The Economics of Unified AI Operations: Workflow Costs, Coordination Overhead, and Enterprise Value Attribution

November 2026

Abstract

The cost of an AI workflow extends beyond model usage to implementation, supervision, exception handling, maintenance, and unresolved work. This paper develops an economic evaluation framework for enterprise AI operating systems using verified business completion as the unit of analysis. It distinguishes released staff capacity from realised financial benefit and separates recurring operating costs from investment in implementation. The analysis draws on research concerning AI productivity, complementary organisational investment, and agent evaluation. A constructed numerical example demonstrates how review effort, completion rates, and allocation of setup costs affect unit economics. The paper proposes a measurement protocol for attributing outcomes while accounting for demand, process changes, and selection effects. MatruOS provides a deployment context, without implying measured commercial performance. The contribution is an auditable cost and benefit account at workflow level. No return on investment forecast or observed deployment result is reported.

Keywords: AI economics, workflow costing, value attribution, human review, implementation investment, enterprise productivity.

1. Choosing an economic unit that represents work

A model call has a measurable price, but it is not a business outcome. A quotation can require several calls, database queries, and approvals. Some quotations are abandoned, corrected, or repeated. A cost calculation that divides model expenditure by the number of generated documents can omit much of the work required to produce valid commitments.

The appropriate unit depends on the process. For order administration, it might be an order correctly recorded and reconciled. For support, it might be an issue resolved under an agreed quality definition. For procurement, it might be a valid purchase commitment with the required approvals. The unit should be defined before measurement begins.

This paper asks how an enterprise should assess the economics of an AI operating system when several functions contribute to one outcome. The focus is operational evaluation rather than pricing a software product. The framework applies to decisions about pilot scope, continued operation, and expansion of delegated responsibility.

A central distinction is between capacity released and money realised. An employee who spends fewer minutes preparing a document may use that time for other work, absorb previously unmanaged demand, or remain constrained by another part of the process. The financial consequence requires observation rather than automatic conversion of every minute into a cash saving.

2. Research basis and analytical method

Brynjolfsson, Li, and Raymond study the introduction of a generative AI assistant in customer support and report effects that differ across workers. Their findings demonstrate the importance of a defined task setting, measured outcomes, and variation among users. They do not establish a universal productivity coefficient for enterprise AI operating systems. [1]

Brynjolfsson, Rock, and Syverson examine complementary intangible investment and its implications for measured productivity. Their work supports including organisational change and implementation effort when interpreting the timing of benefits. An early operating result should be considered alongside the investment required to make it possible. [2]

Kapoor and colleagues identify limitations in agent evaluation when accuracy receives attention without corresponding cost analysis, alongside concerns about holdout data and reproducibility. This supports evaluating the joint relationship between valid outcomes and resources consumed. [3]

The present paper develops a costing model and applies it to a constructed numerical example. All amounts, volumes, times, and outcomes in that example are assumptions for analysis. They are not MatruOS results, customer observations, or market benchmarks. The proposed contribution is a method for keeping the assumptions visible and testing them against deployment evidence.

3. Defining the workflow boundary

The workflow boundary should include the activities required to reach the selected outcome. If a quotation requires stock verification and credit approval, those activities belong in the measured process. Moving them to another employee does not remove their cost.

The boundary also needs a time horizon. Some errors become visible only after invoicing, fulfilment, or reconciliation. A measurement period ending at document generation may miss corrections that occur later. The evaluation should state how long outcomes remain open to review and how late corrections are attributed.

The denominator should include eligible attempts and verified completions as separate counts. Eligible attempts represent demand presented to the workflow. Verified completions represent outcomes satisfying the defined conditions. Rejected requests, abandoned cases, and unresolved work should remain visible rather than being discarded from the analysis.

A stable definition prevents apparent improvement caused by changing which tasks are counted. If a pilot excludes complex customers, its unit cost should not be presented as representative of all customers. The report should identify the eligible population and the characteristics of excluded work.

4. A complete cost account

4.1 Direct operating costs

Direct costs include model usage, retrieval, compute, storage, external service charges, and workflow-specific software charges. These should be measured from actual usage where possible. Estimated allocation should identify the basis used, such as transactions, storage consumption, or processing time.

Model usage should include unsuccessful attempts, retries, and verification calls. A low-cost successful run does not represent the average workflow if many runs require several attempts. Background summarisation

and maintenance jobs also consume resources and should be included where they support the measured process.

