VIEW POINT

AI TOKENOMICS

WHAT EVERY EXECUTIVE NEEDS TO KNOW

Your AI program's largest financial risk isn't the model you picked or the people you hired. It's a unit of consumption that never made it into the budget.

Infosys | CONSULTING

WHAT'S NEW IN THIS EDITION

This edition updates the April 2026 brief to Framework v5. Version 5 rechecked every headline number and corrected the ones that didn't hold up. Several figures came down, and what replaced them stands up better in a budget review. Four changes matter to anyone signing off on spend:

Savings is a range, not a promise. The five optimization levers overlap; they cut into the same spend rather than stacking cleanly. Realized cost reduction lands around 45% on average. The 49% figure from the case study is a strong result near the top of the range, not the default, and the old 64–71% number turns out to be a ceiling the levers can't reach together. Plan against the middle of the range, not the best case.

The supporting statistic is now sourced and verifiable. The earlier "67% over budget" figure has been dropped for Gartner's April 2026 infrastructure-and-operations survey: 28% of AI use cases fully meet ROI expectations, one in five fail outright, and 57% of leaders report at least one AI failure.

The two-lever concentration is smaller than first stated. Retrieval governance and model routing account for roughly 58% of achievable savings, revised down from 78%. Still the first place to spend effort. No longer the whole story.

Budget to the tail. Because the range is wide, fund the program to its 95th-percentile cost (~$2.42M/month at full scale) rather than the median (~$1.87M). Fund to the average and you come up short in roughly half of all outcomes.

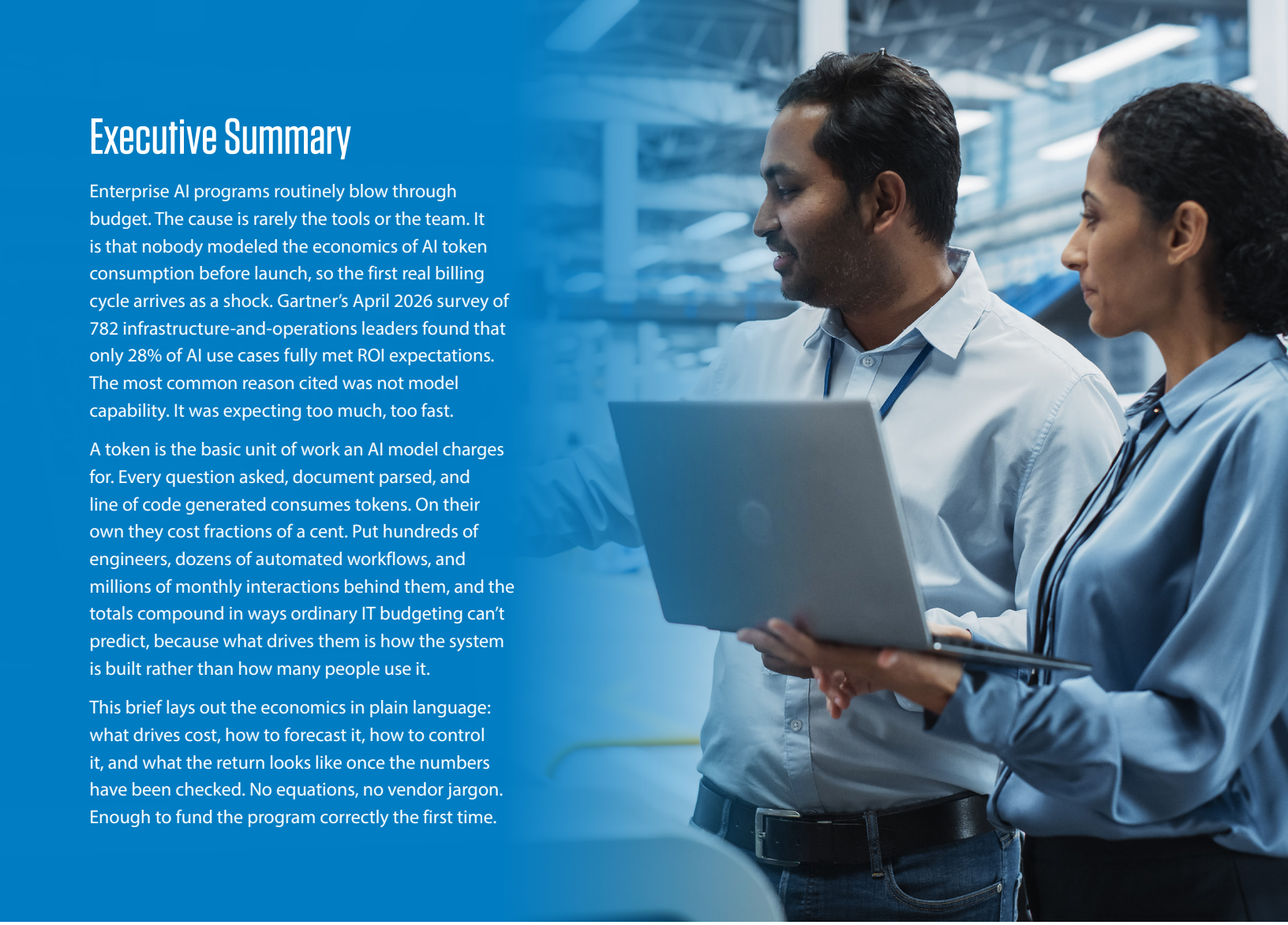
External Document © 2026 Infosys Limited

