML technical literature and practitioner documentation (2025-26)
TRACKING	Field practitioners characterise historical zkVM/zkML overhead as ~1,000,000x, improving through ~100,000x toward ~10,000x - an order-of-magnitude trajectory attested across teams rather than a single measured constant	ICME Labs, "The Definitive Guide to zkML" (2025); corroborating practitioner accounts (2025-26)
CONTESTED	No cryptographic proof of a frontier-scale model inference at production latency and cost has been demonstrated by any team as of July 2026; several teams describe it as imminent	Survey of public zkML results through Jun 2026 - absence of demonstration is the finding
VERIFIED	NVIDIA characterises confidential-computing throughput overhead for transformer inference at 2-5% on Hopper-generation GPUs; attestation costs 1-3 seconds once at provisioning	NVIDIA CC benchmark characterisation; deployment documentation (2025-26)
VERIFIED	Blackwell-generation parts introduce trusted I/O; published engineering measurements show matrix workloads under confidential mode at effectively native speed, with residual serving losses traced to a serialised CVM-GPU bridge (one documented model-load falling 287s to 8.4s with a CC-aware loader)	Phala Network engineering publication, "The Serialized Bridge" (2026); NVIDIA Blackwell materials
VERIFIED	Composite CPU+GPU remote attestation using Intel TDX with NVIDIA Confidential Computing, verified through Intel Trust Authority, is described as production infrastructure on current HGX-class platforms	SEC filing describing HGX B200 confidential-computing deployment (2026)
VERIFIED	TEE attestation verifies the measured environment, not the correctness of the computation; the root of trust is the silicon vendor, and side-channel risk is mitigated rather than eliminated	Vendor security documentation; practitioner security analyses (2025-26)

VERIFIED	No technical mechanism exists by which a model developer can prove to an external party how a model was trained; training runs are not bit-reproducible and audited artefacts cannot be mechanically bound to deployed ones	Future of Life Institute, "Verifiable Training of AI Models" (2025); standing research literature
VERIFIED	Software supply-chain attestation (SLSA, Sigstore) has been extended to AI model artefacts, carrying signed provenance of checkpoints without proving training claims	SLSA/Sigstore AI-artefact documentation; supply-chain attestation literature (2025-26)
VERIFIED	The Pixel 10 (September 2025) signs every photo by default with hardware-backed keys and was the first smartphone certified under the C2PA Conformance Program at Assurance Level 2	Google security publication (Sep 2025); C2PA Conformance Program records
VERIFIED	Five camera makers ship in-body C2PA signing on current models, with Apple support announced for autumn 2026	Adoption trackers and vendor announcements through Apr 2026
VERIFIED	Nikon suspended its C2PA service after a signing vulnerability and revoked all issued certificates, invalidating previously produced credentials; service unrestored as of early 2026	Vendor notices; provenance-ecosystem reporting (2025-26); The Verifier provenance coverage
VERIFIED	The C2PA Interim Trust List was frozen on 1 January 2026 in favour of the formal Trust List; conformance certification remains far rarer than claimed support	C2PA Trust List and Conformance Program governance documents (2025-26)
CONTESTED	C2PA specification versioning is reported inconsistently across careful secondary sources this year (2.2/2.3/2.4 as "current," with differing dates); we record the discrepancy rather than adjudicate it	Comparative review of provenance trackers and guides (Feb-Jun 2026)
TRACKING	Google reports more than twenty billion images watermarked with SynthID; TikTok reports more than 1.3 billion videos labelled with AI provenance data - platform-reported figures	Google and platform disclosures (2025-26)
VERIFIED	OpenAI announced a layered C2PA-plus-watermark verification approach in May 2026; Google is rolling out SynthID and C2PA verification surfaces across Gemini, Search and Chrome	Vendor announcements (May-Jun 2026)
VERIFIED	The Web Bot Auth architecture draft (HTTP Message Signatures, Ed25519, well-known key directories) reached revision -05 on 2 March 2026; an IETF working group is chartered with milestones of April 2026 (core specs to IESG) and August 2026 (operational BCP)	draft-meunier-web-bot-auth-architecture-05; IETF WebBotAuth WG charter and milestones
VERIFIED	Cloudflare activated edge verification in March 2026 and launched a Verified AI Agent category in June 2026 with nineteen agents at debut; AWS WAF, major CDNs and commerce platforms ship support; Visa and Mastercard adopted the protocol as the authentication foundation of their agent-payment programmes	Cloudflare platform updates (Mar-Jun 2026); ecosystem adoption reporting
CONTESTED	At least one secondary source reported the agent-signing standard "finalised in May 2026" by the W3C; the primary record shows IETF standardisation in progress with milestones ahead	Comparison of secondary reporting against IETF datatracker record (May-Jun 2026)
VERIFIED	The MCP-I agent-identity specification was donated to the Decentralized Identity Foundation in March 2026	DIF announcement; specification record (Mar 2026)
TRACKING	AI bot requests exceed ten billion weekly on Cloudflare's network; its CEO predicts bot traffic will exceed human traffic by 2027 - operator-reported figure and a prediction, logged not endorsed	Cloudflare statements (Mar 2026)
VERIFIED	Roughly 88% of organisations report using AI in at least one function while McKinsey finds about 6% capturing significant enterprise value	McKinsey State of AI research; The Verifier, "The Agent Adoption Number Is Doing Three Jobs" (Jul 2026)
TRACKING	A vendor-cited reading of executive surveys reports 71% of enterprise leaders saying they will not scale AI without proof of correctness - vendor-quoted, direction consistent with independent adoption data	Vendor release citing McKinsey-attributed survey figure (Jun 2026)

