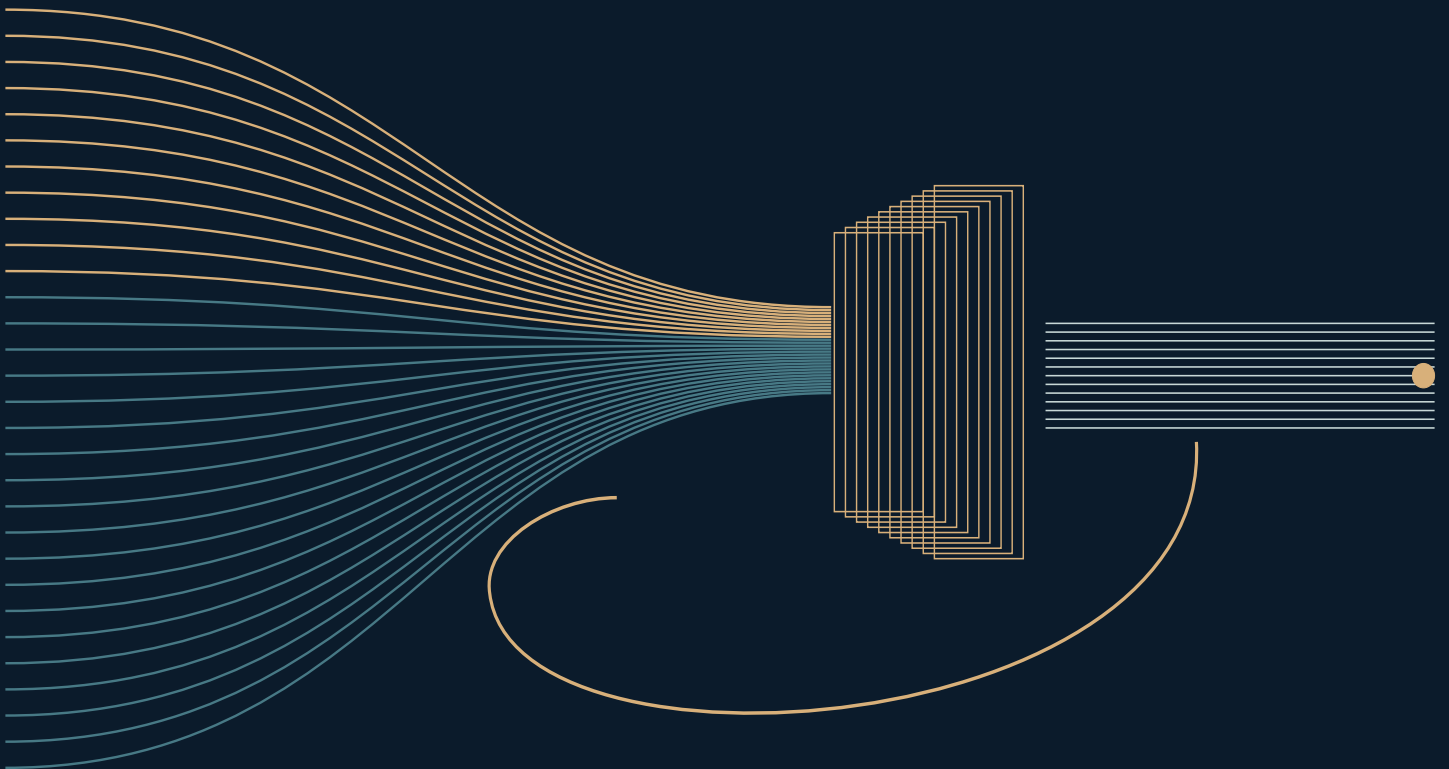


STRATEGY / OCTOBER 2026

Price the accepted outcome

The economics of AI beyond tokens, seats and headline savings



THE DECISION IN BRIEF

Price the complete accepted outcome, including failed attempts, review, recovery and operating capacity; then choose the configuration that meets the required standard at a sustainable cost.

- 01

The model invoice is one component of the economics. A cheaper response can require more review and rework, while a higher-priced model may or may not reduce the cost of accepted work.
- 02

Use the same completion standard, coverage and period for each alternative. Retain failed attempts in total cost, and distinguish marginal decisions from full-service investment decisions.
- 03

Test acceptance, review load, volume and supplier dependence together. Release capacity only becomes cash savings when expenditure changes; quality and coverage constraints remain separate from the cost ratio.

