RJL SOLUTIONS

OPERATORS' CORNER / FINANCE AND TECHNOLOGY

AI Token Economics

Forecast AI spend by workflow. Measure accepted work. Separate capacity from planned cash benefit.

CFO · CIO · CTO

Richard White
RJL Solutions · rjlsol.com

VERSION 2 · PUBLISHED EDITION
OCTOBER 2026

Executive summary

AI spending is easier to approve than to justify. A model call has a price. A useful business outcome has a cost that also includes integration, retrieval, evaluation, human review and ongoing support. And the value of that outcome depends on whether the business changes how work gets done.

So finance, technology and business leaders need a shared economic model. Finance needs a forecast that reconciles to spending, and a clear path from capacity to cash. Technology needs quality and service constraints to optimize within. Business owners need accepted outcomes, a measurable baseline and responsibility for realizing the benefit.

My recommendation: manage AI by the cost and value of accepted work, from the business case onward.

> **Forecast by workflow, from drivers, before the build.** Demand, tokens per task, model tier and price build the forecast. It rolls up to persona, product and provider, and gets refreshed against actuals every month.

> **Limit by persona, measure by workflow.** The persona sets what a role may use. Each workflow is routed and measured on its own economics. Caps are controls, with defined behavior when they bind.

> **Decide with three cost views and an owned realization plan.** Look at fully allocated, incremental and avoidable cost side by side. Fund benefits that a named business owner has committed to convert into cash or avoided cost.

In the illustrative base case, year-one run cost is $2.52 million against a $3.0 million budget. The work releases capacity worth $14.8 million at loaded cost. The owners, though, plan to convert only $4.71 million of it into cash or avoided cost in the year. That's a benefit-cost ratio of 1.87, below the illustrative 2.0 hurdle, because benefits start months after the costs. Under a business downside (lower acceptance, more review, slower and smaller realization) the ratio falls to 0.34. Allocation choices, review effort and benefit realization change the investment decision more than token prices do.

FOR THE CFO Ask for a driver-based forecast that reconciles to invoices, cost per accepted task by workflow, fully allocated and avoidable views side by side, and a benefits plan with a named owner for each workflow.

FOR THE CIO AND CTO Meter every call by workflow, persona and product; enforce limits at a gateway; and put an economics gate in the release pipeline so cost, quality and controls are tested together.

SCOPE AND COMPANION MODEL

API-based assistants and agents. SaaS seats and self-hosted models need additional drivers. The companion Excel workbook uses illustrative assumptions; its forecast is not a record of delivered cash returns. It forecasts run cost and planned benefit by workflow; it is not an investment cash-flow, NPV or IRR model, so add implementation and transition costs for a funding decision.

---

1. Why AI spend surprises finance

Prices for a given level of model capability have fallen fast. Bills still rise. Usage grows faster than prices fall, and the way people and agents use models multiplies the tokens behind each piece of work.

What the evidence shows

> **Estimates can miss badly without cost understanding.** Gartner warned in October 2024 that organizations that don't understand how AI costs scale could make AI cost estimate errors of 500 to 1,000 percent [1]. That's a conditional warning. It isn't a typical overrun.

> **Overruns are expected.** Gartner predicts that through 2028 at least half of generative AI projects will overrun their budgeted costs, citing poor architectural choices and a lack of operational know-how [2]. It's a prediction, not a measured failure rate.

> **Finance teams are engaging.** Among respondents to the FinOps Foundation's State of FinOps surveys, the share managing AI spend rose from 31 percent (2024) to 63 percent (2025) and 98 percent (2026). FinOps for AI is the top forward-looking priority, while workload optimization remains a top current one [3][4]. These are survey respondents, not all enterprises.

> **Agents can consume far more and vary widely.** Bai and colleagues studied agentic coding tasks. They found agentic tasks consuming roughly 1,000 times the tokens of code reasoning and chat, with cost on the same task varying up to 30 times between runs [5]. That's one studied setting. It isn't a typical multiplier for all agents.

