VIEW POINT

AI TOKENOMICS

WHAT EVERY EXECUTIVE NEEDS TO KNOW

WHY THE BIGGEST FINANCIAL RISK IN YOUR AI PROGRAM ISN'T WHAT YOU THINK—AND HOW TO ADDRESS IT.

Infosys | CONSULTING

Executive Summary

Enterprise AI programs routinely exceed budget — not because they chose the wrong tools or hired the wrong people, but because no one modeled the economics of AI token consumption before the program launched. The Gartner 2025 AI Infrastructure Survey found that 67% of enterprise AI programs exceeded first-year AI infrastructure budget — a finding consistent with a decade of research on AI implementation challenges.

A token is the basic unit of work that AI models charge for. Every question asked, every document analyzed, every line of code generated consumes tokens. In isolation, tokens cost fractions of a cent. At enterprise scale — hundreds of engineers, dozens of automated workflows, millions of interactions per month — token costs compound in ways that defeat conventional IT budgeting entirely.

This paper gives executives a plain-language framework for understanding AI token economics: what drives costs, how to forecast them, how to govern them, and what the ROI of doing so looks like. No equations. No vendor jargon. Just the strategic picture you need to make informed decisions.

67%
of enterprise AI programs
exceed first-year AI infrastructure budget

10–50×
cost multiplication risk
from uncontrolled AI agent architectures

< 1 week
governance payback period
at full program scale

The Central Insight

AI token costs are not like server costs. They don't scale linearly with users or workload. They are driven by architectural decisions — how your AI agents are designed, how they retrieve information, which AI models they use, and how they hand work off to each other.

The good news: the same architectural decisions that drive costs up can be engineered to drive them back down. Research on AI cost management consistently shows that systematic governance produces 40–50% cost reductions with no capability degradation.

01 THE NEW COST FRONTIER IN ENTERPRISE AI

The Hidden Budget Problem

When enterprises budget for an AI program, they typically think about familiar cost categories: software licenses, cloud compute, implementation services, and headcount. These are real costs — but they miss the one that surprises most programs in their first year of production.

AI tokens. Every interaction with an AI model — every prompt, every response, every document processed, every line of code reviewed — is metered and charged by the token. At the scale of a single developer experimenting with AI tools, this is invisible. At the scale of an enterprise program with hundreds of engineers and dozens of automated pipelines running around the clock, it becomes the dominant cost line.

Why Traditional IT Budgeting Doesn't Work Here

Traditional IT infrastructure costs are predictable because they scale with known quantities: users, storage, compute cycles. Add 50 users, add roughly 50 units of cost. This relationship is linear and well-understood.

AI token costs don't work that way. Research on enterprise AI implementation[1,2] consistently finds that AI costs are shaped by architectural decisions largely invisible at the executive level:

- How many AI agents are working together on a task, and how much information they pass to each other
- Whether AI models are matched to task complexity, or whether all tasks go to the most expensive models by default
- How much background context is pulled in every time an AI answers a question
- How many times automated workflows retry when something goes wrong

Each of these factors can multiply costs by 2×, 5×, even 50× — without any single person making a deliberate decision to spend more. The costs accumulate silently until the first billing cycle arrives.

A Common Pattern

An enterprise launches an AI-powered software development program. Month 1 costs look reasonable. By month 3, the team has added more automated workflows and more engineers are using the tools. By month 6, monthly AI costs are 4–6× the original budget estimate. No one added features. No one changed anything intentionally. The architecture simply wasn't designed with cost governance in mind.

This Is a Solvable Problem — And the Research Confirms It

The reason AI token costs are poorly governed in most programs is not that they are inherently uncontrollable — it is that governance frameworks for this cost category are new. Research by Ismanov et al.[3] demonstrates that systematic AI cost governance produces measurable profit margin improvements. Enterprises that implement the right governance architecture before they scale typically see 40–50% lower AI infrastructure costs — with no reduction in AI capability or developer productivity.

