FlowX.AI

TECHNICAL PAPER

VERIFIABLE, EVIDENCE-GRADED RETURN ON AGENTS (VERA)

Evidence-Graded ROI for Production Agents

Bringing counterfactual measurement, causal validation, and confidence-graded accounting to the business case for enterprise AI agents.

Bogdan Răduță
Head of Research, FlowX.AI
bogdan.raduta@flowx.ai

FlowX.AI · Applied Research

August 2026 · Version 1.1

ABSTRACT

Enterprises are deploying AI agents faster than they can prove the agents pay for themselves. Industry research reports that 95% of generative-AI pilots produce no measurable profit-and-loss impact[1], and analysts project that more than 40% of agentic-AI projects will be cancelled by 2027, primarily for unclear business value[2]. We argue that this is not a value crisis but a *measurement-credibility* crisis: prevailing ROI practice rests on flat productivity assumptions and self-reported time savings that finance functions correctly decline to capitalize. We present **VERA (Verifiable, Evidence-Graded Return on Agents)**, a measurement framework that embeds the causal-inference gold standard (counterfactual estimation and randomized holdout validation[4,5]) directly into the agent observability layer, and grades every reported dollar by the strength of its evidence. VERA computes return across three reconciled dimensions: a deterministic per-agent baseline, a per-execution counterfactual estimator calibrated against an organization's own labelled history, and a realized-outcome ledger validated by causal experiments. Each figure is then classified as **Verified**, **Modeled**, or **Assumed** (an evidentiary grade analogous to revenue-recognition tiers in financial accounting[10]), with verified value credited conservatively at the lower bound of its confidence interval. We describe the architecture, the estimator and calibration mechanics, the confidence-grading and causal-validation protocols, and the governance surface that turns graded ROI into portfolio decisions. We illustrate the framework on a regulated-insurance deployment and discuss how evidence-graded accounting changes the conversation between AI teams and the CFO.

KEYWORDS agent ROI · counterfactual estimation · causal validation · confidence grading · productivity measurement · value outcome units · AI governance · FinOps for AI

1 Introduction

The enterprise has decided to buy agents. What it has not decided is whether they are worth it. In 2026, the board-level question shifted from *can we deploy AI agents?* to *can we prove they pay?* And the prevailing answer is uncomfortable. A widely cited study of 300 public deployments and several hundred enterprise leaders found that roughly 95% of generative-AI pilots delivered no measurable impact on the profit-and-loss statement[1]. Industry analysts forecast that more than 40% of agentic-AI projects will be cancelled by the end of 2027, citing escalating costs, inadequate risk controls, and, above all, unclear business value[2]. Surveys put the share of executives who can confidently quantify AI return at under a third.

It is tempting to read these numbers as a verdict on the technology. We read them differently. The same period produced rigorous field evidence that AI assistance produces real, large productivity effects: randomized experiments measured task-time reductions near 40% for professional writing[4], and a study of more than 5,000 customer-support agents found a 15% average gain in issues resolved per hour, concentrated among less-experienced workers[5]. Value is being created. What is missing is a way to *measure* that value that a chief financial officer will accept onto the books.

The gap is methodological. Most agent ROI reporting today rests on one of two weak foundations: a flat per-agent assumption ("each run saves an analyst 30 minutes") multiplied by volume, or self-reported time savings collected after the fact. Neither establishes a counterfactual (what would have happened *without* the agent), and neither carries any indication of how strongly the claim is supported. Finance functions are right to discount such figures. They have, in effect, no audit trail. The causal-inference literature has long held that the credible way to measure an intervention's effect is a controlled comparison against a counterfactual[9], and the productivity studies cited above derive their authority precisely from randomized or staggered holdout designs. That rigor lives in academic papers and one-off internal studies; it does not live in the tools enterprises use to run agents day to day.

> *The agent-ROI crisis is not a value crisis. It is a measurement-credibility crisis, and credibility is an engineering property of the measurement system, not a rhetorical one.*
>
> VERA · CORE THESIS

This paper argues that ROI measurement for agents should be treated as a first-class, instrumented discipline embedded in the observability layer, and that every reported figure should carry an explicit grade of evidence. We draw a deliberate analogy to financial accounting: revenue is not recognized uniformly, but according to rules that distinguish realized revenue from estimates and from contingent expectations[10]. ROI for agents deserves the same discipline. A dollar backed by a randomized holdout experiment is not the same asset as a dollar inferred from a model assumption, and a reporting system that conflates the two cannot be trusted by the people who control budgets.

We call the resulting framework **VERA (Verifiable, Evidence-Graded Return on Agents)**. VERA is the measurement framework that the **ROI Hub** of the FlowX.AI Observatory implements. The Observat-

ory is the platform's observability and governance plane for production agents; the ROI Hub is its financial-justification surface.

1.1 Contributions

- A reframing of the agent-ROI problem as a measurement-credibility problem, and a design principle, *grade every dollar by its evidence*, borrowed from revenue recognition.
- A three-dimension measurement architecture that computes a deterministic baseline, a per-execution counterfactual estimator, and a realized-outcome ledger side by side, with explicit reconciliation between them.
- A per-execution counterfactual estimator that converts agent output into estimated human-hours and calibrates against an organization's own labelled history with a log-space correction, designed to be conservative and unbiased in aggregate.
- A confidence-grading protocol (Verified / Modeled / Assumed) in which causally verified value is credited conservatively at the lower bound of its confidence interval, and in which each unit of value is counted under exactly one tier.
- A governance surface that converts graded ROI into portfolio decisions, plus anti-gaming mechanisms (an acceptance signal and a per-operation inclusion policy) that keep the measurement honest.