Executive Summary

Enterprise AI programs routinely blow through budget. The cause is rarely the tools or the team. It is that nobody modeled the economics of AI token consumption before launch, so the first real billing cycle arrives as a shock. Gartner's April 2026 survey of 782 infrastructure-and-operations leaders found that only 28% of AI use cases fully met ROI expectations. The most common reason cited was not model capability. It was expecting too much, too fast.

A token is the basic unit of work an AI model charges for. Every question asked, document parsed, and line of code generated consumes tokens. On their own they cost fractions of a cent. Put hundreds of engineers, dozens of automated workflows, and millions of monthly interactions behind them, and the totals compound in ways ordinary IT budgeting can't predict, because what drives them is how the system is built rather than how many people use it.

This brief lays out the economics in plain language: what drives cost, how to forecast it, how to control it, and what the return looks like once the numbers have been checked. No equations, no vendor jargon. Enough to fund the program correctly the first time.



28%
of AI use cases fully meet ROI expectations (Gartner, Apr 2026)

~45%
median cost reduction from full governance; ~49% in a strong case

≈ 3 days
governance payback at full program scale

THE CENTRAL INSIGHT

AI token costs don't behave like server costs. They don't scale with users or workload. They scale with architecture: how your AI agents are designed, how they retrieve information, which models they call, and how they pass work between each other. That is also the opening. The same decisions that push cost up can be engineered to pull it back down. Systematic governance yields roughly a 45% reduction with no loss of capability. Treat that as the number to plan around, not a figure to bank on.

External Document © 2026 Infosys Limited

01 THE NEW COST FRONTIER IN ENTERPRISE AI

THE HIDDEN BUDGET PROBLEM

When enterprises budget for an AI program, they reach for familiar categories: software licenses, cloud compute, implementation services, headcount. Each one is real and each one gets counted. Together they still miss the cost that surprises most programs in their first year of production.

AI tokens. Every interaction with a model is metered and billed by the token: every prompt, every response, every document processed, every review completed. For one developer trying out a tool, the cost is invisible. For a program running hundreds of engineers and dozens of pipelines around the clock, it becomes the dominant line item.

Why Traditional IT Budgeting Breaks Here

Conventional infrastructure cost is predictable because it scales with things you can count: users, storage, compute cycles. Add fifty users and you add roughly fifty units of cost. Linear, legible, easy to forecast.

Token cost doesn't behave that way. It is shaped by architectural decisions that stay invisible at the executive level:

- How many AI agents collaborate on a task, and how much context each one hands to the next.
- Whether models are matched to task complexity, or whether every task defaults to the most expensive model available.
- How much background information is pulled in each time an AI answers a question.
- How often automated workflows retry when a step fails, paying full price for each restart.