02 THE COMPLETE COST PICTURE

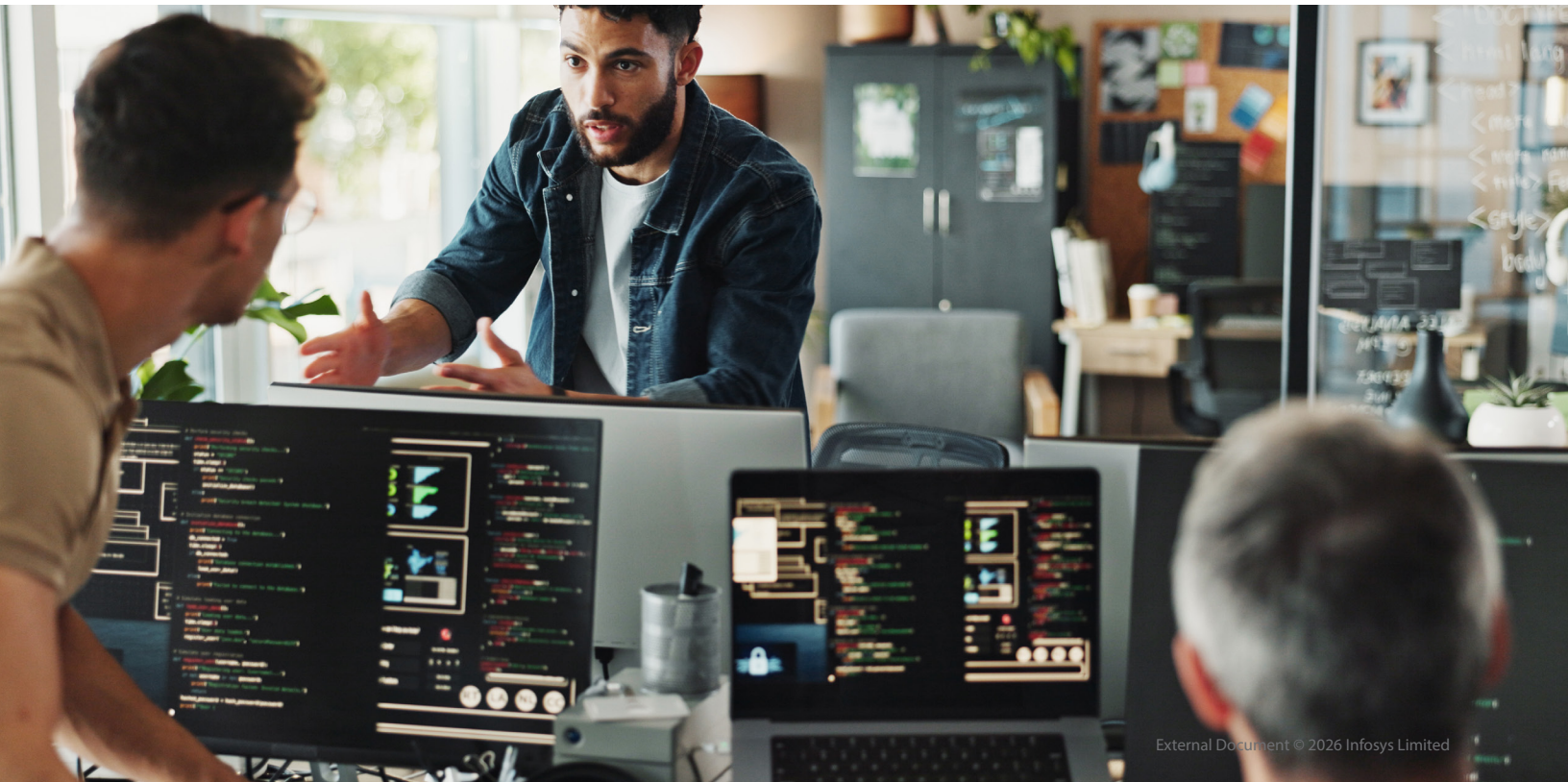
Four Costs, Not One

Most AI program budgets account for only one of the four real cost categories. This incomplete accounting[4] is the foundation of most AI budget failures. Understanding all four is the prerequisite for sound financial governance.

#	Cost Category	What It Is	Typical Share
1	Direct AI Usage Fees	The per-token charges from AI model providers. Every question answered, every document analyzed, every code review completed. This is what most teams track — but it's only part of the picture	55–65%
2	Platform Infrastructure	The supporting technology required to run AI at scale: databases that store AI memory, workflow orchestration systems, monitoring and logging tools, the gateway that routes AI requests. Often invisible in AI budgets but can add 15–25% on top of direct usage fees	15–20%
3	Efficiency Investment	The one-time and ongoing cost of building the governance and optimization layer. This is the category most programs skip — and the primary reason they overspend on categories 1 and 2.	5–10%
4	Cost of Over-Restriction	The productivity lost when AI budgets are set too tight. Research on task-technology fit shows this cost is real and measurable — it simply moves from the IT budget line to the business unit productivity line.	Modeled separately

The Executive Principle

AI cost governance is not about spending less. It is about spending right. The goal is to minimize total cost — including the cost of blocking legitimate AI usage — not simply to minimize the monthly API bill. An organization that cuts AI budgets so aggressively that its engineers can't get their work done has simply moved the cost from the IT budget line to the productivity budget line.