Why the bill moves

DRIVER	WHAT MOVES IT	WHO CONTROLS IT
Volume	Users, adoption, tasks per user, new workflows	Business and product owners
Tokens per task	Context, retrieved documents, history, reasoning, agent steps, retries	Engineering and product design
Model mix	Which tier each step uses; routing	Platform and architecture
Unit price	Rate card, contract, caching, batch, commitments, and provider pricing strategy	Procurement and platform
Recurring costs	Gateway, evaluation, retrieval, data upkeep, security, people, change	CIO, CTO and finance

2. Today's prices are not tomorrow's: the cloud lesson

Cloud computing taught finance a lesson that applies here. Early cloud was easy to start and cheap at small scale. As workloads grew, cloud bills became one of the biggest costs in many software businesses. In a 2021 analysis of public software companies, Andreessen Horowitz estimated committed cloud spend at about half of cost of revenue, and argued that the resulting margin pressure was worth hundreds of billions of dollars in market value: "you're crazy if you don't start in the cloud; you're crazy if you stay on it" [6].

Why AI prices may not keep falling for you

> **Contract terms can change the effective price.** Renewals, minimum commitments, model tier and context length change what a workflow costs. Treat falling unit prices as an assumption to test, rather than a permanent budget guarantee. Provider pricing documentation informs procurement; it is not a long-term commitment [7].

> **Yet prices for a given capability have fallen fast.** Epoch AI found the price of reaching a fixed level of performance fell between 9 and 900 times per year across the benchmarks it studied, while cautioning that the fastest recent declines may not persist [8].

> **Both can be true for the same customer.** The price of yesterday's capability falls. Meanwhile the capability you need, the volume you use and the terms you can negotiate move in other directions. Contract renewals, minimum commitments and the end of introductory pricing are where the change lands.

What to do about it

> Forecast at contracted rates, and test a rising-price case. The workbook's Price rise scenario raises unit prices 25 percent a year and lengthens context. Year-one benefit-cost falls from 1.87 to 1.61.

> Size commitments to a defensible floor of eligible demand. The base forecast is too optimistic a basis. Know the break-even utilization of each commitment.

> Keep workflows portable. Approve alternatives per tier, keep evaluation packs that can test a substitute model, and lock no business logic into one provider's prompts.

> Track effective price per accepted task, not list price per token. Then a provider's pricing change shows up where finance looks.

3. What a token bill is

A short primer, so finance and technology use the same words. The formula below is a planning approximation. Actual cost follows each provider's rate card and your contract.

TERM	MEANING	WHY IT MATTERS TO COST
Token	A fragment of text, roughly three quarters of an English word, counted in and out	The billing unit for most model APIs
Input tokens	Instructions, retrieved documents, history, tool results	Usually the largest volume; grows with context and agent steps
Output tokens	What the model writes, including reasoning tokens where billed as output	Priced several times higher than input
Model tier	Small, standard and frontier models at very different prices	A major lever on cost per task
Prompt caching	Reusing an unchanged prefix at a reduced input price, where eligible	Large savings when prompts share a stable prefix
Batch	Asynchronous processing at a discount, for work that can wait	Often about half price for eligible work
Calls per task	How many model calls one piece of work takes	Multiplies everything above

COST PER TASK calls per task × (1 + retry overhead) × [input tokens × blended input price + output tokens × output price] ÷ 1,000,000, less the batch discount on eligible work.

For the illustrative claim summary workflow, three calls of 12,000 input and 800 output tokens on a standard tier, with half the input cached, cost about $0.10 per task.

4. From accepted work to cash

A useful task is one the business accepts under stated criteria. Retries, failures and abandoned sessions cost money and produce nothing, so the unit is the accepted task. From there, value moves through four distinct stages. Every headline number has to say which stage it describes.

STAGE	DEFINITION	BASE CASE, YEAR 1
Accepted work	Tasks the business accepts	5.5 million useful tasks
Capacity released	Hours saved on accepted tasks, net of review (unless review is a new paid cost)	224,000 hours; $14.8 million at loaded cost
Planned benefit	The part a named owner plans to convert to cash or avoided cost, from a stated month	$4.71 million
Cash flow	Planned benefit and costs by month, with any up-front investment	Not modeled; add for an investment decision

Four kinds of benefit, kept apart

> **Capacity released:** net hours after review, rework and exceptions, bounded by the work that actually exists.

> **Cash savings or cost avoidance:** specific staffing, overtime, contractor, vendor or hiring changes, with timing and an owner. Reductions to current spend and avoided future spend are both valid; label which.

> **Incremental commercial value:** contribution after fulfillment costs, with evidence the AI changed the outcome. Revenue alone is not contribution.