2 Background and the Measurement Gap

2.1 How agent ROI is measured today

The dominant practice is the *flat baseline*: a fixed estimate of the manual effort a task would take, multiplied by the number of times an agent performs it. It is cheap, deterministic, and requires no per-run computation, which makes it a reasonable default for aggregate reporting. Its weakness is that it is unconditional: it credits the same value to a run that produced a flawless result and one whose output was discarded, and it cannot distinguish an agent doing hard work from an agent doing trivial work. The alternative commonly seen in vendor materials, self-reported productivity surveys, is more conditional but less defensible: it is subject to recall bias, selection effects, and the well-documented tendency of respondents to over-attribute gains to a new tool. Recent commentary has warned that reducing AI value to "minutes saved" misses both quality and the counterfactual entirely[8].

2.2 The counterfactual gold standard, and why it stays in the lab

The credible estimand for any intervention is the difference between observed outcomes and the outcomes that would have obtained without it[9]. In practice this is approximated by a control group: workers, tasks, or accounts that do not receive the agent. The strongest evidence on AI productivity is built this way: randomized assignment in controlled writing tasks[4], staggered rollout with a pilot randomized trial across thousands of support agents[5], and randomized developer-productivity trials that have repeatedly surprised their own authors[6]. The lesson of this literature is not only that AI helps, but that the size of the effect is frequently mis-estimated without a holdout. Yet almost no production platform runs holdouts as a standing capability; causal measurement is treated as a research project, not as plumbing.

2.3 Estimating effort is hard, but unbiased in aggregate

A second strand of work asks how much human effort a given piece of machine output represents. Published analyses have shown that even seemingly strong proxies are weak: a line-of-code predictor of engineering effort explains only about a quarter of the variance in log-effort[7], and a single per-agent constant is weaker still. The practical implication is important and slightly counter-intuitive: a per-execution estimate that conditions on what actually happened in the run is noisy at the level of any individual run, but if it is roughly unbiased, those errors cancel as runs accumulate. The aggregate, the organization-wide number a CFO cares about, can therefore be trustworthy even when no single estimate is. This is the statistical foundation of VERA's estimator dimension.

2.4 The assurance gap

Financial reporting solved an analogous trust problem a century ago, not by making every number certain, but by grading and auditing it: realized revenue, estimates, and contingencies are recognized under different rules and disclosed separately[10]. Emerging AI-management standards push organizations toward measurable, auditable processes[11], and regulation in high-risk domains increasingly demands evidence rather than assertion[12]. No widely used agent platform currently attaches an evidentiary grade to the ROI it reports. Table 1 summarizes how the common approaches

compare against the properties a finance function actually requires; VERA is designed to satisfy all of them at once.

Table 1. ROI measurement approaches against the properties finance requires.

Approach	Conditional on the run?	Counterfactual?	Graded by evidence?	Self-correcting?
Flat per-agent baseline	No	No	No	No
Self-reported time savings	Partly	No	No	No
One-off RCT / holdout study	Yes	Yes	Implicit	No (point-in-time)
VERA (this work)	Yes	Yes (standing)	Yes (explicit)	Yes (calibration loop)

3 The VERA Framework

VERA computes return across three dimensions of increasing evidentiary strength and reconciles them under a single confidence grade. The dimensions are not alternatives to choose between; they are layers that co-exist, each visible, each feeding a confidence engine that decides how much of the value may be credited and at what tier. Figure 1 gives the architecture.