Shared platform costs require an allocation rule. The purpose is not to create artificial precision but to avoid assigning every shared expense to whichever workflow is being reviewed. The rule should be consistent across periods and disclosed alongside the result.

4.2 Human effort

Human effort includes routine review, exception resolution, policy maintenance, incident investigation, and support. Time should be recorded by activity rather than inferred only from the system’s response duration. An employee can spend substantial time checking an answer after the model finishes.

A loaded labour rate can translate time into an economic cost for comparison, but it should not be presented automatically as a cash expense avoided. If staffing and paid hours remain unchanged, released time represents capacity that must be deployed usefully before a financial effect can be claimed.

Review effort may vary by employee experience and task difficulty. Averages should therefore be accompanied by the distribution across important categories. A pilot relying on unusually experienced reviewers may understate the support required in wider deployment.

4.3 Implementation and maintenance investment

Implementation includes data preparation, integration, testing, training, and development of business rules. It may also include time spent establishing ownership of records and resolving conflicting processes. These activities can be necessary even when the software itself is readily available.

For an economic comparison, setup cost can be allocated over a stated evaluation horizon. The horizon is an analytical assumption, not an accounting policy. A separate cash flow view should show when expenditure is actually incurred.

Maintenance includes model changes, connector changes, policy updates, and ongoing evaluation. A design that works at initial deployment may need adjustment as the business changes. Excluding maintenance creates an incomplete account of sustained operation.

4.4 Errors and unresolved obligations

Error costs include correction effort, duplicate actions, customer remediation, and other observed consequences. Some incidents are infrequent but severe. A short pilot may not provide enough evidence to estimate their expected cost reliably.

The report should distinguish observed losses from scenario analysis of possible losses. Inventing a small expected loss number can create false confidence. Where evidence is insufficient, the evaluation should state the uncertainty and examine sensitivity to plausible incident costs chosen by responsible stakeholders.

Unresolved work also consumes resources. Stock may remain reserved, payments may remain uncertain, and employees may need to monitor pending cases. These costs should be attributed to the workflow that created the obligation even if resolution falls outside the initial measurement period.

5. A constructed numerical example

Assume a monthly workload of 1,000 eligible order administration cases. In the baseline process, each case consumes twelve minutes of staff time. At an assumed loaded rate of INR 300 per hour, that represents 200 hours and INR 60,000 of labour capacity. Assume baseline software allocation is INR 5,000 and observed correction cost is INR 3,000. The total economic cost is INR 68,000.

Assume 950 cases reach verified completion under the chosen definition. The baseline cost per verified completion is therefore INR 68,000 divided by 950, or approximately INR 71.58. The remaining fifty cases should be reported by outcome, including unresolved and rejected work. Their effort is already included in the total cost.

For the proposed AI workflow, assume all 1,000 cases receive four minutes of routine review. Assume 150 cases need an additional ten minutes of exception handling. Routine review uses approximately 66.67 hours, and exceptions add 25 hours. At the same loaded rate, total review effort costs INR 27,500.

Assume model and infrastructure cost is INR 8,000, platform allocation is INR 7,000, operational maintenance is INR 6,000, and correction cost is INR 4,000. These categories total INR 52,500 including review. Assume a setup investment of INR 120,000 allocated over twelve months, adding INR 10,000 per month to the economic comparison.

The resulting monthly economic cost is INR 62,500. If 970 cases reach verified completion, cost per verified completion is approximately INR 64.43. The calculation shows what follows from the stated assumptions. It does not establish that a real deployment will achieve these times, costs, or outcomes.

The assumed monthly difference is INR 5,500 after the setup allocation. Staff effort falls from 200 hours to approximately 91.67 hours, releasing about 108.33 hours of capacity. That time has an economic valuation, but its cash effect depends on what the organisation does with it and whether other costs change.

The initial month also requires the setup payment. Treating the allocation as if it were the actual cash timing would misrepresent the investment. An enterprise should inspect both the recurring operating account and the cash flow schedule before making a deployment decision.

5.1 Sensitivity to review effort

Hold the other assumptions constant. One additional minute of review across 1,000 cases adds approximately 16.67 hours, or INR 5,000. Increasing routine review from four to five minutes would bring the monthly economic cost to INR 67,500. This almost removes the assumed difference from the baseline total.

The example illustrates why review effort requires measurement. A small error in estimating routine supervision can materially change the economic conclusion. Model expenditure may be easy to observe while review effort remains hidden in normal staff activity.

5.2 Sensitivity to verified completion

If the AI workflow’s total cost remains INR 62,500 but only 850 cases are verified as complete, the unit cost becomes approximately INR 73.53. A low total operating cost can coexist with an unacceptable completion rate or a growing backlog.