> **Risk or service improvement:** measured separately unless a defensible financial conversion exists. Never count the same hours twice.

The realization plan is owned by the business and validated by finance

Finance can challenge and reconcile the assumptions. It can't make a department redeploy capacity. So for each workflow, require a short benefits contract: baseline volume and time, acceptance criteria, review and exception burden, measured change, where the released capacity goes, the cash or commercial action, the accountable owner, the timing and the evidence for the next funding gate. Keep the forecast, the measured pilot effect and the realized result separate.

The base case nets review time against capacity. Where review needs new paid resources, the workbook counts it as a cost instead, so it never gets counted twice. Saved hours have to fit the work people actually do, so the workbook checks hours per active user against a plausible maximum.

5. Three cost views for every decision

COST VIEW	QUESTION IT ANSWERS
Fully allocated	What share of the portfolio's total cost is assigned to this workflow?
Incremental	What additional cost and benefit arise if we expand it, for example by lifting a cap?
Avoidable	What cost really disappears, and when, if we stop it?

Call summaries: below the hurdle, and still worth keeping

The call summary workflow costs about $3,000 a year in tokens. Because it completes three quarters of the portfolio's tasks, a task-based allocation assigns it about $1.03 million of shared cost, putting its benefit-cost ratio at 0.93: below the illustrative stop hurdle. But only part of that shared cost would disappear if it stopped. Stopping it would remove about $189,000 of avoidable cost and lose $964,000 of planned benefit, making the portfolio $775,000 worse.

Allocation basis matters too. Allocate by token cost instead of tasks and call summaries jump to a ratio above 100, while the code change agent drops to 0.72. Neither basis is wrong. Show the sensitivity, allocate distinct cost pools by defensible drivers where you can, and decide on avoidable and incremental economics.

Code change agent: an incremental decision

The code change agent produces the most benefit per task and carries the highest cost per task; its $600 monthly cap leaves about a third of demand unmet. Lifting the cap with no other change would add about $544,000 of token cost and $826,000 of planned benefit: a 1.52 incremental ratio. That is below the illustrative 2.0 hurdle, but it is an improvement, not zero. The better first step may be to route routine steps to a smaller tier and trim context, then reconsider the cap.

RULE Use hurdles as an executive discussion aid. Falling below a hurdle doesn't prove that stopping helps. Clearing it isn't enough authorization to scale. Check quality evidence, the realization plan, incremental and avoidable economics, service obligations and the exit path.

6. Budget, forecast and actuals

Showback reports what was spent after the month closes. It's a useful first step and a weak control. Keep three numbers distinct: the approved budget (frozen), the current forecast (updated with drivers) and actuals reconciled to invoices and the ledger. A forecast informs funding. It doesn't authorize it by itself.

STAGE	WHAT FINANCE GETS	WHAT IT CANNOT DO
Showback	Spend by team after the month	Separate growth from waste, or predict next quarter
Chargeback	Accountability in the consuming budget	Show whether the spend paid off
Driver-based forecast	Expected spend by workflow and month	Prove value without measured drivers
Forecast with actuals and unit economics	Spend, accepted output, realized benefit and an explained variance	This is the target state

Explain variance by driver

The workbook's Actuals tab compares monthly actuals with the frozen budget and the current forecast. It splits variance into a volume effect (more or fewer accepted tasks than forecast) and a rate and efficiency effect (prices, tokens per task, mix and recurring cost). It's a simple, documented bridge. It isn't a full mix analysis.

RULE Investigate rising tokens per task against accepted completion, review effort, latency and business value. Fund extra consumption when it improves outcome economics within approved controls. Correct consumption that adds no matching value. Treat higher volume the same way: is it incremental accepted work with incremental value, or duplicate work and retries?

7. Limit by persona, forecast by workflow

A persona is a useful permission boundary. It defines the data, tools and actions a role may use. It isn't always the right economic unit, though. A finance analyst reconciles accounts and drafts commentary, and those workflows need different models, with different acceptance criteria and failure costs. Forecast and route by workflow, then roll up to persona, product, team and provider.