FIG. 1 · VERA THREE-DIMENSION MEASUREMENT ARCHITECTURE

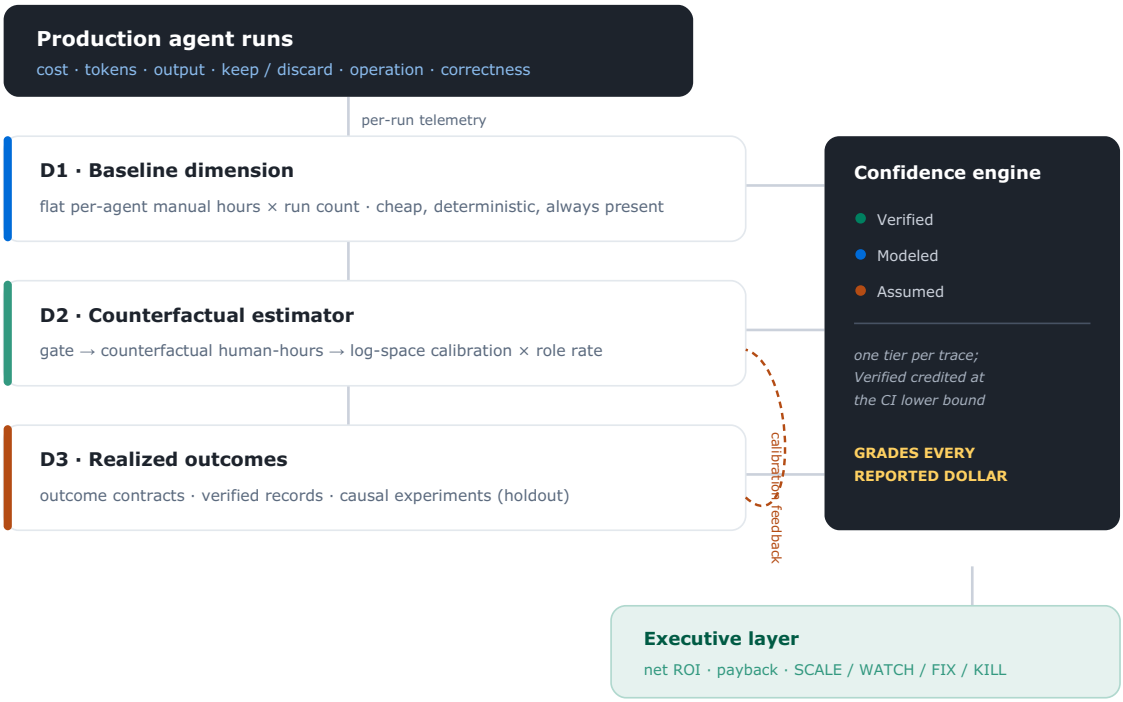


Figure 1. The three measurement dimensions co-exist over a single stream of agent-run telemetry. Each feeds a confidence engine that assigns one evidentiary tier per unit of value (Verified, Modeled, or Assumed) and credits verified value conservatively. Realized outcomes feed back to re-calibrate the estimator, closing the loop. The executive layer consumes only graded value.

3.1 The baseline dimension (D1)

The baseline is intentionally simple and is never removed. For each agent, a configured manual-effort figure (hours) and an hourly role rate produce a per-run value that, multiplied by volume, yields an aggregate. The baseline requires no model calls and no calibration, which makes it a stable anchor: it is always available, it is the same number every time, and existing reporting built on it continues to work. Its role in VERA is to be the floor and the sanity check, the *Assumed* tier, against which the conditional dimensions are reconciled. Crucially, VERA exposes the reconciliation ratio between the baseline and the estimator so reviewers can see precisely where the flat assumption over- or under-credits real work.

3.2 The counterfactual estimator (D2)

The second dimension answers a conditional question for every run: *how many human-hours would this specific output have taken to produce?* It is a three-stage pipeline (gate, estimate, calibrate) summarized in Appendix B.

Usefulness gate. Before any value is estimated, a gate decides whether the run produced useful output at all. Deterministic predicates run first (for example, an output that failed validation, or produced nothing, is discarded); only genuinely ambiguous runs fall through to an optional lightweight classifier. The gate is recorded with the run, so every estimate carries the method by which it was admitted (deterministic, classifier, policy, or estimator) and is fully auditable.

Counterfactual hours. An admitted run is passed to a single estimation call that returns the estimated human-hours the output represents, a self-reported confidence, and a short rationale, bounded by a fixed trace-token budget. The metric is explicitly the counterfactual *human* effort, never the agent's own wall-clock or step count, a distinction that matters, because an agent that runs for two seconds may replace two hours of human work, or two minutes of it.

Calibration. Raw estimates are corrected against the organization's own labelled samples with a log-space fit, $h = a \cdot raw^{b}$, resolved most-specific-first (organization-and-type, then organization, then type, then a global identity of $a = b = 1$). A near-unit exponent indicates the raw estimator is already close to a constant multiplier of truth; the fit absorbs systematic bias while leaving the per-run noise to cancel in aggregate. The system ships disabled and falls back to identity until a fit exists, so it can never silently inflate a number. An organization promotes the estimator from "experimental" to a reporting default only after it clears an acceptance bar on held-out labels: a minimum log-space fit quality, no systematic residual trend, and an aggregate ratio in a sane band.

3.3 The realized-outcome ledger (D3)

The third dimension records value that actually occurred. An *outcome contract* declares, for a given business outcome, what counts as a realized event, where the truth comes from (a platform event, a signed webhook, or a batch upload), how it is economically modelled (labor, error-avoidance, revenue, velocity, or compliance), and what it is worth. An *outcome record* is one occurrence, linked back to the originating run, with a status of recorded, verified, disputed, or voided. Where the baseline and estimator *model* value, D3 *observes* it, and it is only this dimension that can produce Verified credit.

3.4 Confidence grading: Verified, Modeled, Assumed

The element that ties the three dimensions together (and the framework's central contribution) is that every figure of value is assigned exactly one evidentiary tier, and the tiers are mutually exclusive per trace so that value is never double-counted. Figure 2 shows the grading.