THE MAYDEN VIEW

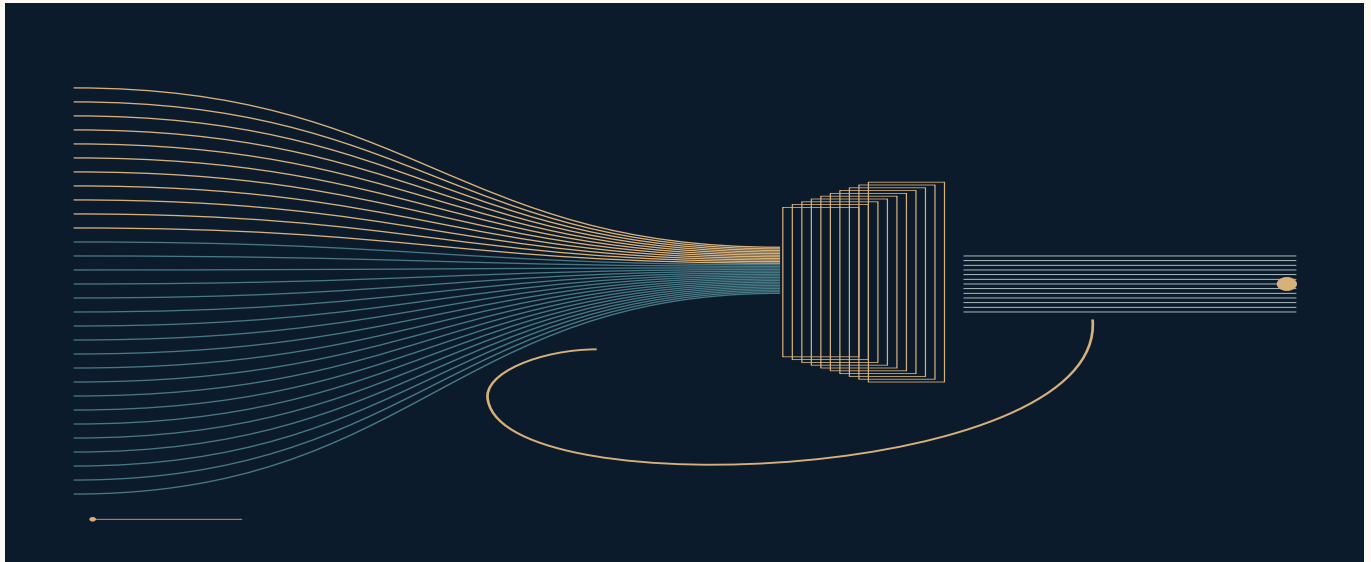
The relevant price is what the business pays for work it can accept.

INSIDE THIS PERSPECTIVE

The argument at a glance	3
01 / Define the cost of accepted work	4
02 / Test the decision boundary	5-6
Exhibit 01 / Compare the cost of an accepted outcome	7
03 / Include capacity and transition	8
Exhibit 02 / Reconcile the numerator and test the break-even boundary	9
04 / Fund a defensible outcome	10
The leadership agenda	11
Method and references	12

ONE DECISION, FOUR CONNECTED QUESTIONS

Cost meets acceptance



ACCEPTED OUTCOME ACCOUNT / CONCEPTUAL ARTWORK

01 Define the cost of accepted work

Choose the complete arrangement, not the cheapest response

02 Test the decision boundary

A simple sensitivity test exposes the hidden economics

03 Include capacity and transition

Separate transition costs from the cost of continuing to operate

04 Fund a defensible outcome

Cost per outcome has limits that leaders should respect

The relevant price is what the business pays for work it can accept.

CHAPTER 01

01

Define the cost of accepted work

Choose the complete arrangement, not the cheapest response

The cheapest model can produce the most expensive accepted outcome. Its output may need repeated attempts, expert review and correction after use. A more capable model may reduce that burden - or merely add expense. Compare the complete operating arrangement against the same acceptance standard before choosing a component.

The FinOps Foundation distinguishes resource-efficiency measures from business-unit measures. ^[1] That distinction is useful for AI: cost per token helps engineers manage a component, while cost per resolved case helps an operating leader judge a service. The measures should connect without being confused. A system can improve its token efficiency while producing more failed attempts or longer reviews. Optimizing the first ratio does not establish improvement in the second.

Define completion first. A draft needs an accountable recipient's acceptance; a service case may need a reopening window; a transaction may need reconciliation. Apply the same rule to the current and proposed process. A looser completion standard can create apparent savings by buying a worse service.