ELEMENT	SET BY PERSONA (AUTHORIZATION)	SET BY WORKFLOW (ECONOMICS)
Model	Tiers the role may use	Tier routed per step, within that permission
Data and tools	Products, indexes and tools allowed; read or write	Which of those the workflow calls
Context and calls	Upper limits	Measured tokens and calls per task
Spend	Monthly cap per user, with alerts	Forecast, cost per accepted task, unmet demand
When a limit binds	Escalation route and who can raise it	Fallback: priority queue, manual path or deferral, measured

Caps are controls. They prove nothing about value or safety. When a cap binds, record whether the work was delayed, done by hand, abandoned or moved to another tool, so cost doesn't quietly leave the measured system. Access controls, data isolation and approval for consequential actions stay separate from token limits.

8. An economics gate in the release pipeline

A model, prompt or retrieval change can alter cost without changing the business-facing application. Put economics into the release decision, alongside quality and controls.

> Version the workflow, prompt, model, tools, retrieval configuration and pricing assumptions.

> Evaluate representative tasks against business acceptance criteria, including difficult cases and failure paths.

> Measure cost per accepted task, review time, latency, retry rate and tail cost; reconcile evaluator judgments with real business acceptance.

> Compare with the existing workflow and with a simpler automation or search solution. Promote only when the quality, economic and risk criteria are met.

> Release to a small share of traffic, monitor task-level outcomes and spend, and roll back when the release exceeds agreed tolerances.

Budget for the evaluation and maintenance work this needs. Record release versions with telemetry, so the monthly variance review can tie cost movements to actual changes. A more capable model may cost more per attempt and less per accepted outcome. The gate is how you find out.

9. The forecast model

The companion workbook is an illustrative planning model, simple enough for finance to audit. Every input is a labeled cell, and every output is a formula. Its checks tab flags out-of-range inputs, unallocated cost, zero accepted output and implausible hours before anyone relies on the numbers.

TAB	PURPOSE
Dashboard	Run cost, capacity value, planned benefit, budget comparison, decisions by workflow, persona rollup, charts
Assumptions	Scenario, allocation basis, budget, hurdles, first forecast month, plausibility limit
Prices	Provider, model, effective date, currency and contracted rates by tier; caching and batch terms
Workflows	Demand, tokens, limits, quality, value drivers, realization share and start month, benefit owner
Fixed costs	Recurring cost lines, timing, and the share avoidable if one workflow stops
Forecast	24 months by workflow: demand, cap effect, accepted tasks, cost, capacity, planned benefit
ROI	Fully allocated, avoidable and incremental views by workflow and year, with the hurdle decision
Actuals	Frozen budget, current forecast, actuals and a volume and rate variance bridge
Checks	Input and reconciliation checks, summarized on the dashboard

9. The forecast model (continued)

Scenarios and what they show

YEAR-ONE CASE	RUN COST	PLANNED BENEFIT	BENEFIT / COST
Low (less adoption and volume)	$1.99M	$3.05M	1.53
Base	$2.52M	$4.71M	1.87
High (more adoption, volume, tokens)	$3.00M	$5.88M	1.96
Price rise (+25% a year, longer context)	$2.57M	$4.13M	1.61
Business downside (lower acceptance, more review, less and later realization, higher recurring cost)	$2.57M	$0.87M	0.34

Run cost barely moves across scenarios; planned benefit moves a great deal. Benefit delivery and review burden matter more to the investment conclusion than token prices. These are deterministic scenarios, not probabilities.

How to calibrate it

> **Frame:** fill the workflow rows from the baseline and planned scope; mark each value as an estimate with an owner and date.

> **Prove:** replace tokens, calls, acceptance and review minutes with measured values from the pilot on representative work; reconcile pilot invoices.

> **Run:** record actuals monthly, keep the approved budget frozen, and explain variance by driver.

10. Dashboards by role

Each role needs a different view of the same numbers, built on one definition of an accepted task and one allocation policy. The exhibits come from the workbook's base case.