The evaluation should therefore report cost and outcome together. It should also examine the reason for incomplete work. A valid rejection of an unauthorised request is different from an abandoned order, even though neither may count as completed under a particular definition.

6. Measuring benefit without double counting

Benefits can include additional valid throughput, reduced correction work, released staff capacity, or changes in working capital timing. Each should be connected to a defined observation. The same effect should not be counted twice under different labels.

For example, if released staff time is used to process additional orders, valuing both all released time as a saving and the full incremental contribution from those orders can overstate benefit. The analysis should identify the capacity used, incremental costs incurred, and contribution attributable to the additional work.

Revenue is also distinct from contribution. Additional sales may require product costs, delivery, support, and collection. A workflow that increases order volume without considering those costs does not establish an equivalent increase in enterprise value.

A reduction in elapsed process time may have operational value without a direct financial effect. Its relevance depends on the business constraint. If delivery capacity remains unchanged, preparing orders earlier may improve visibility while leaving shipment volume unchanged. The report should describe the observed consequence rather than assign a speculative earnings value.

7. Attribution and experimental design

A pilot should establish a baseline before deployment, using the same definitions of eligibility, completion, and cost. Where practical, a controlled comparison can assign similar cases to alternative processes. If random assignment is unsuitable, the analysis should document how groups differ and how those differences affect interpretation.

Seasonality, customer mix, staff experience, and process redesign can influence the result. An AI deployment often occurs alongside data cleanup or policy clarification. These complementary changes should be recorded because they may explain part of the outcome.

The first research question is whether valid completion and total resource use change under a defined workflow configuration. A second is how much review and maintenance effort is required after the initial learning period. A third is whether the released capacity is actually used for a measurable business purpose.

A pilot should retain cases that fail or require manual takeover. Excluding them creates selection bias. The analysis should report both the assigned workflow and the actual handling path, including the cost of moving work back to staff.

Measurement should continue long enough to observe ordinary variation and corrections. The appropriate period depends on transaction frequency and the delay before consequences appear. A successful demonstration provides useful technical evidence but little basis for estimating recurring economics.

8. Application to MatruOS

A MatruOS evaluation should connect each workflow to usage records, review activity, exceptions, and verified outcomes. The platform should allow costs to be attributed without exposing unnecessary customer information. Operational reports should show incomplete work alongside completed work.

The proposed framework also informs model selection. A model should be assessed within the complete workflow, including its effect on review and correction. A low inference price can be offset by additional

human work, while additional reasoning cost may be justified only if it produces a measured operational benefit.

Capacity planning should include concurrent workflows and shared services. A result from one lightly used workflow may not describe the cost or delay when several business functions operate together. The evaluation should record workload and resource constraints so that expansion assumptions are reviewable.

This paper offers no MatruOS pricing claim or return forecast. Its purpose is to establish the evidence needed for those decisions. The same measurement requirements should apply to any platform being considered for delegated business execution.

9. Limitations and conclusion

Economic evaluation depends on the quality of time records, cost allocation, and outcome definitions. Some benefits, such as improved understanding of operations, may be difficult to monetise. They can be reported as observed qualitative outcomes without assigning unsupported financial values.

Rare incidents create additional uncertainty. A short period with no severe failure does not establish negligible risk cost. Scenario analysis can support decisions, provided its assumptions remain separate from observed results.

The economics of an AI operating system should be assessed through complete, verified business workflows. Model usage is one component of a wider account that includes implementation, supervision, correction, and maintenance. Released staff capacity and realised financial benefit should remain distinct until their relationship is observed.

The proposed framework provides an auditable basis for pilot decisions and subsequent expansion. Its practical value lies in making the denominator, cost boundary, assumptions, and attribution method explicit before an enterprise interprets a deployment as financially successful.

Author affiliation and disclosure

The author is affiliated with Code Colonies Private Limited, Ahmedabad, and leads the MatruOS team. The numerical example is hypothetical and is not a product forecast or customer result.

References

[1] Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at Work. The Quarterly Journal of Economics, 140(2), 889-942.

https://doi.org/10.1093/qje/qjae044

[2] Brynjolfsson, E., Rock, D., and Syverson, C. (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. American Economic Journal: Macroeconomics, 13(1), 333-372.

https://doi.org/10.1257/mac.20180386

[3] Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. (2024). AI Agents That Matter. arXiv:2407.01502.

https://arxiv.org/abs/2407.01502