The Verification Gap

An Independent Analysis of the Distance Between Reported and Verified AI Performance

MarginSignal OS · Independent Findings Report · 2026 · v1.0 *Audience: risk-bearing operators (insurance, healthcare RCM, BPO, financial services) and the executives accountable for AI outcomes. Every citation independently verified against primary sources before publication (see Methodology). Living document — figures are point-in-time and revisited.*

Executive summary

Enterprises are deploying autonomous AI into high-stakes operations faster than anyone is checking whether it works. The numbers that justify those deployments — accuracy rates, resolution rates, benchmark scores — are, almost without exception, produced by the same party that builds or sells the system. That is not evidence. It is self-attestation with a dashboard.

This report documents the resulting **verification gap**: the measurable distance between what AI is *reported* to do and what it independently *verifies* as doing. We draw only on documented, primary-sourced events. Five findings:

1. **The people paid to verify are failing at verification.** A Big Four firm shipped a government report built on AI-fabricated citations; the same firms' own audits carry record deficiency rates.
2. **You are legally liable for your AI's output — the "the model did it" defense has already failed in court.**
3. **The best-looking dashboards hide the worst outcomes.** A marquee AI-customer-service success story reversed after quality quietly collapsed on the interactions that mattered.
4. **AI is confidently wrong more often than a year ago — and the confidence never wavers.** Independent audit puts leading chatbots' false-claim rate near one in three, and rising.
5. **The scoreboard itself is unreliable.** The benchmarks the industry cites rest on documented methodological flaws.

The through-line is structural, not moral: **the entity that builds or runs an AI system cannot be the one that certifies it worked.** Independence is not a feature to add later. It is the missing layer.

1. What the verification gap is

Every deployed AI system emits a number: "resolved 80% of tickets," "94% accurate," "state-of-the-art." That number is generated by the system being measured, or by the vendor selling it, using a method the buyer never sees. It answers the question *"does it look like it worked?"* — not *"did it work?"*

The gap between those two questions is where operational risk lives. It is quiet by design: a dashboard reports success the moment an interaction closes, not when the customer's problem is actually solved; a model reports the same fluent confidence whether it is right or hallucinating. Nothing in the system is built to surface its own failure. The gap only becomes visible when someone independent looks — a regulator, a court, a journalist, a researcher — usually after the cost has already been paid.

2. The evidence

The verifiers can't verify. In 2025, Deloitte's Australian firm delivered a ~A$290,000 report to the Department of Employment and Workplace Relations that contained citations to academic papers that do not exist and a fabricated quote attributed to a federal judge; the firm agreed to repay part of the fee. A separate Deloitte report for a Canadian provincial government was later found to contain similar AI-fabricated research. These are not fringe vendors — they are among the most trusted assurance brands on earth, caught publishing unverified AI output as fact.

It is not isolated. In its 2022 inspection cycle, the U.S. audit regulator (PCAOB) found deficiencies (insufficient audit evidence, Part I.A) in roughly **40%** of *all inspected* audits — up from 34% in 2021 and 29% in 2020 — and in the firm-specific inspection reports, EY's rate reached **46.3%**, more than double its prior year. The firms whose entire product is assurance are missing on their own core work at record rates. If professional verification is this fragile, a vendor's self-graded number deserves *less* trust, not more.

You own the output. In *Moffatt v. Air Canada* (2024 BCCRT 149), the British Columbia Civil Resolution Tribunal held Air Canada liable after its website chatbot hallucinated a bereavement-fare policy that did not exist. The airline argued the chatbot was a "separate legal entity" responsible for its own actions. The tribunal rejected that outright: the company is responsible for all information on its site, chatbot included, and had failed to take reasonable care to ensure accuracy. Damages were **$650.88** — small; the precedent is not. Accountability for autonomous output sits with the deployer, full stop.

The dashboard that lied. In 2024, Klarna became the reference case for AI customer service: its assistant, the company said, was doing the work of ~700 agents and handling roughly two-thirds of chats within a month. By May 2025, CEO Sebastian Siemiatkowski reversed course — *"we*

went too far" — as quality dropped and customers grew dissatisfied; Klarna began rehiring humans and moved to a hybrid model. The mechanism is the lesson: the bot excelled at high-volume, low-stakes queries, so the top-line metrics looked excellent, while it failed on the low-volume, high-consequence interactions — disputes, fraud, escalations — that barely moved the average. The failures hid inside a healthy-looking number.

Confidently wrong, and getting wronger. An independent watchdog (NewsGuard) audited the ten leading AI chatbots over a year and found they repeated false claims about news roughly **35%** of the time — nearly double the ~18% rate a year earlier. The rate rose *because* the models became more willing to answer, pulling from a polluted information ecosystem and treating unreliable sources as credible. Capability rose while reliability fell — and the model's tone of certainty never changed.

The scoreboard is broken. The benchmarks the industry cites to prove progress are themselves methodologically shaky. A 2024 review (Bean et al.) of **445** benchmark papers found only **16%** used any statistical or uncertainty test to support their comparisons, and 27% relied on convenience sampling. Stanford's 2026 AI Index, citing benchmark-reliability research, reported invalid-question rates up to **42%** on some widely-used benchmarks (e.g., GSM8K). A benchmark where nearly half the questions are flawed is not a strict benchmark — it is marketing shaped like measurement.