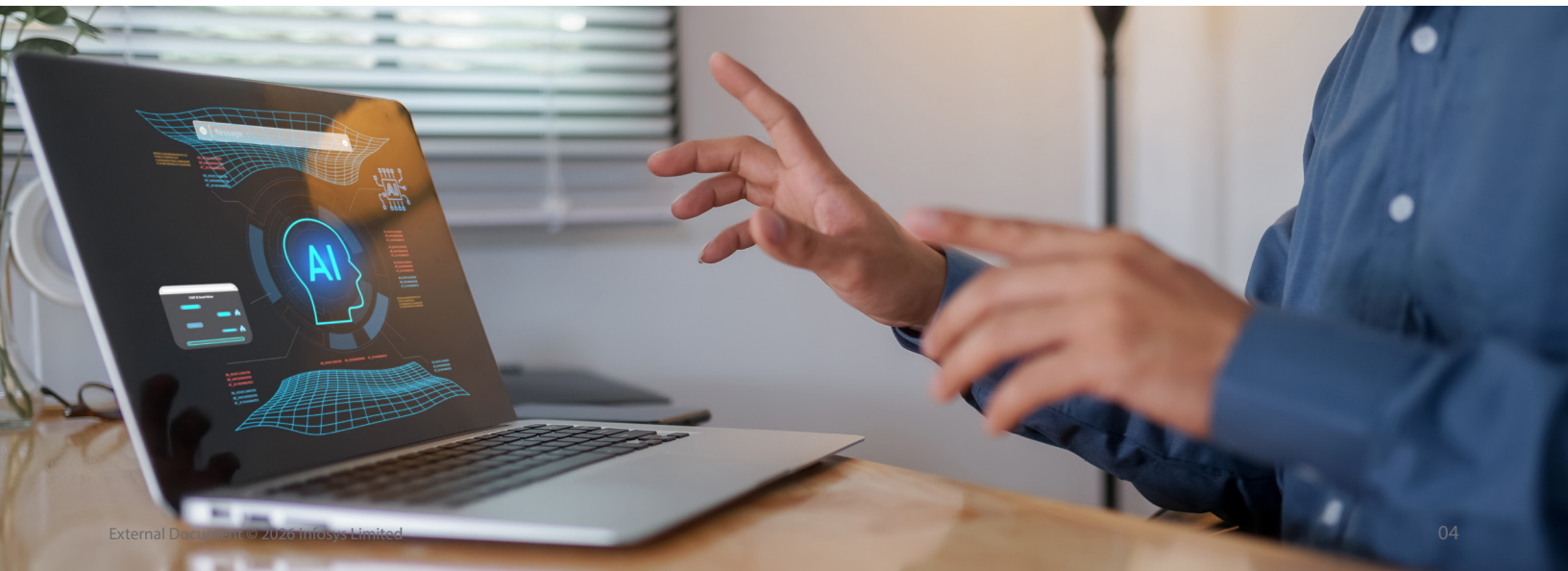
Any of these can multiply cost by 2×, 5×, even 50×, with no one ever deciding to spend more. The total builds quietly until the first serious billing cycle lands.

A PATTERN YOU'LL RECOGNIZE

A program launches an AI-assisted software-development initiative. Month 1 looks reasonable. By month 3 more workflows are automated and more engineers are on board. By month 6 the monthly AI bill is several times the original estimate. Nobody added features. Nobody chose to spend more. The architecture was never designed with cost governance in mind, and nothing in the tooling forced the question.

This Is Solvable, and the Field Data Says So

Token cost is poorly governed in most programs for one reason: the frameworks for governing it are new. Gartner's 2026 data locates the failures precisely. They cluster in auto-remediation, self-healing infrastructure, and agent-led workflow management, which is exactly where token cost concentrates. The open question was never whether AI can reduce cost. It is whether an organization can forecast and govern that reduction well enough to survive production. Programs that build the right controls before they scale typically run their AI infrastructure at roughly 45% lower cost, with no loss of capability or developer productivity.



