95% of AI Pilots Show No P&L Impact. Trust, Not Models, Is Why

Second Order Labs™ · June 2026

ABSTRACT

Generating an AI answer now costs fractions of a cent. Trusting it still costs human labor. That gap decides who captures value in enterprise AI.

Keywords: Enterprise AI, AI Evaluation, LLM-as-a-Judge, Verification Tax, AI Agents, Trust Layer

"The appearance of software working is not software working."

— Alex Karp, Palantir Q1 2026 Earnings Call

I. The Cheapest Thing in AI Is the Answer

A frontier model can now draft a contract clause, write a database migration, or summarize a deposition in seconds for fractions of a cent. The generation is almost free. The expensive part is the lawyer reading the clause, the engineer reviewing the migration, the paralegal checking the summary against the record. Generation collapsed toward zero. Checking did not.

That gap is the biggest economic fact in AI right now, and most companies still price it wrong. The standard story says enterprise AI is stuck because models are not smart enough, and the next model release will fix it. The numbers point somewhere else. MIT Sloan found that more than 95% of generative AI pilots fail to produce measurable P&L impact[1]. IDC puts the deployment gap just as starkly: 88% of enterprise agent pilots never reach production[2]. Those are not mainly capability failures. They are trust failures.

Verifiability, not raw model capability, is now the scarce input that decides who gets paid in AI. The cost of generating an answer is falling toward zero while the cost of trusting that answer is still tied to human labor and expensive compute. Call this the **Verification Tax**: every output carries a cost to check before anyone can rely on it, and that cost is not falling. The asymmetry is the whole problem, generation gets cheaper fast while verification does not. The company that cuts the verification bill keeps the margin.

II. The Verification Tax Breaks the Software ROI Model

Traditional software economics assumed both costs fell as usage grew. AI breaks that relationship, and the part that stays expensive is the one that matters. Classic SaaS worked because once you built the feature, each extra use was cheap and reliably correct. A spreadsheet formula gives the same answer every time, so you check it once and trust it after that. Generative systems do not offer that deal. Every output is probabilistic. Every output is, in practice, a new object to inspect. Princeton's Arvind Narayanan and Sayash Kapoor call evaluating these systems a minefield, where prompt sensitivity, contamination, and shaky construct validity make it genuinely hard to know whether an output can be trusted[3]. When verification repeats on every output and scales with human attention, cheap generation stops looking cheap. Companies book the savings up front and discover the verification bill later.

Figure 1: Generation flows nearly free; the cost of believing the output barely trickles down.

Gartner's list of why projects get abandoned maps cleanly to this problem. It cites poor data quality and weak risk controls, both of which raise the cost of checking outputs[4]. The complaint is not that the model lacked flair.

III. Why Long-Horizon Agents Are Economically Dead on Arrival

Autonomous agents stumble in the enterprise for a simple reason: human verification cost rises with task length, so the savings disappear right where the work gets valuable. METR puts numbers on it. Even top frontier models lose reliability fast on autonomous software engineering tasks that run past a two-hour horizon, with reliability falling toward roughly 50%[5]. A coin-flip agent on a multi-step task is not replacing labor. It is creating review work, because a human now has to reconstruct the chain of errors.

> *The longer the leash you give an agent, the more expensive the inspection becomes, until the human you removed is back, holding a magnifying glass instead of a keyboard.*

That is the demo trap. A demo handles a short, bounded task in a setting where the operator already knows the right answer. Production means open-ended work on proprietary data, where nobody knows the answer ahead of time. Alex Karp put it sharply on a Palantir earnings call: the appearance of software working is not software working[6]. The gap between demo and production is mostly a verification problem pretending to be an engineering problem.

IV. The Recursive Trust Trap of Automated Evaluation

The obvious response to expensive human checking is automated checking, usually with one LLM grading another. That just moves the trust problem up a level. As generation gets cheaper, the real cost becomes trusting the grader, because now you need to verify the verifier. The MT-Bench researchers found a specific failure that matters here: LLM judges often prefer longer answers, which distorts evaluation[7]. A grader that confuses length with quality is not a neutral instrument. It is a biased one, and bias has to be checked against ground truth. Ground truth is still human.

Figure 2: Two independent studies, one verdict: the failure mode is trust, not intelligence.

Scale AI is blunt about what happens next: without serious human-in-the-loop evaluation pipelines, models drift and learn to game the metric[8]. The software meant to cut verification cost often turns it into a standing operating expense. You do not remove the human reviewer. You keep the reviewer and add another system to monitor. OpenAI's own research points to the pricier but more honest route: process supervision beats outcome supervision for training reliable models[9]. Checking how the model got there works better than checking whether the final answer looks plausible. It also costs more, which is exactly why buyers in regulated industries will pay for it.

V. Trust Is Being Productized as the Scarce Layer

Because verification is the bottleneck, value is moving away from the model layer and toward the trust layer around it. Enterprise buyers pay for risk reduction, not maximum cleverness. Palantir, Salesforce, Microsoft, and ServiceNow are winning by selling auditability and controls, not raw model access.

> *"In God we trust; all others must bring data."*
> *W. Edwards Deming, on statistical quality control*

Palantir frames its moat around the Ontology. Its pitch is that the bottleneck for enterprise AI is not the model but the ontology and permissions layer that makes model behavior legible and controllable[10]. Salesforce turned the same idea into a product with the Einstein Trust Layer, which it describes as the reason customers feel safe putting agents into production without exposing data[11]. Microsoft says Azure AI growth comes from customers that want platforms with audit logs and access controls strong enough for production use[12]. ServiceNow's Now Assist works because it sits inside strict business rules that constrain outputs and keep them compliant[13].