Reconcile the complete cost with accepted work

mayden's Accepted Outcome Account groups cost into execution, human effort, recovery and shared operating capacity. Include model calls and retrieval; preparation, review and correction; failure handling; and monitoring, evaluation, support and infrastructure. Divide costs by accepted outcomes from the same scope and period. Failed attempts remain in the numerator.

Specify the decision the account supports. A routing choice may depend on marginal cost; building a service requires transition investment and sustained capacity. Finance should identify incremental, allocated and sunk costs. Show the relevant views together rather than selecting the one that makes the proposal look cheapest.

Use a case identifier to connect attempts, versions, human interventions and acceptance without unnecessary personal data. The operating owner can then explain changes in unit cost through actual cases: different demand, harder work or a deteriorating configuration. An aggregate model bill cannot resolve those questions.

CHAPTER 02

02

Test the decision boundary

A simple sensitivity test exposes the hidden economics

Consider two wholly illustrative ways to process 100 cases. Configuration A incurs 10 monetary units in model cost, 500 in review, 40 in recovery and 50 in shared operation: 600 divided by 80 accepted outcomes equals 7.50 each. B incurs 40, 200, 10 and 50: 300 divided by 95 equals 3.1579, rounded to 3.16. Exhibit 01 foregrounds this outcome metric; Exhibit 02 reconciles its inputs. These are invented scenario values, not prices, client data or a benchmark. All attempts stay in the numerator; accepted outcomes use the same service boundary, quality standard and observation window.

Hold A's illustrative 7.50 outcome cost fixed. For a positive number of B accepted outcomes, equality occurs when B total cost = 7.50 multiplied by B accepted outcomes. B is cheaper only below that boundary. At a total cost of 300, 40 accepted outcomes produce exactly 7.50,

not an advantage; more than 40 produce a lower unit cost if all other conditions hold. At 95 accepted outcomes, 712.50 is the total-cost ceiling for equality. These are one-variable break-even calculations around fixed assumptions, not a recommendation to accept only 40% of cases. Zero accepted outcomes leaves cost per accepted outcome undefined.

The admissible choice must satisfy the quality and coverage conditions before this cost comparison is used. A cheaper ratio cannot compensate for prohibited errors or silently exclude difficult cases. If the higher model price fails to change acceptance or effort, B's assumed advantage may disappear. If review is a fixed staffing commitment, an effort reduction is not immediate cash savings. Test connected changes such as lower acceptance and longer review jointly; volume may spread fixed costs while overloading exceptions. The boundary identifies what evidence changes the decision, not a forecast that the illustrative inputs will occur.

CHAPTER 02

02

Test the decision boundary

Model quality and cost must be evaluated jointly

Research on AI agents identifies the limitations of accuracy-only benchmarks and argues for assessing cost alongside performance. ^[2] For a company, the corresponding decision is to compare configurations on its own accepted work. A high benchmark score does not establish that a larger model is economical for routine cases. A cheap model does not establish that escalation and correction costs are manageable. The cost-performance trade-off is specific to the workload and its acceptance standard.

A June 2026 study compared online memory, workflow and skill modules with a simpler web agent given an approximately matched total token budget. Across the tested models and three WebArena domains, the simpler baseline matched or exceeded the augmented methods in aggregate success. ^[8] The study used step limits as an approximate budget control, rather than exact per-task token matching. The result does not rule out reusable skills; it shows why their creation, retrieval and context costs belong inside the comparison. Ask whether added architecture earns its cost against a properly

resourced simpler alternative. The simpler baseline also uses accessibility-tree pruning; this is a comparison of tested configurations, not isolation of a universal causal penalty from memory.

Routing can improve economics when cases can be distinguished reliably. A simpler model may handle a bounded routine category, with a more capable system or person handling the remainder. But routing creates its own failure mode: a difficult case can be classified as easy and receive insufficient attention. Evaluate the complete routing policy, including false confidence and escalation capacity. Savings attributed to the cheap route are incomplete if the costly mistakes are booked elsewhere.

Historical inference-price declines are relevant but insufficient. Stanford's 2025 AI Index tracks sharply lower prices at a fixed benchmark capability level over a defined historical period. ^[3] It does not establish that a particular enterprise workload will become cheaper at the same rate. More demanding outputs, larger context and higher usage can change total expenditure. Budget from measured consumption and explicit scenarios, while retaining the ability to benefit from better components when they arrive.