External Document © 2026 Infosys Limited

02 THE COMPLETE COST PICTURE

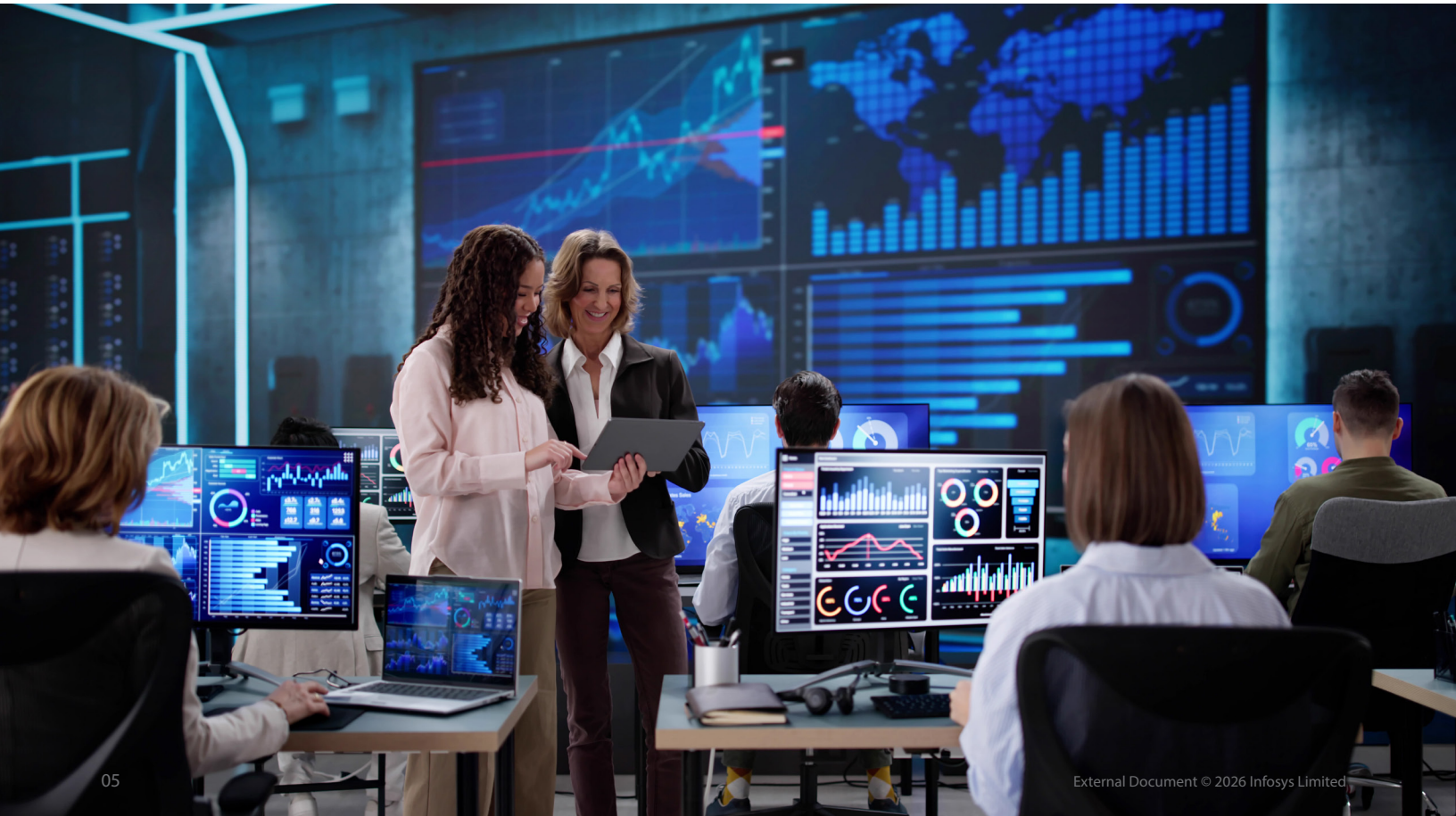
FOUR COSTS, NOT ONE

Most AI budgets track one of the four real cost categories. That partial view is where most AI budget failures begin. Seeing all four is the prerequisite for controlling any of them.

#	Cost Category	What It Is	Typical Share
1	**Direct AI Usage Fees**	Per-token charges from model providers: every question answered, document analyzed, review completed. What most teams track, and only part of the picture.	55–65%
2	**Platform Infrastructure**	The supporting stack: memory stores, workflow orchestration, monitoring and logging, the gateway that routes requests. Often invisible in AI budgets; adds 15–25% on top of usage fees.	15–20%
3	**Efficiency Investment**	The one-time and ongoing cost of building the governance and optimization layer. The category most programs skip, and the reason they overspend on the first two.	5–10%
4	**Cost of Over-Restriction**	The productivity lost when AI budgets are set too tight. Real and measurable, it simply moves from the IT line to the business-unit productivity line.	Modeled separately

THE EXECUTIVE PRINCIPLE

AI cost governance isn't about spending less. It's about spending in the right places. The goal is the lowest total cost, which includes the cost of blocking legitimate, high-value use, not the lowest monthly API bill. Cut budgets so hard that engineers can't get work done and you haven't saved money. You've moved the cost onto the productivity line, where it is larger and harder to see.



External Document © 2026 Infosys Limited

03 WHAT DRIVES AI COSTS

SIX THINGS THAT DETERMINE WHAT YOU PAY

Token cost isn't random. Six identifiable factors drive it, and each one can be measured, forecast, and managed. Knowing them is the difference between a program that holds its budget and one that surprises you every month.

	Cost Driver	In Plain English	The Risk If Unmanaged
D1	**Pipeline Depth**	Every task breaks into steps, and each step costs tokens. Longer pipelines cost more, and a failed step retries at full context cost.	An 8-step pipeline at a 12% retry rate uses nearly 30% more tokens than the same pipeline with no failures.
D2	**Agent Teamwork**	When agents collaborate, each one receives a full summary of what came before. More agents in the chain means more tokens, before any of them produce useful output.	A 3-agent chain can consume several times the tokens of a single agent doing the same task, with no gain in quality.
D3	**Information Retrieval**	AI often pulls in background information to answer. How much it injects per query is the largest single variable cost in most programs.	Ungoverned retrieval uses 7–8× more tokens than governed retrieval. This one factor can double monthly spend.
D4	**Model Selection**	Models don't cost the same. Frontier models can cost 10–15× more per query than lightweight ones, and most programs default to the expensive one for everything.	Frontier-for-everything typically costs 2–3× more than matching model to task.
D5	**Adoption Speed**	As more people use the tools, consumption grows along an S-curve, not a straight line. How fast adoption spreads decides whether cost growth is predictable or sudden.	Programs that onboard faster than governance matures overspend in months 3–6, then scramble to impose restrictions that hurt productivity.
D6	**Quality Guardrails**	Every output should be validated before it's acted on. Validation adds a small overhead; skipping it costs far more, because errors caught late are the expensive ones.	The cost of validation is recovered many times over in prevented remediation. The return dwarfs the overhead.

