

**Beyond Token Costs:
Measuring the Cost of a Verified AI Outcome**

Sruthi Rachamalla

Southern Illinois University, Carbondale

ORCID iD: 0000-0001-9312-507X

Abstract

Background: A low inference bill does not establish that an AI workflow is economical. The relevant expense includes the work required to check, correct, and complete a business task.

Methods: This paper develops a conceptual accounting framework from research on agent evaluation and guidance on technology unit economics. It connects a fixed cohort of requests to execution records, human effort, shared costs, and independently checked outcomes.

Results: The proposed measure, cost per verified outcome, retains expenditure on unsuccessful and unresolved cases while counting each accepted business result once. An illustrative invoice-processing comparison shows how a workflow with higher model expenditure can have lower total cost because it requires less correction. A reporting architecture preserves the evidence behind both the numerator and denominator.

Conclusion: Procurement and deployment decisions should consider verified completion, quality, elapsed time, and cost together. The framework offers an auditable comparison method, not evidence that a particular model or workflow is cheaper. Its usefulness depends on credible verification and consistent accounting boundaries.

Keywords: AI economics; outcome verification; workflow evaluation; cost allocation; enterprise automation

1 Introduction

An inexpensive answer can become an expensive task. Someone may need to retrieve missing evidence, correct a calculation, repeat a tool call, or finish the work manually. A token invoice records only part of that process.

Cost-aware evaluation is not a new idea. Kapoor et al. argue for assessing agent accuracy and cost together [1]. Curcio's LCOAI preprint considers lifecycle expenditure per productive output [2], while the FinOps Foundation distinguishes resource-efficiency measures from business-unit measures [3]. This paper applies those principles to a narrower question: what does it cost to deliver one independently verified business outcome? Its contribution is a practical accounting boundary that connects attempts, human intervention, and unresolved work to the same request.

2 Methods

2.1 Define the outcome before measuring cost

The method is a design synthesis, supported by a constructed example rather than a deployment experiment. Begin with a request type, acceptance rules, and a fixed observation window. For invoice processing, success might require a correct, authorized accounting entry reconciled to source records without duplication. Correctly rejecting an invalid invoice can count only when rejection is a predefined acceptable outcome.

Verification must inspect business evidence, not merely accept the assistant's completion message. Anthropic's evaluation guidance similarly separates an agent's transcript from the resulting state of its environment [4]. Automated checks can verify structured conditions; qualified reviewers resolve ambiguity. Human-rescued cases may count as completed, but must be labeled separately from completion without human intervention.

2.2 Account for the entire request cohort

This request-level measure applies established unit-economics principles [2], [3], using verified business outcomes rather than individual inferences as the denominator.

Let N_{verified} be the number of distinct requests meeting the acceptance rules by the cohort's cutoff.

Define:

$$\text{CPVO} = (C_{\text{run}} + C_{\text{people}} + C_{\text{shared}} + C_{\text{loss}}) / N_{\text{verified}} \quad (1)$$

C_{run} includes model calls, retrieval, tools, infrastructure, and automated checks across every attempt. C_{people} records review, correction, escalation, and manual completion using measured time and documented loaded labor rates. C_{shared} allocates setup and ongoing monitoring, security, and evaluation costs; document allocation rules and the amortization period for setup. C_{loss} records realized error losses net of recoveries; uncertain exposure is reported separately. Valid invoice payments are business transfers, not processing expenses. Keep cost categories mutually exclusive and reconcile them to billing and time records.

Failed, abandoned, and unresolved requests retain their incurred costs but contribute no verified outcome. Retries retain the original request identifier. Follow comparable cohorts for equal periods; report pending cases and later restatements instead of quietly moving unfinished work elsewhere. When no outcome is verified, report CPVO as undefined alongside total expenditure.

2.3 Preserve an auditable cost trail

Figure 1 separates the business workflow from its measurement record. Each request links to attempt identifiers, model and policy versions, resource usage, reviewer time, outcome evidence, and verification status. The outcome ledger joins these records before aggregation. Reopened cases keep their identity, accumulate additional costs, and trigger a versioned status correction. Access controls protect sensitive evidence; summaries need not expose source documents.