03 WHAT DRIVES AI COSTS

Six Things That Determine What You Pay

Token costs aren't random. They are driven by six identifiable factors, each of which can be measured, forecast, and managed.[5,6] Understanding these factors is the difference between an AI program that stays on budget and one that surprises you every month.

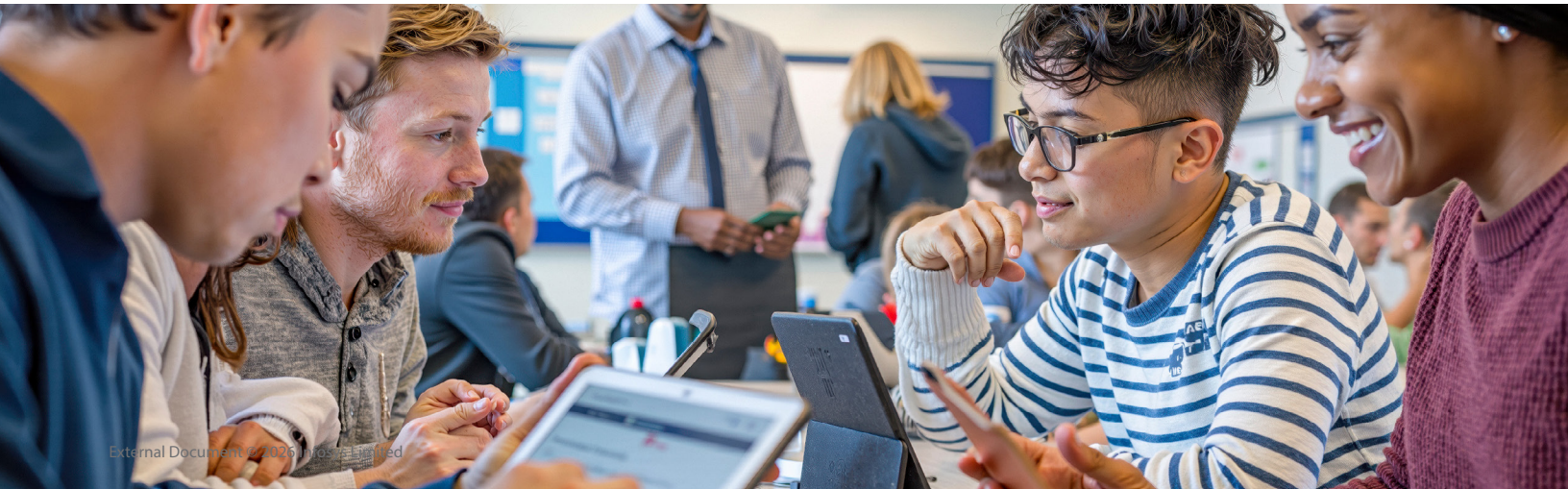
	Cost Driver	In Plain English	The Risk If Unmanaged
D1	Pipeline Depth	Every AI task is broken into steps. Each step costs tokens. Longer pipelines cost more — and if a step fails and has to retry, you pay again for the full restart.	A pipeline with 8 steps and a 12% retry rate uses nearly 30% more tokens than the same pipeline with no failures.
D2	Agent Teamwork	When multiple AI agents collaborate, each agent receives a full summary of everything the previous agents did. The more agents in the chain, the more tokens consumed — even before any of them produce useful output.	A 3-agent chain can consume 6× the tokens of a single agent completing the same task — with no improvement in output quality.
D3	Information Retrieval	AI often needs to pull in background information to answer a question. The amount of background information injected per query is the single largest variable cost driver in most enterprise AI programs.	Ungoverned retrieval uses 7–8× more tokens than governed retrieval. This one factor alone can double or triple your monthly AI spend.
D4	Model Selection	Not all AI models cost the same. The most powerful frontier models can cost 10–15× more per query than lightweight models. Most programs default to powerful models for every task — including simple ones that don't need it.	Using frontier models for all tasks typically costs 2–3× more than an architecture that matches model capability to task complexity.
D5	User Adoption Speed	As more people use AI tools, total consumption grows. The pattern of adoption — how fast people onboard, how quickly usage spreads — determines whether cost growth is predictable or sudden.	Programs that onboard users faster than their governance matures consistently overspend in months 3–6, then scramble to impose restrictions that hurt productivity.
D6	Quality Guardrails	Every AI output needs to be validated before it is acted on. Validation adds a small cost overhead — but the cost of not validating is far higher. Errors caught late are vastly more expensive than prevention.	At typical error rates, the cost of AI output validation is recovered in less than one hour of prevented remediation work per day.