First Insight: Two Drivers Lead, but They Aren't the Whole Prize

You don't fix everything at once. Two drivers account for roughly 58% of all achievable savings, revised down in the current framework from an earlier and too-generous 78%:

- Information retrieval governance: controlling how much background context agents pull in per task.
- Model routing: matching model capability to task complexity automatically, instead of defaulting to expensive frontier models.

Both are configuration changes, not architectural overhauls. Neither one changes what the AI does or how engineers use it, and both produce immediate reductions. Start here. The revised 58% is the correction that matters, though: the remaining ~42% is spread across agent coordination, caching, context management, and memory, which is why this work is a portfolio rather than a single fix.

SECOND INSIGHT: OPTIMIZATION MOVES COST, IT DOESN'T DELETE IT

The most important structural lesson in the current framework is that context-management "savings" usually relocate cost rather than remove it. Trim an agent's memory to save context tokens and it re-derives what it lost, so the cost comes back as retries. Offload context into a retrieval system and the saving comes back as retrieval spend, at the 7–8× ungoverned multiplier. Every apparent efficiency lands somewhere. The job at the executive level is to insist on knowing where. An optimization you can't trace is a cost you've hidden from yourself, not one you've removed.

External Document © 2026 Infosys Limited

04 FORECASTING: KNOWING WHAT YOU'LL SPEND BEFORE YOU SPEND IT

FROM SURPRISE BILLS TO DEFENSIBLE BUDGETS

The most common frustration in enterprise AI is the gap between the projection and the invoice. It isn't that AI cost can't be forecast. It's that programs forecast it the way they forecast servers, estimating users and multiplying by a per-user rate, while the six drivers interact non-linearly. Closing the gap takes a model built for that.

The Three-Layer Forecast

Layer	What It Does	How to Think About It
1 Baseline	Establishes AI consumption under standard conditions, task by task, across every phase of the development lifecycle.	Your "do-nothing" reference: the cost you're engineering below.
2 Adjustment	Multiplies the baseline by real architectural variables: agent coordination, model selection, retrieval volume, caching efficiency.	Turns the baseline into a projection of your actual architecture.
3 Scenarios	Runs conservative, base, and worst-case futures so leadership sees the financial consequence of each governance choice.	Converts technical complexity into business decisions with quantified stakes.

Your Forecast Is a Range, So Budget to Its Edge

This is the most consequential change in the current framework, and it changes how the number is calculated, not the economics underneath. The five optimization levers were once multiplied together as if independent, which produced a headline reduction north of 60%. They aren't independent. They compete for the same spend. Account for that overlap and the realized reduction becomes a distribution:

Outcome	Cost Reduction	Governed Monthly Cost (full scale)	How to Read It
Conservative (P5)	~29%	~$2.42M	The unlucky tail. Budget here.
Typical (median)	~45%	~$1.87M	The outcome to expect.
Strong (case study)	~49%	~$1.8M	A good, above-median result, not the default.
Ceiling (theoretical)	~64%	—	What the levers would deliver if they didn't overlap. Out of reach.

BUDGET TO THE TAIL

Because the range is wide, provision the program to its 95th-percentile cost (~$2.42M/month at full scale) rather than the median (~$1.87M). Fund to the average and you come up short in roughly half of all outcomes, which is how a well-governed program still ends up looking like a runaway to finance. The width of the range isn't a flaw in the model. It's the signal. Use it to set a budget line you can defend.