CFO: run cost, capacity and planned benefit

CFO: spending, capacity and planned benefit

Year 1 · Base scenario · Illustrative forecast · Workbook version 2

Run cost, year 1	Capacity value	Planned benefit
$2.52M	**$14.81M**	**$4.71M**
Fully allocated	Hours at loaded cost; not cash	Owned conversion to cash or avoided cost

Benefit / cost	Net planned return	Budget headroom
1.87	**87%**	**$482k**
Ratio; illustrative scale hurdle 2.00	Planned benefit less run cost, divided by run cost	Against a $3M illustrative budget

Workflow decisions: compare allocation with stopping

Workflow	Benefit / cost	Hurdle	Net change if stopped
Claim case summary	6.43	Scale	-$1.24M
Finance reconciliation	9.22	Scale	-$281k
Finance commentary draft	10.24	Scale	-$96k
Code change agent	1.53	Fix	-$607k
Call summary (batch)	0.93	Stop	-$775k
Document questions	2.34	Scale	-$309k

A negative stopping value means the forecast loses more benefit than it saves avoidable cost. The hurdle is a review trigger, not an automatic stop decision.

Exhibit 1. Capacity value is shown, but kept apart from planned benefit. Net change if stopped sits beside the hurdle. Illustrative.

10. Dashboards by role (continued)

CIO and CTO: what drives the run cost

CIO and CTO: what drives run cost

Year 1 · Base scenario · Illustrative forecast · Workbook version 2

Token cost	Recurring cost	Frontier share
$1.15M	**$1.37M**	**92%**
46% of run cost	Gateway, evaluation, retrieval, data and people	Token spend concentrated in one workflow

Token cost, tier and limits by workflow

Workflow	Tier	Cap / user / month	Cap binds	Token cost	Fully allocated
Claim case summary	Standard	$60	No	$62k	$207k
Finance reconciliation	Standard	$60	No	$17k	$33k
Finance commentary draft	Small	$15	No	<$1k	$10k
Code change agent	Frontier	$600	Yes	$1.06M	$1.09M
Call summary (batch)	Small	$20	No	$3k	$1.03M
Document questions	Standard	$25	No	$11k	$147k

The code change agent uses the frontier tier and its cap leaves about a third of demand unmet. Test routing, context size and acceptance before changing the cap.

Exhibit 2. Token cost by workflow, tier, caps and unmet demand. The frontier tier accounts for 92 percent of token spend through one workflow. Illustrative.

10. Dashboards by role (continued)

Product owner: one workflow, incremental choices

Product owner: code change agent

Year 1 · Base scenario · Illustrative forecast · Workbook version 2

Useful tasks	Cost per useful task	Planned benefit per task
95,482	**$11.39**	**$17.47**
Of 146,895 completed	Fully allocated cost	Forecast conversion, net of review

Code change agent: incremental choices

Option	Extra cost	Extra planned benefit	Ratio
Lift cap, no other change	$544k	$826k	1.52
Route routine steps to a smaller tier	Measure	Evaluate acceptance	Test
Trim context and summarize history	Measure	Evaluate review time	Test

Lifting the cap improves the forecast at a ratio below the illustrative 2.00 scale hurdle. Compare redesign and quality constraints before deciding.

Exhibit 3. The code change agent sits at Fix; lifting its cap is a policy trade-off to compare with redesign. Illustrative.

11. Value and return without the hype

WORKFLOW	FULLY ALLOCATED COST	PLANNED BENEFIT	BENEFIT / COST	HURDLE	NET CHANGE IF STOPPED
Claim case summary	$207k	$1.33M	6.43	Scale	-$1.24M
Finance reconciliation	$33k	$300k	9.22	Scale	-$281k
Finance commentary draft	$10k	$98k	10.24	Scale	-$96k
Code change agent	$1.09M	$1.67M	1.53	Fix	-$607k
Call summary (batch)	$1.03M	$964k	0.93	Stop	-$775k
Document questions	$147k	$344k	2.34	Scale	-$309k
Portfolio	$2.52M	$4.71M	1.87	Fix	n/a

The benefit-cost ratio is not ROI. The portfolio's net return, (planned benefit minus run cost) divided by run cost, is 87 percent under these assumptions; neither figure is a demonstrated cash return.

What the illustrative case shows

> **Non-token recurring costs are 54 percent of run cost** in this case: gateway, evaluation, retrieval, data upkeep, people, change and security.

> **Capacity is not cash.** $14.8 million of capacity becomes $4.71 million of planned benefit, and only from the months the owners commit to.

> **Allocation can mislead a stop decision.** Every workflow, including the one below the stop hurdle, makes the portfolio worse if stopped, because most shared cost is not avoidable.

> **Some apparent improvement is just price.** Year-two ratios rise partly because assumed price declines let capped workflows do more work. That is not demonstrated optimization; test flat and rising prices.