ILLUSTRATIVE SCENARIO · HYPOTHETICAL INPUTS

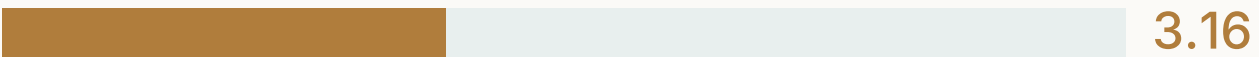
Compare the cost of an accepted outcome

Hypothetical monetary units per accepted outcome

Configuration A



Configuration B



THE MANAGEMENT IMPLICATION

Choose only among quality-compliant configurations; lower component price does not determine outcome cost.

Wholly illustrative: A: $600 / 80 = 7.50$; B: $300 / 95 = 3.1579$, rounded to 3.16. Each attempts 100 cases. Same acceptance, quality, coverage and observation window; failed attempts stay in total cost. Transition investment excluded; not cash savings. Source: mayden analysis. Proposed framework or illustrative scenario.

CHAPTER 03

03

Include capacity and transition

Separate transition costs from the cost of continuing to operate

Separate three accounts: experiments buy information, transition changes the process, and steady operation delivers recurring outcomes. A small experiment can be valuable despite a high cost per case. A mature service can look artificially cheap if maintenance staff remain charged to the project. Make the transfer of cost and ownership explicit when the service becomes an operating dependency.

FinOps guidance for AI emphasizes allocation, forecasting and the distinctive variability of AI spending. ^[4] mayden's application is to budget a baseline load, a plausible growth case and a constrained case in which acceptance or supply deteriorates. Each should include the human support needed to keep the service usable. A spending cap alone can protect a budget while causing abrupt service failure. The operating response to the cap belongs in the design.

Normalize supplier billing where possible, but retain business context. FOCUS provides an open specification for cost and usage data. ^[5] A normalized invoice can make comparison easier; it cannot decide whether an output was useful or which customer outcome it supported. Link financial records to an independently defined outcome record. Otherwise the organization becomes precise about consumption while remaining uncertain about the value purchased by that consumption.

Economics change when demand and capacity respond

Lower effort can attract more demand. Rising expenditure may therefore accompany valuable expansion. Test whether additional accepted outcomes justify their incremental cost and whether shared capacity can absorb them. More generated material is not evidence of value when nobody uses it.

Physical and supplier constraints also matter. The IEA's Energy and AI report examines data-centre electricity demand and local infrastructure pressures. ^[6] The company-level implication is conditional: deployments that depend on particular hosting capacity or locations should test availability and continuity assumptions. The report's global projections should not be converted into a surcharge on an individual workload without relevant evidence. Aggregate energy trends inform scenario design; actual contracts and measurements determine the account.

Keep an exit view beside the operating view. The CMA's market review highlights the relevance of choice and dependence in foundation-model markets. ^[7] A service can have attractive current unit economics while accumulating expensive switching requirements. Record what must be migrated, re-evaluated or renegotiated if the supplier changes. The additional flexibility may or may not be worth buying today, but omitting the question understates the strategic commitment embedded in the unit price.

ILLUSTRATIVE SCENARIO · HYPOTHETICAL INPUTS

Reconcile the numerator and test the break-even boundary

Hypothetical monetary units

Illustrative input	Configuration A	Configuration B
Cases attempted	100	100
Model cost	10	40
Human review	500	200
Recovery	40	10
Shared operation	50	50
Total cost	600	300
Accepted outcomes	80	95
Cost per accepted outcome	7.50	3.16

THE MANAGEMENT IMPLICATION

At cost 300, B ties at 40 accepted outcomes; at 95 accepted, equality permits 712.50 total cost. These thresholds do not relax quality or coverage.

All figures are deliberately illustrative, not measured results or vendor prices. $600/80=7.50$; $300/95=3.1579$, rounded to 3.16. Excludes transition investment; costs of all attempts remain in numerator. Equality: $B \text{ total cost} = 7.50 \times B \text{ accepted outcomes}$, with a positive denominator and fixed A baseline. Quality is a separate condition. Source: mayden analysis. Proposed framework or illustrative scenario.

CHAPTER 04

04

Fund a defensible outcome

Cost per outcome has limits that leaders should respect

Some valuable outcomes are rare, delayed or difficult to quantify. A strategic decision, a prevented incident or improved customer confidence may resist a credible per-unit calculation. Do not manufacture a monetary benefit merely to fit the account. Use explicit decision criteria, evidence of quality and bounded cost while the organization learns. A transparent qualitative judgement is more useful than a precise return based on an unsupported value for trust or avoided harm.

The metric can also encourage gaming. Teams may redefine easy cases as separate outcomes, reject difficult requests or narrow the observation window until reopening costs disappear. Preserve the case definition and report coverage alongside efficiency. A lower unit cost with a shrinking service boundary may be a rational choice, but it should be presented as such. Leadership needs to see who is served, who is excluded and where their work goes next.