External Document © 2026 Infosys Limited

Scenario Planning: Separate the Two Things That Move the Number

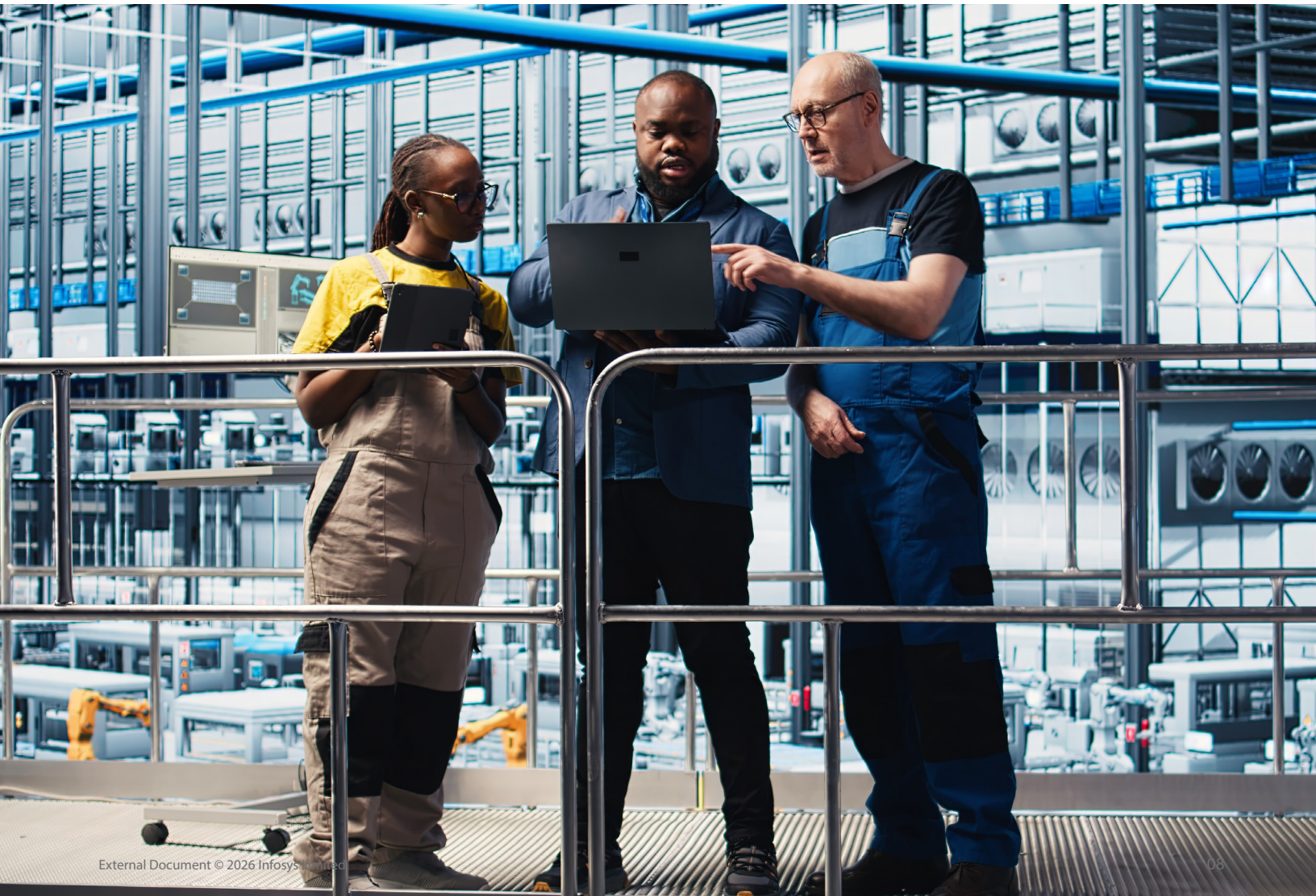
Three scenarios bound the range at full program scale, drawn from a representative deployment at a large semiconductor and infrastructure technology company. Read them carefully. The worst case varies two things at once, how fast users onboard and whether the program is governed, and those are separate levers.

Scenario	Adoption	Governance	Cost @ Mo. 6	Cost @ Mo. 18
Conservative	Slow ramp	Full	~$285K	~$820K
Base Case	On-plan	Full	~$318K	~$1.1M
Ungoverned / Fast	Aggressive	Delayed / skipped	~$1.24M	~$3.6M

The 3.3× gap between the base and ungoverned cases at month 18 is really two effects. About 1.5× comes from faster adoption, meaning more users, and about 2.2× from absent governance. Hold the user count constant and the governance-only penalty is roughly 2.2×. That is the number to put in front of leadership: the cost of the governance decision by itself, separated from the cost of the growth you actually want.

FORECASTS SHARPEN WITH DATA

Early forecasts carry real uncertainty, typically ±60–80% in the first month. That is normal, and it narrows quickly as production data accumulates: ±20–30% by month 3, ±10–15% by month 6. Executive dashboards should show this convergence directly, so leadership matches its confidence to the stage of the program instead of treating a month-1 estimate as a commitment.



External Document © 2026 Infosys Limited

05 THE GOVERNANCE ARCHITECTURE

FOUR CONTROLS THAT DO THE WORK

Knowing what drives cost is necessary but not enough. The governance architecture is the infrastructure that makes control automatic. None of it changes what the AI does. All of it runs in the background, invisible to the people using the tools.