VERIFIED	A controlled experiment covered by this publication found LLM judges exhibiting measurable self-preference when grading models including their own lineage	The Verifier, "Who Grades the Graders" (Jul 2026), and the underlying study
VERIFIED	Base-layer real-time proving improved from roughly sixteen minutes to sixteen seconds across the year to mid-2026	EthProofs and Ethereum Foundation milestones; The Verifier proving coverage (2026)
TRACKING	Deloitte projected synthetic content could account for up to 90% of online media by 2026 - a projection, logged as such	Deloitte TMT Predictions (2025)

APPENDIX B

Method and fingerprint

The fingerprint on the integrity page is computed as follows, using the same recipe The Verifier publishes for its articles. Take the canonical text file published alongside this report (report-01-canonical.txt). Decode any character entities, remove every whitespace character of any kind, convert the result to lower case, and compute the SHA-256 digest of the UTF-8 bytes. The hexadecimal digest must match the fingerprint printed in this document exactly. The canonical text is the report's manuscript of record - front matter through Appendix C, structural markers included - published alongside the PDF; the PDF is a typeset presentation of it. Any correction to the report produces a new canonical text, a new fingerprint, and a dated entry in the corrections log on the report's page. One line of scope: the fingerprint proves the text you hold is the text we published; it does not prove the text is true. The ledger, the sources and the markers are for that - and so, in time, is the scoreboard we have promised to publish against our own predictions.

APPENDIX C

Sources

Primary and regulatory: Regulation (EU) 2024/1689 (the AI Act), Articles 50, 99 and 113 and the GPAI code of practice; California SB 942; CISA advisory on multimedia integrity (Jan 2025); IETF draft-meunier-web-bot-auth-architecture-05 and the WebBotAuth working group charter and milestones; RFC 9421 (HTTP Message Signatures); C2PA specification family, Conformance Program and Trust List governance; SEC filing describing confidential-computing deployment on HGX B200 (2026). Vendors and standards bodies: Lagrange Labs (DeepProve-1, Feb 2026 release, Jun 2026 open-sourcing); zkPyTorch release documentation; EZKL project documentation; NVIDIA confidential-computing documentation and Blackwell materials; Intel Trust Authority documentation; Google security blog (Pixel 10 C2PA) and SynthID disclosures; OpenAI provenance announcements (May 2026); Cloudflare platform updates (Feb-Jun 2026) and Web Bot Auth documentation; Amazon Bedrock AgentCore registry announcement; Decentralized Identity Foundation (MCP-I, Mar 2026); Sigstore and SLSA AI-artefact documentation. Research and civil society: Future of Life Institute, "Verifiable Training of AI Models"; the Data Provenance Initiative (Longpre et al.); the Foundation Model Transparency Index (Stanford CRFM); ZKP-based verifiable machine learning survey literature (2025); attestation-objects analysis (Vincent, Apr 2026); Phala Network, "The Serialized Bridge" (2026); ICME Labs, "The Definitive Guide to zkML" (2025). Adoption tracking and secondary: provenance adoption trackers and conformance analyses (Feb-Jun 2026); C2PA camera-support guides; agent-verification ecosystem reporting (Apr-Jun 2026); McKinsey State of AI; Deloitte TMT Predictions (2025). The Verifier's own reporting drawn on and cross-referenced throughout: "Twenty-Five Days to August," "Who Grades the Graders," "The Agent Adoption Number Is Doing Three Jobs," "Content Credentials and the Last Hop," "The Year Proving Got Fast," and the desk manuals for benchmarks, demos and filings.