Quality constraints should remain separate from expected-cost arithmetic when consequences are unacceptable. A workflow should not be approved merely because a low probability multiplied by an estimated loss looks affordable. Some obligations, customer commitments and risk limits are non-negotiable. The account supports a decision within those boundaries. It does not replace the judgement required to establish them or the assurance needed for consequential action.

Rare failures require a different treatment from routine rework. A short pilot may observe none while remaining too small to establish that the risk is negligible. Keep an explicit uncertainty register for consequential failures, including the exposure, detection route and practical recovery

limit. Do not assign a zero cost simply because the event has not yet occurred. Equally, avoid manufacturing a precise probability from an unrelated industry example. The investment decision can use bounded scenarios and a narrower release boundary until stronger evidence is available. The cost account should make this unresolved exposure visible beside the measured operating figures.

Make the configuration choice on one shared account

Ask finance and operations to compare the existing process, a simpler improvement and the proposed AI arrangement on the same accepted outcome. Show transition cost, continuing cost, quality and coverage. Name the staffing or commercial decision that converts released capacity into value. If the case depends on new demand, show who will establish it.

Technology should explain which configuration choices drive the result and how they were tested. Operations should own completion, exceptions and staffing. Procurement should surface terms that change future cost or flexibility. The account is ready for a decision when those views reconcile.

Choose the arrangement that meets the service standard at a sustainable cost. Where uncertainty remains, test the variable that could reverse the choice: acceptance, review effort, routing, volume or recovery. An expensive capability can be economical; a cheap component can be wasteful. The accepted outcome is the unit that distinguishes them.

FROM INSIGHT TO ACTION

Decisions to take forward

FINANCE

Use break-even within quality constraints

Evidence to require: Reconciled numerator, positive denominator and fixed baseline

OPERATING OWNER

Reject savings created by weakening the service

Evidence to require: Acceptance, exclusions, reopening and review effort

TECHNOLOGY LEADER

Choose capability that earns its full cost

Evidence to require: Whole-configuration comparison and simpler alternative

mayden connects senior judgement, research and engineering to turn strategic choices into working capabilities.

EVIDENCE YOU CAN INSPECT

Research notes

Method and boundaries

Narrative synthesis of eight primary research, institutional and practitioner-standard sources with original mayden management accounting analysis.

Sources cover cost-performance evaluation, unit economics, billing normalization, infrastructure and supplier choice; opened 10 October 2026.

No proprietary cost dataset or empirical ROI estimate. The numerical exhibit is explicitly hypothetical. Frameworks require workload-specific acceptance, cost and risk definitions; historical prices are not current quotations.

Primary sources and author research

[1] Unit Economics

FinOps Foundation / Living guidance; accessed October 2026. Definition; resource-efficiency and business-unit metrics. Normative guidance; not empirical proof of ROI.

[Read the source](#)

[2] AI Agents That Matter

Kapoor et al.; arXiv / 1 July 2024. Abstract. Benchmark critique does not quantify enterprise economics.

[Read the source](#)

[3] AI Index 2025: Research and Development

Stanford HAI / 2025. Inference costs. Not current prices or a forecast of total enterprise spending.

[Read the source](#)

[4] FinOps for AI

FinOps Foundation / Living guidance; accessed October 2026. Overview; allocation, forecasting and optimization. No causal or representative ROI estimate.

[Read the source](#)

[5] FinOps Open Cost and Usage Specification

FOCUS project; FinOps Foundation / Living specification; accessed October 2026. Specification overview. Billing normalization does not establish business acceptance or value.

[Read the source](#)

[6] Energy and AI

International Energy Agency / 2025. Executive summary; data-centre demand and infrastructure. Aggregate scenarios are not workload-level measurements or company-specific forecasts.

[Read the source](#)

[7] AI Foundation Models: Technical update report

Competition and Markets Authority / 16 April 2024. Choice pp 53-56; competition risks. Historical market assessment; not a finding about any current individual contract.

[Read the source](#)

[8] Are Online Skill and Memory Modules Always Worth Their Tokens? A Budget-Constrained Study of Web Agents

Hajimiri et al.; arXiv / 12 June 2026, v1. Sections 3.3, 4, 7 and 8. Limited domains and model set; online overhead only. Offline learning can amortize differently. Tokens do not equal complete enterprise cost. Budget control uses step limits and is approximate rather than exact per-task token matching.

[Read the source](#)