Governance Component	What It Does	Business Analogy
The AI Gateway	The traffic controller for every model call. Routes each task to the right model tier, enforces team-level spending limits, attributes every dollar to the workflow that spent it, and shows spend in real time.	An expense system that routes purchases to the right cost center and flags overspending before it's a problem.
Quality Guardrails	Validates every output before it reaches an engineer or a downstream system. Catches errors and policy violations at the source, before they turn into expensive remediation.	Quality-control inspection on a production line, at a fraction of the cost of shipping the defect.
Workflow Budget Controls	Encodes spending limits into the automated workflows themselves. A step about to exceed budget reroutes to a cheaper model instead of failing or overspending.	Spending rules built into procurement, so compliance is automatic rather than left to individual judgment.
Observability & Measurement	Real-time and historical visibility into every token, model, and dollar, broken out by team, workflow, and lifecycle phase. The data foundation for every other control.	The financial dashboard executives already have for headcount and licensing, extended to AI spend.

THE IMPLEMENTATION SEQUENCE MATTERS

Build observability first. You cannot govern what you cannot see.

Implement model routing right away, before onboarding scales. Lowest effort, highest return.

Deploy quality guardrails in audit mode, then enforce once the data shows where the risk is.

Add workflow budget controls as automated pipelines mature.

ON THE HORIZON: THE NEXT COST FRONTIER IS MEMORY

These four controls discipline what a model call costs. They don't yet discipline what the system remembers between runs. In a mature multi-agent program, each pipeline execution starts from scratch, re-deriving at full cost the lessons earlier runs already paid for. The emerging design principle is to add a priced memory layer beneath the workflow orchestrator, where every read and write is metered like any other consumption. This isn't something to action today. It's worth knowing the frontier is moving, and that the same rule applies to it: an ungoverned memory layer becomes the next place cost hides.

External Document © 2026 Infosys Limited

06 THE BUSINESS CASE

THE ROI OF GETTING THIS RIGHT

Executive conversations about governance often get framed as a trade-off: invest in controls and slow down, or move fast and carry the risk. That framing is wrong. The governance layer runs in the background and adds no friction to how engineers work. The real question isn't whether to build it. It's when, and the answer is before you scale.

The Investment

The full governance architecture costs roughly $178,000 one-time in engineering and tooling, plus about $22,000 a month to operate. Both figures line up with documented AI cost-governance implementations.

The Return

At full scale, the gap between a governed and an ungoverned architecture runs about $1.8 million a month in avoided infrastructure cost.

~$178K
one-time governance investment
(engineering + tooling)

~$1.78M
monthly savings at full scale,
governed vs. ungoverned

≈ 3 days
payback at full program volume

WHY THE RETURN HOLDS EVEN WHEN THE SAVINGS DON'T

The ~$1.78M figure is the case-study point, a strong and above-median outcome. The question worth asking is whether the business case survives a weaker one. It does, and comfortably. Even under the framework's more conservative modeled outcomes, payback stays inside a couple of weeks rather than stretching into quarters. Few IT infrastructure investments pay back in days, and this one holds that property across the whole range. That is what makes it fundable in front of a skeptical finance team.

The Risk of Waiting

Every month of delay is a month of avoidable overspend. At full scale, a single ungoverned month costs more than the entire governance implementation. The programs that get this wrong are rarely the ones that tried to govern cost and failed. They are the ones that put governance off until cost was already a crisis, then clamped down with restrictions that hurt both productivity and trust in the program.

What Good Looks Like: A Phased Approach

Phase	Timeline	Governance Focus	Expected Monthly Cost
Foundation	Months 1–3	Deploy the gateway and observability stack. Establish baseline measurements. Understand where tokens go before imposing controls.	$420K–$580K (pre-optimization)
Optimization	Months 4–6	Activate model routing and retrieval governance. Move guardrails into enforcement. First real reductions appear.	$780K–$1.1M (first optimizations)
Full Scale	Months 7–18	All controls active. Workflows operating under budget constraints. Forecasts accurate to within 10–15%.	$1.5M–$2.0M governed vs. $3.2M–$3.6M ungoverned

External Document © 2026 Infosys Limited

07 SIX DECISIONS EVERY EXECUTIVE MUST OWN

YOUR ROLE IN AI COST GOVERNANCE

Token economics is a technical discipline, but the decisions that decide success or failure are executive ones. Research on managing AI at enterprise scale consistently finds that executive ownership of these decisions is the strongest predictor of program cost outcomes. Six matter most. The sixth is new to this edition.