When to stop

Retire or redesign a workflow below the stop hurdle after a fix cycle, and only when the avoidable view supports it. Release only costs that are actually avoidable. Before announcing savings, identify stranded commitments, transition costs and service consequences.

12. Levers to test

Their ranking depends on workload, quality requirements and cost structure, so test each against cost and value per accepted task.

LEVER	WHAT TO MEASURE	OWNER
Route each step to the cheapest tier that meets its quality bar	Cost per accepted task and acceptance rate	Platform, product
Trim context: fewer, better passages; summarize history	Input tokens per call, accuracy, review minutes	Engineering
Cap calls per task and retries; stop loops early	Tail cost and variance	Engineering
Cache stable prompt prefixes, where eligible	Cache hit rate and effective input price	Engineering
Batch eligible work that can wait	Effective price and turnaround	Product, operations
Negotiate commitments sized to a demand floor, with price protection	Utilization and break-even	Procurement, finance
Retire low-value workflows and duplicate tools	Avoidable cost released	CFO, CIO

A cheaper model that lowers acceptance can raise the cost per accepted task. A pricier one that cuts review can lower it.

13. Operating model and cadence

ROLE	ACCOUNTABLE FOR
CFO and FP&A	Budget, forecast versions, reconciliation to invoices and ledger, hurdles, validation of realization plans
CIO and CTO	Metering, gateway, routing, limits, the economics gate, provider contracts and platform cost lines
Business owner	Acceptance criteria, baseline, the benefits contract and the actions that convert capacity to cash
Product owner	Workflow design, measured drivers, unmet demand and the case to scale, fix or stop
FinOps and procurement	Cost pools, allocation drivers, commitment utilization, anomaly alerts, optimization backlog
Risk and audit	Versioned acceptance criteria, evaluation evidence, override logs, traceability from task to charge

> **Daily:** anomaly alerts on spend per workflow and caps about to bind.

> **Monthly:** actuals reconciled, forecast refreshed, variance explained by driver.

> **Quarterly:** funding decisions per workflow with the CFO, the CIO or CTO and business owners, on all three cost views.

> **Every new workflow:** a forecast, a persona permission set and a benefits contract before the build.

14. The first 90 days

DAYS	WORK	EVIDENCE AT THE END
1 to 30	Set cost scope; meter all model calls through a gateway by workflow and persona; baseline each workflow; agree accepted-task definitions, allocation policy and accountable owners	Baseline cost per accepted task; named owners and definitions
31 to 60	Instrument pilots; validate quality, cost and review effort; build the workflow forecast; set persona limits and cap behavior; draft benefits contracts	A forecast finance accepts, reconciled to pilot invoices
61 to 90	Reconcile actual spending; test benefit realization; stand up the dashboards and economics gate; make staged funding decisions with a documented downside and exit path	Funding decisions recorded on all three cost views

Ten questions for the CFO to ask

> What is our cost per accepted task, by workflow, and who defines accepted?

> How much of run cost is tokens, and what is everything else?

> Does the forecast reconcile to invoices and the ledger, and which drivers explain the variance?

> How much capacity are we releasing, and how much will be realized as cash, by whom and when?

> Which costs would really disappear if we stopped a workflow?

> What would lifting each binding cap cost, and return?

> Is review time netted once, and is any of it a new paid cost?

> What happens to our economics if prices rise, or introductory terms end?

> What do our provider contracts commit us to, and at what utilization do they break even?

> What does the business downside look like, and are we still comfortable?

Sources

Prepared by Richard White, October 2026. Version 2, published edition. Workbook figures are illustrative forecasts, not measured results. Sources checked 13 October 2026.

[1] Gartner, "Four Emerging Challenges to Delivering Value from AI Safely and at Scale," 21 October 2024. https://www.gartner.com/en/newsroom/press-releases/2024-10-21-gartner-identifies-four-emerging-challenges-to-delivering-value-from-ai-safely-and-at-scale

[2] Gartner, "10 Best Practices for Optimizing Generative and Agentic AI Costs," June 2026, as reported by Campus Technology, 22 June 2026: at least 50 percent of GenAI projects will overrun budgeted costs through 2028. A prediction. https://campustechnology.com/articles/2026/06/22/gartner-half-of-gen-ai-projects-could-exceed-budget-by-2028.aspx See also Gartner, "The 2026 Hype Cycle for GenAI," 22 July 2026. https://www.gartner.com/en/articles/hype-cycle-for-genai

[3] FinOps Foundation, State of FinOps 2025: 861 respondents; 63 percent managing AI spend, up from 31 percent. https://data.finops.org/2025-report/

[4] FinOps Foundation, State of FinOps 2026: 1,192 total respondents; the AI section reports 98 percent managing AI spend and N=693. Question samples vary. https://data.finops.org/

[5] Bai, L. et al., "How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks," arXiv:2604.22750, 2026; summarized by the Stanford Digital Economy Lab, 5 May 2026.