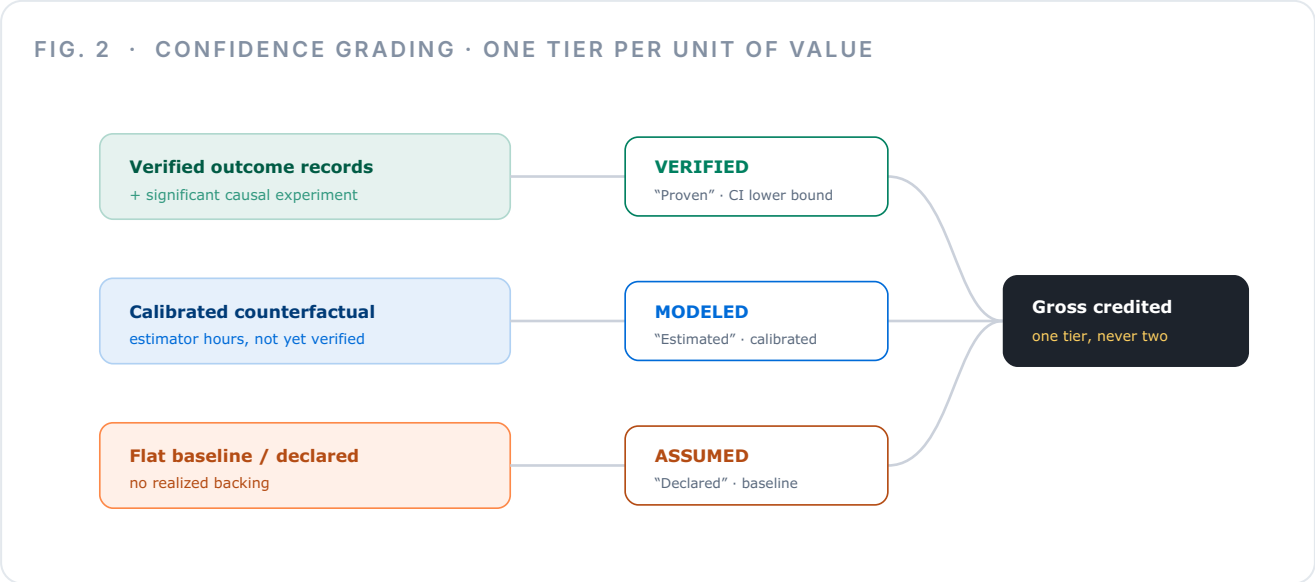


Figure 2. A unit of value is graded by the strongest evidence available for it and credited under that tier only. **Verified** value requires an outcome contract with verified records or a statistically significant causal experiment, and is credited at the lower bound of the confidence interval rather than the point estimate. **Modeled** value is the calibrated estimator output; **Assumed** is the flat baseline. Executive views relabel the tiers Proven / Estimated / Declared.

Two design choices give the grade its integrity. First, **verified value is credited conservatively**: when a causal experiment establishes a lift, VERA books the lower bound of the confidence interval, not the central estimate, and records the suppressed remainder separately. The reported Verified number is therefore one the organization can defend under scrutiny: it is the value that survives even a pessimistic reading of the experiment. Second, **tiers are exclusive**: a trace credited as Verified is removed from the Modeled and Assumed pools, so the headline figure is a clean sum rather than an overlapping one. The executive surface relabels the tiers in plain language (Proven, Estimated, Declared), but the underlying discipline is unchanged.

3.5 Causal validation as standing infrastructure

What lets a figure reach the Verified tier is a causal experiment that VERA runs in production rather than in a study. Traces are tagged into arms, agent or holdout, and an experiment, randomized (Tier A) or quasi-experimental (Tier B), measures the realized lift of the contracted outcome against the holdout, returning the lift, its confidence interval, a p-value, and the sample size. The statistics are deliberately simple and transparent. The significance of this is architectural: the gold-standard method that the productivity literature treats as a one-time research undertaking[4,5,6] becomes a button in the platform, with the holdout maintained continuously and the result wired directly into the confidence grade. An organization does not commission a study to find out whether its claims-triage agent works; it reads the experiment that has been running the whole time.

3.6 The calibration loop

The estimator is kept honest by the outcomes it is eventually measured against. As verified outcome records accumulate, VERA re-fits the estimator's calibration constants against realized value and tracks the drift between what the estimator predicted and what actually occurred. The Modeled dimension thus converges toward the Verified one over time, and a widening drift is itself a signal: that the workload has shifted, that the estimator is stale, or that a contract is mis-specified. This is the self-correcting property that one-off studies structurally lack: VERA does not produce a number and walk away; it produces a number and then spends the rest of the deployment checking it.

4 Trustworthy by Construction

A measurement system is only as credible as its resistance to the incentives acting on it. Because ROI numbers are used to justify budgets, they attract optimism; VERA therefore builds in ground truth and anti-inflation controls rather than relying on good intentions.

4.1 A load-bearing acceptance signal

Where agents produce output that a human accepts or rejects (a generated draft kept or discarded, a suggestion approved or dismissed), VERA captures that decision as a first-class signal. Because the user must choose one or the other to proceed, coverage is high and the signal is not a survey but an observed action. Acceptance feeds both the usefulness gate (discarded work contributes no value) and an acceptance-rate view that lets reviewers see how much of the reported ROI rests on work that was actually kept and judged correct. It is the cheapest honest counterfactual available: the human told us, in the act of working, whether the output was worth anything.

4.2 Per-operation inclusion policy

Not every action an agent takes represents replaced human work. Routing, intent-classification, and inspection operations are internal plumbing; counting them would inflate ROI with motion that no human would ever have performed. VERA exposes a per-operation inclusion policy so that producer operations opt in and pure-overhead operations opt out. Excluded work is still recorded (auditable, with an explicit reason) but contributes zero hours and zero dollars, and is filtered before it ever reaches an estimation call. The effect is that the denominator and numerator of the business case are scoped to genuine value, by policy, not by accident.

4.3 Value Outcome Units and portfolio verdicts

VERA groups agents by the business outcome they serve (a Value Outcome Unit) rather than reporting per-agent vanity metrics. This is what allows the executive layer to issue a verdict for each unit: **SCALE** when graded ROI and adoption are strong, **WATCH** when the case is promising but thin, **FIX** when value exists but is being eroded by cost or low acceptance, and **KILL-CANDIDATE** when neither the evidence nor the economics support continuation. Each verdict carries a one-line reason traceable to the underlying tiers. The portfolio view turns a pile of agent telemetry into the small set of decisions a leadership team actually has to make.