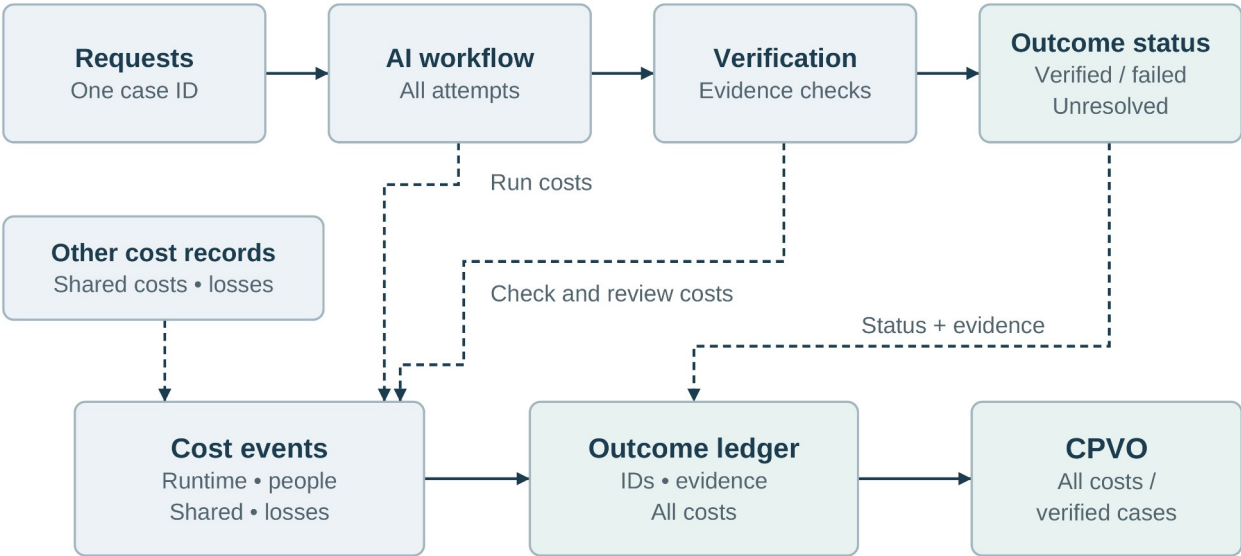


Figure 1. Request-level evidence connects workflow costs to verified outcomes.

3 Results

Consider two workflows handling equivalent cohorts of 1,000 invoice cases over the same observation window. Workflow A uses lower-cost inference; Workflow B spends more on models and automated validation. Both apply identical acceptance rules. Every amount in Table 1 is an assumption, not a measured result. Runtime costs cover infrastructure, repeated attempts, and automated checks. Human work costs USD 50 per hour: 12 hours for A and 4.8 hours for B. These hours cover aggregate effort across each cohort, not a review of every case. Neither workflow incurs a realized error loss.

Table 1. Illustrative cohort comparison; monetary amounts are USD.

Measure	Workflow A	Workflow B
Runtime costs	80	240
Human work	600	240
Allocated shared costs	120	120
Realized error losses	0	0
Total cost	800	600
Verified / failed / unresolved	900 / 80 / 20	960 / 25 / 15
Verified completion rate	90%	96%
Cost per verified outcome	0.89	0.63

A's smaller technology bill conceals greater human effort. Its total cost is USD 800 for 900 verified cases, compared with USD 600 for 960 under B. Using unrounded ratios, B's cost per verified outcome is about 30% lower. This arithmetic illustrates a possible reversal in rankings; it does not demonstrate that higher-priced models cause better results.

The difference is sensitive to review effort. With B's completion count and other costs unchanged, B reaches A's unit cost at approximately 9.9 human hours. Above that level, the apparent advantage disappears. Procurement should therefore test review-time assumptions, not simply negotiate token rates.

4 Discussion

A lower CPVO is not sufficient grounds for deployment. Report completion rate, unresolved demand, serious errors, human-rescue frequency, and end-to-end latency alongside it. Otherwise, a system can appear efficient by accepting easy work, weakening verification, or delaying

difficult cases. NIST's generative AI guidance supports evaluating risks throughout the system lifecycle [5].

A pilot should compare equivalent case mixes under fixed acceptance rules, preferably with randomized assignment. Independently audit accepted cases and count the audit cost. Report uncertainty and results by complexity; averages can conceal expensive exceptions. Report the proportion receiving human intervention and time per intervention, separating review, correction, escalation, and manual completion. Verification errors can reverse the ranking.

Start with one bounded workflow. Operations defines acceptable outcomes, engineering supplies trace records, and finance approves cost allocations. Reconcile a sample manually before relying on the dashboard; reassess after material model, staffing, or policy changes.

Report marginal and fully allocated costs separately. The example establishes neither return on investment nor cash savings: it lacks a manual baseline and benefit valuation. Released employee time is capacity, not a saving, unless spending is actually avoided or reduced.

5 Conclusion

Token prices remain useful engineering signals, but they cannot establish the economics of a completed business task. A defensible comparison follows each request through verification, counts unsuccessful work, and makes human effort visible. The preferred workflow is the one that meets required quality and service conditions at a supportable total cost, not necessarily the one producing the cheapest response.

References

- [1] S. Kapoor, B. Stroebel, Z. S. Siegel, N. Nadgir, and A. Narayanan, "AI agents that matter," arXiv:2407.01502, 2024. doi: 10.48550/arXiv.2407.01502.
- [2] E. Curcio, "Introducing LCOAI: A standardized economic metric for evaluating AI deployment costs," arXiv:2509.02596, 2025. doi: 10.48550/arXiv.2509.02596.
- [3] FinOps Foundation, "Unit economics," FinOps Framework. Accessed: Oct. 27, 2026. [Online]. Available: <https://www.finops.org/framework/capabilities/unit-economics/>
- [4] Anthropic, "Demystifying evals for AI agents," Jan. 9, 2026. [Online]. Available: <https://www.anthropic.com/research/demystifying-evals-for-ai-agents>
- [5] C. Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, Jul. 2024. doi: 10.6028/NIST.AI.600-1.