#	The Decision	Why It Matters	The Risk of Getting It Wrong
1	Fund governance before the program scales.	Payback is days, not quarters. Every month of delay is avoidable overspend.	Deferring governance routinely means 2–4× overspend in months 3–6, then emergency restrictions that damage adoption and trust.
2	Require cost attribution from day one.	You cannot govern what you cannot see. Knowing which teams and workflows spend what is the prerequisite for everything else.	Without attribution, cost reduction is guesswork and budget talks with engineering turn adversarial.
3	Set token budgets as ranges, not hard caps.	Budgets set too tight block legitimate work. Set them at ~150% of baseline during ramp-up, then tighten quarterly as data improves.	Over-restriction cuts productivity and drives abandonment, shifting cost from IT to the business-unit line.
4	Gate scaling on governance maturity, not adoption metrics.	Fast onboarding without mature governance is the most common cause of budget overrun.	Approving 200+ users before routing, caching, and observability are live turns a controllable trajectory into an emergency.
5	Treat the forecast as a living document.	Models sharpen as data accumulates. A quarterly forecast-vs-actual reconciliation is a strategic asset.	Static forecasts drift from reality fast; programs without reconciliation stay permanently reactive.
6	Provision to the tail, not the average.	The forecast is a range. Fund to the 95th-percentile cost (~$2.42M/month) so a normal unlucky outcome doesn't read as a failure.	Budget to the median and you underfund the program in half of outcomes: the well-governed program that still looks like a runaway.

08 CONCLUSION

THE PROGRAMS THAT WIN

Enterprise AI is past the proof-of-concept stage. Whether AI belongs in enterprise software development is settled; it does. What's still open is which organizations build their programs on a foundation that lets them scale with confidence, and which spend the next two years reacting to billing surprises.

The winners will treat AI token economics as a first-class strategic concern from the start, not a technical detail to sort out after cost becomes a crisis. They will also hold the numbers to the standard this edition applies: a governed program saves roughly 45%, not a guaranteed 50; the savings are a range, not a point; and the budget belongs at the edge of that range, not its center. None of that is pessimism. It's the difference between a budget you can stand behind and one you spend the year explaining.

External Document © 2026 Infosys Limited

THE SUMMARY IN FOUR SENTENCES

AI token costs are architecture-driven, not user-driven. How you route model calls and govern retrieval accounts for roughly 58% of achievable savings, and the rest is a portfolio, not a single fix.

Full governance cuts cost by roughly 45% at the median. The strong case reaches ~49%, and the theoretical ceiling of ~64% stays out of reach because the levers overlap.

The governance infrastructure that controls these costs pays back in days and keeps paying back across the range of outcomes. There is no credible reason to defer it.

Budget to the tail, not the average. Fund the program to its 95th-percentile cost, so the width of the forecast becomes a line item you can defend instead of a quarterly surprise.

Selected References

1. Gartner. (2026). Gartner says AI projects in infrastructure and operations stall ahead of meaningful ROI returns [Press release, April 7, 2026].
2. Ismanov, I., et al. (2024). AI and cost management: Strategies for reducing expenses and improving profit margins. 2024 International Conference on Knowledge Engineering and Communication Systems. IEEE.
3. Hassan, S. Z., et al. (2025). From tokens to tactics: Operationalizing generative AI in enterprise workflows. IEEE ICITEICS 2025.
4. Sorensen, S. (2026). The token industrialist. Human–AI Collaborative Research Series Working Paper.
5. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. MIS Quarterly, 45(3), 1433–1450.
6. Przegalinska, A., et al. (2025). Collaborative AI in the workplace: Resource-based and task-technology fit perspectives. International Journal of Information Management, 81, 102853.
7. Enholm, I. M., et al. (2022). Artificial intelligence and business value: A literature review. Information Systems Frontiers, 24(5), 1709–1734.
8. Vial, G., et al. (2023). Managing artificial intelligence projects. Information Systems Journal, 33(3), 669–691.
9. Peretz-Andersson, E., et al. (2024). AI implementation in manufacturing SMEs: A resource orchestration approach. International Journal of Information Management, 77, 102781.
10. Akpan, M. (2024). Beyond attention: Advancing AI token valuation through user engagement and market dynamics. Corporate Ownership and Control, 21(4), 8–14.
11. Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.
12. Palomares Carrascosa, I. (2026). Context window management for long-running agents: Strategies and tradeoffs. MachineLearningMastery.com [practitioner source].

ABOUT INFOSYS CONSULTING

Infosys Consulting is a next-generation consulting partner that bridges strategy and execution. With an AI-first mindset, deep industry knowledge, and the combined strengths of business and technology consulting, it helps enterprises turn bold vision into tangible outcomes, faster, smarter, and at scale. Infosys Consulting is helping some of the world's most recognizable brands transform and innovate. Our consultants are industry experts that lead complex change agendas driven by disruptive technology. With offices in 20 countries and backed by the power of the global Infosys brand, our teams help the C-suite navigate today's digital landscape to win market share and create shareholder value for lasting competitive advantage.

For more information, contact consulting@infosys.com

Infosys® | CONSULTING

© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.