4.4 Risk framing for the board

Finally, VERA reports value as a distribution, not a point. A Monte-Carlo sweep over the uncertain inputs (model cost, infrastructure, labor saved, revenue attributed, compliance exposure) produces conservative, base, and optimistic scenarios with percentile statistics and the probability that ROI is positive at all. A five-year total-cost-of-ownership projection sets the agent option against the realistic alternatives a board would otherwise fund: robotic process automation, outsourcing, additional hiring, or the status quo. The point is not to manufacture precision but to express honest uncertainty in the language finance already uses for capital decisions.

5 Illustrative Application

To make the mechanics concrete we trace a single quarter for an illustrative regulated insurer, "Meridian Insurance," running a fleet of build-side and execution-side agents on the platform. The figures here are illustrative and chosen to expose the method, not to report a customer result; the purpose is to show *how a headline number is constructed and graded*, which is the part that survives changes in magnitude.

Aggregated naively, the quarter looks spectacular: roughly €4,200 of platform and infrastructure cost against on the order of €380,000 of modelled value, an 88× return. Under prevailing practice that single number would go to the board, and the board would be right to distrust it. VERA does not report it that way. It decomposes the gross value by evidentiary tier and credits each under its own rule, as in Table 2.

Table 2. Illustrative decomposition of a quarterly figure by evidentiary tier.

Tier	Source of the value	Crediting rule	Confidence
Verified · Proven	Claims-triage outcome contract, validated against a randomized holdout	CI lower bound of measured lift	High
Modeled · Estimated	Per-execution counterfactual hours, calibrated to Meridian's history	Calibrated point estimate	Medium
Assumed · Declared	Flat baselines for agents without realized backing yet	Baseline value	Low

The conversation this enables is the entire point. Instead of defending a fragile 88×, the AI team presents a Proven figure it can stand behind under audit (smaller, conservative, credited at the lower bound) alongside a larger Estimated figure clearly labelled as model-derived, and a Declared remainder flagged as not yet backed by realized outcomes. The CFO now sees not a claim but a balance sheet of evidence, and can treat each tier exactly as its strength warrants. As the quarter's holdout experiments mature and outcome records accumulate, value migrates from Declared to Estimated to Proven, and the calibration loop tightens the estimator against what actually happened. The number does not merely get reported; it gets *earned*.

6 Empirical Evaluation

Section 5's figures are explicitly illustrative. The statistical machinery behind them, the log-space calibration fit, the acceptance bar, conservative crediting at the CI lower bound, and the calibration loop's response to drift, is not: it is testable directly against synthetic data with a known ground truth, with no LLM in the loop, since what is under test is the statistics, not a model. We ran four such tests. The simulation code, raw output, and a pytest suite are released as supplementary material.

6.1 Does calibration help, and does the acceptance bar actually gate?

We simulated two scenarios for the raw counterfactual estimator against a known true-hours value: a *well-behaved* case where raw is a constant power-law multiplier of truth plus lognormal noise, the case Section 3.2's $h = a \cdot raw^b$ fit targets, and a *misspecified* case with a genuine regime change (different a, b for short vs. long tasks) that a single power-law fit cannot capture. Table 1 reports the fit on a held-out set and whether the acceptance bar (fit quality above a threshold, plus an aggregate ratio in a sane band) promotes each.

Table 1. Calibration fit and promotion decision, held-out MAPE (mean absolute percentage error).

Scenario	Raw MAPE	Calibrated MAPE	Fit R^2	Promoted?
Well-behaved	0.534	**0.324**	0.748	Yes
Misspecified	1.267	**0.408**	0.645	No

Calibration reduces error in both scenarios, including the misspecified one (MAPE 1.267 → 0.408): a single power-law fit is a real improvement over the raw estimate even when it cannot fully capture a regime change, which is itself worth knowing, calibration is not all-or-nothing. The acceptance bar correctly separates the two cases on fit quality alone (R^2 0.748 vs. 0.645 against a 0.70 threshold): a promote-gate that always says yes is not a gate, and here it says no exactly when the estimator's own diagnostic says it should. We note a methodological loose end rather than paper over it: an additional "no systematic residual trend" check the paper's design calls for turned out to have no statistical power against this regime-change misspecification when operationalized as a monotonic rank correlation, and a Ramsey-RESET-style nonlinearity test we tried instead flagged spurious curvature caused by measurement error in the regressor (an errors-in-variables artifact) even on the well-behaved scenario. We report the fit-quality and aggregate-ratio checks, which behaved as intended, and leave the residual-trend check's correct operationalization as an open question rather than a solved one.

6.2 Is Verified crediting actually conservative?

Section 3.4 credits Verified value at the lower bound of the confidence interval, not the point estimate. We simulated 2,000 randomized causal experiments (Section 3.5) with a known true lift and computed the fraction of trials in which the credited lower bound was at or below the true value, the operational meaning of "conservative." At a nominal 95% two-sided CI, the one-sided miscoverage this claim needs to hold is 97.5%; we observed **96.75%**, within Monte Carlo noise of nominal. The average

credited value across trials was 0.487 of the true 1.2-hour lift, meaning the conservative-crediting design leaves, on average, roughly half the true value on the table at this sample size (60 runs per arm) in exchange for a number that rarely overstates. That is the design working as intended, not a flaw, but it is worth stating as a number rather than an adjective.