[6] Wang, S. and Casado, M., "The Cost of Cloud, a Trillion Dollar Paradox," Andreessen Horowitz, 27 May 2021. https://a16z.com/the-cost-of-cloud-a-trillion-dollar-paradox/

[7] Anthropic, Claude API pricing documentation. Model, context, caching and processing choices affect effective rates; check the applicable contract. https://platform.claude.com/docs/en/about-claude/pricing

[8] Epoch AI, "LLM inference prices have fallen rapidly but unequally across tasks," 12 March 2025. https://epoch.ai/data-insights/llm-inference-price-trends

[9] FinOps Foundation, Framework capabilities: Allocation, and Unit Economics. https://www.finops.org/framework/

Related RJL papers: Why Enterprise AI Stalls After the Demo; The Connected Enterprise Data & AI Blueprint; Agent Onboarding. Proposed operating guidance; not financial, accounting or legal advice.

PRODUCT CONSTRUCTION / REUSE ECONOMICS

A second consumer is a business case, not free value.

Evaluate reuse against the real alternative: supported reuse, a derived interface, a separate build or no investment.

COST OR BENEFIT	HOW TO TREAT IT
First product investment	Include source integration, validation, definition work, controls, testing and initial deployment; separate sunk costs from costs relevant to the next decision.
Continuing service	Count compute/storage, monitoring, support, access review, corrections, incident response, contract maintenance and retirement obligations.
Second-consumer increment	Estimate onboarding, permissions, semantic changes, performance isolation, additional evaluation and compatibility/migration work.
Avoided duplication	Compare against a credible separately built alternative over the same horizon and workload. Planned avoided future effort is not automatically realized cash savings.
Business benefit	Measure the named reporting or collection outcome against its baseline. Time released becomes cash benefit only through a specific owned realization plan.

For the synthetic receivables product, a second report may reuse the existing cutoff and definition. A live collections assistant may require a different API, stronger latency guarantees and action controls. Include those differences before attributing savings to a shared data label.

Track total portfolio cost and the marginal cost of each additional consumer. Any allocation of shared platform cost should be transparent. Do not count the same time saving in both the foundational product and its derived assistant benefit.

This supplement adds decision guidance; it does not change the companion workbook formulas or turn its illustrative forecast into an NPV, IRR or investment cash-flow model. See Data Products in Practice for the product construction sequence.

Design references reviewed 29 October 2026: Databricks: medallion architecture; Designing data products; Google SRE: service objectives. The worked records, controls and targets are illustrative RJL recommendations.

ABOUT

Richard White

I was global firm chief data officer at McKinsey & Company, owning data, AI and platforms end to end, and before that chief data officer at New York Life, where I built the chief data office and reported data risk to risk committees at multiple levels. Earlier I led data at the Federal Reserve Bank of New York and enterprise data governance at Citigroup. I founded RJL Solutions in 2024. My book, *The Day After the Demo*, is free at rjlsol.com/book.

RJL Solutions connects strategy, architecture, engineering, governance and operation through Frame, Prove, Build, and Run & Transfer. Each phase is separately scoped and priced in a written proposal before you commit. The first sixty-minute working session is free.

DELIVERY **Why Enterprise AI Stalls**
What must change before people rely on AI.

ARCHITECTURE **The Connected Enterprise Blueprint**
Business outcomes, ownership and release evidence.

AGENTS **Agent Onboarding**
Contracts for accepted work and controlled actions.

RJL SOLUTIONS

WEB rjlsol.com

EMAIL rwhite@rjlsol.com

Companion model: rjlsol.com/library/token-economics. Values are illustrative forecasts. Use measured workflow data and an owned benefits plan before making an investment decision.