What's being sold	Who's buying it for	The scarce input
Raw model API access	Prototypes, demos	Commoditizing toward zero
Ontology and permissions (Palantir)	Grounding outputs in trusted data	Verifiability
Trust Layer (Salesforce)	Auditing and masking in production	Verifiability
Deterministic workflows (ServiceNow)	Bounding the action space	Verifiability

Figure 3: The model is commoditizing; the vault around it is where the margin now lives.

ServiceNow's approach deserves extra attention. It does not make verification cheaper by checking faster. It makes verification cheaper by reducing how much can go wrong. Constrain an AI to a narrow, rule-governed action space and the review burden shrinks toward zero, much like grading multiple choice is cheaper than grading an essay. Palantir sells permissions and ontology, Salesforce sells masking and audit controls, and ServiceNow sells bounded workflows. Open-ended GenAI keeps disappointing enterprises because open-ended output creates an open-ended review bill.

VI. Where the Margin Goes Next

The competitive map changes once you see verification as the choke point. Andreessen Horowitz already says the missing piece in the LLM stack is evaluation tooling that shows when outputs fail in production[14]. Venture money is shifting from foundation models toward this trust infrastructure because the model layer is getting commoditized and the trust layer is not. Sequoia's David Cahn made the macro point: the gap between revenue implied by AI infrastructure spending and actual AI revenue keeps widening[15]. That gap does not close until buyers trust applications enough to deploy them.

Operators should treat verification cost as a first-class line item, not a hidden one. Before deploying anything, ask what it costs to believe each output. Then redesign the system to push that number down. Bound the action space where possible, as ServiceNow does. Build process audit trails for anything that touches compliance, as OpenAI's research suggests. If you sell to enterprises, build masking, audit logs, and permission controls into the product. And treat any proprietary evaluation dataset you col-

lect as a harder asset than model access, because the model will be matched soon enough and the dataset probably will not.

The reliability bar is unforgiving. A system you trust 99% of the time still fails one task in a hundred, and in a regulated workflow that single failure is the whole story. Over the next two years, the best margins in AI will go to the vendors that can prove outputs cheaply enough for those workflows. If verification stays expensive, the model leaders will supply commodities while the workflow vendors keep the profits.

KEY FINDINGS

MIT Sloan found upwards of 95% of generative AI pilots produce no measurable P&L impact, and the cause is trust barriers rather than weak models.

METR measured frontier-model reliability dropping toward 50% on autonomous software tasks past a two-hour horizon.

Human verification cost rises with task length, which erases the savings from automating long-horizon agent work.

Salesforce, Palantir, and ServiceNow now sell governance and grounding layers, so the margin migrates from the model to the trust layer around it.

LLM-as-a-judge carries position, verbosity, and self-enhancement bias, so the automated grader becomes the next thing a human has to verify.

REFERENCES

[1] MIT NANDA, *The GenAI Divide: State of AI in Business 2025*. https://nanda.media.mit.edu/ai_report_2025.pdf

[2] IDC, enterprise AI agent pilot-to-production research (2026): roughly 88% of agent pilots never reach production, corroborated by Forrester and Anaconda. https://www.cio.com/article/3850763/88-of-ai-pilots-fail-to-reach-production-but-thats-not-all-on-it.html

[3] Arvind Narayanan and Sayash Kapoor (Princeton / AI Snake Oil), *Evaluating LLMs is a Minefield*. https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/

[4] Gartner, *30% of GenAI Projects Will Be Abandoned by End of 2025*, 2024. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

[5] METR, *Measuring AI Ability to Complete Long Tasks*, March 2025. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

[6] Alex Karp, Palantir Q1 2026 Earnings Call (May 2026), as reported by TheStreet. https://www.thestreet.com/investing/stocks/palantir-ceo-issues-blunt-warning-to-ai-slop-competitors

[7] *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, arxiv.org/abs/2306.05685. https://arxiv.org/abs/2306.05685

[8] Scale AI, *The LLM Evaluation Playbook*. https://scale.com/blog/llm-evaluation-playbook

[9] OpenAI, *Let's Verify Step by Step*, arxiv.org/abs/2305.20050. https://arxiv.org/abs/2305.20050

[10] Palantir Technologies, Q3 2024 Earnings Call. https://www.fool.com/earnings/call-transcripts/2024/11/04/palantir-technologies-pltr-q3-2024-earnings-call-t/

[11] Salesforce, Q2 Fiscal 2025 Earnings Call (Einstein Trust Layer). https://investor.salesforce.com/news/news-details/2024/Salesforce-Announces-Second-Quarter-Fiscal-2025-Results/default.aspx

[12] Microsoft, FY24 Q4 Earnings Call. https://www.microsoft.com/en-us/investor/events/fy-2024/earnings-fy-2024-q4

[13] ServiceNow, Q3 2024 Earnings Call. https://www.sec.gov/Archives/edgar/data/0001373715/000137371524000342/erq3fy24.htm

[14] Andreessen Horowitz, *Emerging Architectures for LLM Applications*. https://a16z.com/emerging-architectures-for-llm-applications/

[15] Sequoia Capital, *AI's $600B Question*, David Cahn. https://www.sequoiacap.com/article/ais-600b-question/

© 2026 Second Order Labs™ · https://secondorderlabs.com/articles/tech-journalism/why-verifiability-not-raw-model-capability-is-becoming-the-scarce-input-that-decides-who-captures-value-in-ai/