6.3 Does the calibration loop actually respond to drift?

Section 3.6 claims the loop is self-correcting: a widening drift between the estimator and reality is itself a signal. We simulated a workload shift (the true relationship changes at round 5 of 8) and compared a calibration continuously refit on all accumulated data against one frozen at round 4, both evaluated each round on fresh post-shift data. Immediately after the shift, the frozen calibration's error is worse than the continuously-refit one (0.629 vs. 0.534 MAPE) and stays worse through round 8 (0.631 vs. 0.413): the continuously-refit calibration recovers as post-shift data accumulates, while the frozen one, by construction, never sees it and never improves. This is the mechanism working exactly as Section 3.6 describes, drift is visible in the gap between the two, and re-fitting is what closes it.

6.4 What this evaluation does and does not show

This validates the statistical machinery in isolation: the calibration fit, the acceptance bar's two well-behaved checks, conservative CI-lower-bound crediting, and the loop's response to a simulated drift. It does not exercise the usefulness gate, the realized-outcome ledger against real webhook or platform-event data, Value Outcome Units, or a real organization's causal experiments. It also leaves one check, the residual-trend diagnostic, as an open operationalization question rather than a validated one. The simulation code, raw output, and a pytest suite that reproduces Table 1 and the three simulations are released as supplementary material.

7 Discussion

VERA's contribution is less a new statistic than a new *contract* between the people who build agents and the people who fund them. By grading every dollar, it makes the disagreement that usually stays implicit, "I don't believe your ROI," explicit and addressable: a skeptic can point to the Modeled tier and ask for it to be promoted to Verified, which is now a concrete, runnable request (declare a contract, start a holdout) rather than a rhetorical standoff. This reframing generalizes well beyond insurance. Any domain where agents replace bounded human work and where outcomes can be observed (document processing, customer operations, software engineering, back-office finance) fits the three-dimension model directly; only the contracts and role rates change.

The approach also aligns naturally with the direction of regulation and assurance. Evidence-graded value is, structurally, the same artifact an auditor or a regulator wants: a claim, the method behind it, and a record of the data that supports it[11,12]. An organization that measures ROI this way is, almost incidentally, building the documentation that high-risk AI oversight demands. We expect the "grade your evidence" principle to become as ordinary for AI value as it long has been for revenue.

8 Limitations

- **The empirical evaluation is a statistical simulation, not a deployment.** Section 6 validates the calibration fit, the acceptance bar, conservative crediting, and drift response against synthetic data with a known ground truth; it does not exercise the usefulness gate, real outcome-contract data, or a real organization's causal experiments, and one acceptance-bar check (the residual-trend diagnostic) is left as an open operationalization question.
- **Counterfactual estimation remains a model.** The Modeled tier rests on an LLM's judgment of human effort. We mitigate this with calibration and by relying on aggregate cancellation of unbiased noise[7], but a systematically biased workload that the calibration set does not represent will mislead until enough labels accumulate.
- **Holdouts have a cost.** A genuine control arm means deliberately not serving the agent to some traffic, which carries an opportunity cost and is not always organizationally or ethically acceptable. Quasi-experimental designs reduce but do not eliminate this tension.
- **Outcome attribution can be contested.** Linking a realized business outcome to a specific run depends on clean instrumentation; in long, multi-touch processes the contract author still makes attribution judgments that reasonable people may dispute.
- **Conservatism is a choice, not a theorem.** Crediting Verified value at the CI lower bound deliberately under-reports; an organization optimizing for headline numbers could undo this, and the framework's integrity then depends on governance rather than mechanism.

9 Conclusion

The enterprise AI conversation has reached the stage where enthusiasm is no longer sufficient and proof is required. The data that reads as failure (pilots with no P&L impact, projects cancelled for unclear value) is better understood as the predictable result of measuring value with instruments that finance was always going to reject. VERA proposes that the fix is to treat ROI measurement as engineering: instrument it in the observability layer, hold a standing counterfactual, and grade every reported dollar by the strength of its evidence, crediting proven value conservatively and letting modelled value earn its way up the tiers over time. The result is not a bigger ROI number. It is a number a chief financial officer can sign, and, in the present climate, that is worth far more.

References

[1] MIT NANDA Initiative. *The GenAI Divide: State of AI in Business 2025.* Massachusetts Institute of Technology, 2025. (150 leader interviews, 350-employee survey, 300 public deployments; ~95% of pilots without measurable P&L impact.)

[2] Gartner, Inc. *Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027.* Press release, 25 June 2025.

[3] Gartner, Inc. *40% of Enterprise Applications Will Feature Task-Specific AI Agents by 2026, Up From Less Than 5% in 2025.* Press release, 26 August 2025.

[4] Noy, S., & Zhang, W. *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.* Science, 381(6654):187–192, 2023.

[5] Brynjolfsson, E., Li, D., & Raymond, L. R. *Generative AI at Work.* NBER Working Paper 31161, 2023 (Quarterly Journal of Economics, 2025).

[6] Model Evaluation & Threat Research (METR). *Measuring the Impact of AI on Experienced Developer Productivity: Uplift Study Update.* metr.org, 2026.

[7] Cognition. *How Much Time Does AI Actually Save? A Productivity Measurement Methodology.* cognition.ai/blog/ai-productivity (log-effort $R^2 \approx 0.27$ for a lines-of-code proxy), 2025.

[8] LSE Business Review. *AI Productivity Gains Should Be Measured in More Than Minutes Saved.* London School of Economics, February 2026.

[9] Imbens, G. W., & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[10] Financial Accounting Standards Board / IASB. *Revenue from Contracts with Customers (ASC 606 / IFRS 15).* FASB ASU 2014-09; IFRS 15, 2014.

[11] International Organization for Standardization. *ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system.* ISO, 2023.

[12] European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union, 2024.