3. Why it happens

Three structural forces, none of which are anyone's bad intent:

- **The grader is the graded.** When the party that builds or profits from a system also measures it, the measurement bends toward the flattering answer. Every mature industry eventually separates the two — trials from drugmakers, audits from management, aircraft certification from airlines. AI has not yet.
- **Dashboards count what's easy.** Volume, deflection, "handled," closed — all cheap to count and all silent on whether the underlying problem was actually resolved. The expensive failures are rare, complex, and invisible to the metric, exactly as Klarna's were.
- **Confidence is not accuracy.** A model does not know when it is wrong, so it cannot flag it. It reports the same success state for a correct answer and a hallucination. The fluency is not the evidence.

And underneath all three: **there is no single, stable number.** Measured accuracy depends on method, version, and date — the same model scores differently on different evaluations, and the yardsticks themselves are versioned and swapped (Vectara alone runs a versioned HHEM leaderboard and a separate, stricter FaithJudge benchmark that disagree). A figure frozen on a slide is a photograph of a moving target, taken by the vendor.

4. What the rules actually say

The regulatory direction is real, but it is widely overstated — and for a report like this, precision matters.

The EU AI Act does **not** broadly require independent third-party verification of all high-risk AI. Under Article 43, most high-risk systems in Annex III (credit scoring, hiring, and similar) currently follow **self-assessment via internal control** (Annex VI); third-party assessment by a notified body (Annex VII) is mandatory only in specific cases. What the Act *does* enshrine is the principle: **Article 31** requires that conformity-assessment bodies be *independent* of the provider and bars them from conflicting consultancy. In other words, the law's own logic says an assessor must be independent — while today's default still lets most high-risk AI grade itself. That gap between principle and practice is precisely the space independent verification fills.

On liability, the direction is settled: *Moffatt v. Air Canada* confirms that deployers own their AI's outputs. The reckoning is not a future regulatory cliff; it is already arriving through the courts, one case at a time.

5. The cost of the gap

We decline to invent a precise dollar figure — doing so would violate the standard this report is built on. But the shape is clear from the cases above. A metric that counts containment instead of resolution understates the true cost of every unresolved interaction: the repeat contact, the escalation, the downstream work, the churn. Klarna paid for it in a public reversal and a rehiring cycle. Air Canada paid for it in court. Deloitte paid for it in a refund and a reputational hit. The bill for an unverified AI number is rarely on the invoice — it arrives later, larger, and attributed to something else.

6. The path: independent verification

The fix is not more dashboards, a better leaderboard, or a compliance checklist. It is an independent, reproducible verdict on the one question that pays your bills: **did *this* system, in *your* production, on *your* data, deliver what it reported?** No vendor can grade that for you; no lab benchmark answers it.

What "independent" has to mean, concretely:

- **Structurally neutral.** No equity, board seats, or consulting fees from any party being measured. Buyer-paid, not issuer-paid.
- **Reproducible.** A method published and re-runnable, so the number survives a second look.

- **Primary-sourced and adversarially checked.** Every figure traced to ground truth and stress-tested before it stands.
- **Outcome-level, not process-level.** Measuring what the system *delivered in production*, not whether the paperwork is in order.

That is the layer the market is missing, and the one this firm exists to provide.

Methodology & sources

This report uses only documented events traceable to a citable source, prioritizing primary or top-tier sources (regulators, courts, the original report or dataset, the company's own statements). **In keeping with its own thesis, this report was fact-checked by the same engine we apply to any claim** — and where our first sourcing was secondary or imprecise, we corrected it (notably, Air Canada damages are $650.88 per the tribunal decision, not the $812 figure common in secondary coverage). Figures are point-in-time and will be revisited; this is a living document.

#	Claim	Source	Type
1	Deloitte AU ~A$290k report, fabricated citations + fake judge quote, partial refund; second Canadian report	Fortune (Oct 7 & Nov 25, 2025)	Top-tier secondary
2	PCAOB 2022: ~40% of *all inspected* audits deficient, Part I.A (34% '21, 29% '20); EY firm rate 46.3%	PCAOB 2022 inspection release (primary) + EY firm-specific inspection report (via Compliance Week)	Primary + secondary
3	*Moffatt v. Air Canada*: deployer liable; "separate entity" defense rejected; $650.88 damages	BC Civil Resolution Tribunal, 2024 BCCRT 149	Primary
4	Klarna 2024 AI claims; May 2025 CEO reversal ("we went too far"), quality drop, rehiring	Bloomberg (May 2025); Siemiatkowski statements	Top-tier
5	NewsGuard 2025 audit: ~35% false-claim rate on news, up from ~18%	NewsGuard one-year AI audit report, 2025	Primary
6	445 benchmark papers reviewed; only 16% used statistical tests; 27% convenience sampling	Bean et al., 2024 (construct-validity review)	Primary (preprint)
7	Invalid-question rates up to 42% (GSM8K) on some benchmarks	Stanford AI Index 2026, citing benchmark-reliability research (e.g., arXiv 2511.16842)	Primary + preprint
8	EU AI Act: Art. 31 independence for notified bodies; Annex VI self-assessment default; Annex VII third-	EU AI Act text (Articles 31, 43; Annexes VI/VII)	Primary (official text)

#	Claim	Source	Type
	party only in specific cases		
9	Measured AI accuracy depends on method/version/date; versioned HHEM + separate FaithJudge	Vectara (GitHub: hallucination-leaderboard, FaithJudge); cf. MSOS Recount Study 01	Primary