Appendix A Evidentiary Tier Definitions

Tier	Executive label	Evidence required	Crediting rule	Promotes to
Verified	Proven	Outcome contract + verified records, or a statistically significant causal experiment vs. holdout	Credited at the confidence-interval lower bound; suppressed remainder recorded separately	– (top tier)
Modeled	Estimated	Per-execution counterfactual estimate, calibrated to org history, passing the acceptance bar	Calibrated point estimate (calibrated hours × role rate)	Verified, once an outcome/experiment confirms it
Assumed	Declared	Flat per-agent baseline or a declared figure with no realized backing	Baseline value (manual hours × rate × volume)	Modeled, once the estimator covers the agent

Appendix B Worked Example: From One Run to a Board Number

This appendix carries a single claims-triage run at the illustrative insurer of §5 all the way to its contribution to the quarterly figure. Every number is illustrative; the arithmetic is the point.

B.1 A single run

1. **Operation policy.** The run's operation is *claims.triage*, a producer operation that is opted in, so it proceeds. A routing operation such as *coordinator.intent* would have been recorded with zero value and stopped here.

2. **Usefulness gate.** Deterministic rules find a validated, schema-conformant output that the adjuster *kept* (acceptance = Keep). The run is admitted as useful with gate method *deterministic*, no classifier or estimation call is needed to reject it.

3. **Counterfactual estimate.** The estimator returns **raw = 2.5 human-hours** at medium confidence: the effort an adjuster would have spent assembling and triaging this claim unaided.

4. **Calibration.** Meridian's organization-and-type fit is *(a = 1.00, b = 0.90)*. The calibrated estimate is $\hat{h} = a \cdot raw^{b} = 1.00 \cdot 2.5^{0.90} \approx 2.28\ h$, slightly below the raw figure, reflecting the deliberately conservative fit.

5. **Dollarization.** The resolved role is *claims adjuster* at €48/h, so the run is credited $2.28\ h \times €48 = €109$.

This €109 lands in the **Modeled** tier, tagged with its gate method, calibration version, and rationale, so it is fully auditable and can later be promoted to Verified if an outcome record or experiment confirms it.

B.2 Scaling to the quarter

Over the quarter the agent logged 1,600 runs; the gate discarded 180 (no output, failed validation, or discarded by the user), leaving 1,420 useful runs with a mean calibrated estimate of ≈1.7 h. The agent's Modeled contribution is therefore $1{,}420 \times 1.7\ h \times €48 \approx €116{,}000$. Individual estimates are noisy; this total is not, for the reason given in Appendix C.

B.3 The graded board figure

Aggregated naively across the fleet, the quarter shows ≈€380,000 of value against ≈€4,200 of cost, the headline 88× of §5. VERA refuses to report it whole. A randomized holdout on the claims-triage outcome contract establishes a lift whose confidence-interval lower bound credits **€92,000 of Verified value**; the calibrated estimator contributes **€214,000 of Modeled value** across agents not yet outcome-verified; and **€74,000 remains Assumed**, resting on flat baselines. The board receives three numbers, not one, and notably, even the Proven slice alone returns roughly **22×** the quarter's entire cost, a figure that survives audit. The remaining €288,000 is not hidden; it is labelled by exactly how strongly it is supported, with a concrete path to migrate it upward as outcomes and experiments mature.

Appendix C Statistical Note: Why Noisy Estimates Make a Trustworthy Total

C.1 Aggregation

Write the calibrated estimate for run i as $\hat{h}_i = h_i + \varepsilon_i$, where h_i is the true counterfactual effort and ε_i is estimation error with $E[\varepsilon_i] = 0$ after calibration. The reported aggregate is $\hat{H} = \Sigma_{i=1}^{n} \hat{h}_i$. Its expectation is the true total Σh_i, and if the errors are approximately independent with variance σ^2, the standard error of the aggregate grows only as $\sigma\sqrt{n}$ while the total itself grows as n, so the *relative* error shrinks as $\approx 1/\sqrt{n}$. A single estimate may be off by a wide margin, consistent with the weak effort proxies reported in the literature (a lines-of-code predictor explains only ≈27% of log-effort variance[7]), yet the organization-wide figure concentrates on the truth as runs accumulate. This is the formal reason VERA reports aggregates with confidence while flagging single-run estimates as indicative only.

C.2 Calibration

The correction is fit in log space: $\log \hat{h} = \log a + b \cdot \log(raw) + \eta$. Ordinary least squares on labelled *(raw, human-hours)* pairs yields *(a, b)*; an exponent $b \approx 1$ means the raw estimator is already close to a constant multiple of truth, and the fit removes the multiplicative bias that a single per-agent constant cannot. Fits resolve most-specific-first: organization-and-type, then organization, then type, then a global identity $(a = b = 1)$ that keeps the system safe before any labels exist. An estimator is promoted from *experimental* to a reporting default only after it clears a held-out acceptance bar: adequate log-space fit quality, no systematic residual trend, and an aggregate human-to-model ratio within a sane band.

C.3 Conservative crediting

Verified value is never booked at an experiment's point estimate. Given a measured lift L with a $(1-\alpha)$ confidence interval $[L_{lo}, L_{hi}]$, VERA credits L_{lo} and records the difference $L - L_{lo}$ as suppressed. Because the realized lift exceeds L_{lo} with probability at least $1-\alpha/2$ under the interval's assumptions, the Verified figure is, by construction, one the organization is unlikely to have overstated, the property that lets it survive adversarial review. The one assumption aggregation cannot rescue is independence of errors across runs; correlated drift, such as a shift in workload mix, is caught instead by the calibration-drift monitor of §3.6.

Appendix D Sample Evidence-Graded Executive Readout

The executive surface is where graded ROI becomes a decision. Figure D.1 reproduces the one-screen readout a leadership team sees for the illustrative deployment of §5. The gross 88× is shown, but never alone: the confidence donut immediately decomposes it into what is Proven, Estimated, and Declared, the unit economics make the per-outcome cost legible, and every Value Outcome Unit carries a recommendation rather than a vanity metric. This is the artifact the preceding appendices exist to justify: a single page a chief financial officer can sign.

FIG. D.1 · EVIDENCE-GRADED ROI · EXECUTIVE (SIMPLE) VIEW

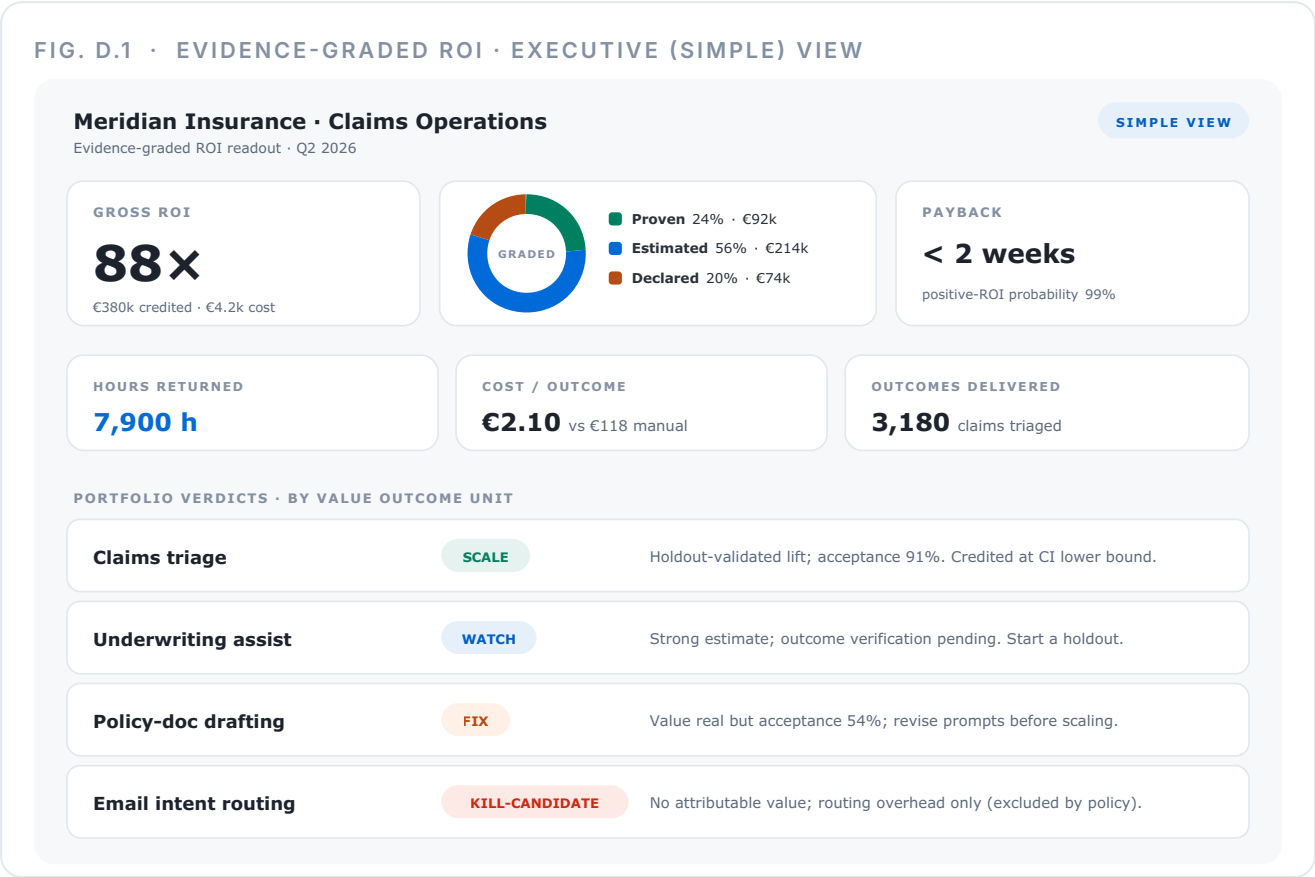


Figure D.1. The headline ROI is never shown unqualified: the confidence donut splits it into **Proven** (holdout-verified, credited at the CI lower bound), **Estimated** (calibrated estimator), and **Declared** (flat baseline). Unit economics express value per outcome against the human alternative, and per-unit verdicts (**SCALE**, **WATCH**, **FIX**, **KILL-CANDIDATE**) reduce the portfolio to the short list of decisions leadership must actually make.

FlowX.AI

FlowX.AI builds and orchestrates enterprise-grade AI agents for regulated industries, with a focus on banking, financial services, and insurance. The platform pairs LangGraph-native agent orchestration with the Observatory governance and observability plane (including its ROI Hub, which implements the VERA framework), so that institutions can deploy autonomous workflows whose value is measurable, auditable, and graded by evidence.

FlowX.AI Technical Paper · VERA v1.1 · Bogdan Răduță, Head of Research · bogdan.raduta@flowx.ai