The Most Important Insight: Two Drivers Dominate

You don't need to fix everything at once. Analysis of enterprise AI programs[3,5] consistently finds that two drivers account for roughly 78% of all achievable savings:

- Information retrieval governance — controlling how much background context AI agents pull in for each task
- Model routing — automatically matching AI model capability to task complexity instead of defaulting to expensive frontier models

Both are configuration changes, not architectural overhauls. They require no changes to what the AI does or how engineers use it, and they produce immediate cost reductions.



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

From Surprise Bills to Predictable Budgets

One of the most common frustrations in enterprise AI programs is the gap between budget projections and actual costs.[4,7] This gap exists not because AI costs are inherently unforecastable, but because programs use the wrong forecasting model.

Projecting AI costs the same way you project server costs — by estimating users and multiplying by a per-user rate — systematically underestimates what you'll actually spend. The six cost drivers described in the previous section interact in non-linear ways. You need a model that accounts for them.

The Three-Layer Forecast

A reliable AI cost forecast has three components that build on each other:

Layer	What It Does	How to Think About It	Business Value
1	**Baseline Cost Model**	Establishes what AI consumption looks like under standard conditions, task by task, across every phase of the software development lifecycle.	Your 'do nothing' cost — the ceiling you're trying to engineer below.
2	**Adjustment Factors**	Multiplies the baseline by real-world architectural variables: agent coordination overhead, model selection, retrieval volume, caching efficiency.	Turns the baseline into an honest projection that reflects your actual architecture.
3	**Scenario Analysis**	Runs three versions of the future — conservative, base case, and worst case — so leadership can see the financial implications of each and make informed decisions.	Converts technical complexity into business decisions with quantified consequences.
D6	**Quality Guardrails**	Every AI output needs to be validated before it is acted on. Validation adds a small cost overhead — but the cost of not validating is far higher. Errors caught late are vastly more expensive than prevention.	At typical error rates, the cost of AI output validation is recovered in less than one hour of prevented remediation work per day.

Scenario Planning: The Governance Imperative

Three scenarios illustrate the range of outcomes at full program scale — drawn from a representative deployment at a leading semiconductor and infrastructure technology company:

Scenario	User Adoption	Governance Maturity	Monthly Cost at Month 6	Monthly Cost at Month 18
Conservative	**Slow ramp**	Full governance, lower adoption	~$285K	~$820K
Base Case	**On-plan ramp**	Full governance, target adoption	~$318K	~$1.1M
Ungoverned / Fast Growth	**Aggressive ramp**	Governance delayed or skipped	~$1.24M	~$3.6M

The difference between the Base Case and the Ungoverned scenario is not the number of users. It is governance maturity. At month 18, that single governance decision is worth roughly $2.5 million per month.

Forecasts Improve Over Time

Early forecasts carry significant uncertainty — typically ±60–80% in the first month of production. This is normal. As real consumption data accumulates, forecasts narrow rapidly: ±20–30% by month 3, ±10–15% by month 6. Executive dashboards should communicate this convergence so leadership can make sound budget decisions at each stage.

05 THE GOVERNANCE ARCHITECTURE

Four Controls That Change Everything

Knowing what drives costs is necessary but not sufficient. The governance architecture is the operational infrastructure that makes cost control automatic. It is grounded in the research finding[8] that AI governance requires purpose-built mechanisms — not adaptations of conventional IT controls.

A well-designed AI governance architecture has four components. None require changes to what the AI does. All operate in the background, invisible to end users.

Governance Component	What It Does	Business Analogy
The AI Gateway	Acts as the traffic controller for all AI model calls. Routes each task to the right model tier, enforces team-level spending limits, attributes every dollar of cost to the workflow or team that spent it, and provides instant visibility into where money is going.	An expense management system that automatically routes purchases to the right cost center and flags overspending before it becomes a problem.
Quality Guardrails	Validates every AI output before it reaches engineers or automated systems. Catches errors, policy violations, and low-quality outputs at the source — before they generate expensive remediation work downstream.	Quality control inspection on a production line. The cost of inspection is a fraction of the cost of shipping defective products.
Workflow Budget Controls	Encodes spending limits directly into automated AI workflows. If a workflow step is about to exceed its budget, it automatically routes to a less expensive model instead of failing or overspending.	Building spending rules directly into procurement workflows — so budget compliance is automatic, not dependent on individual judgment.
Observability & Measurement	Real-time and historical visibility into every token consumed, every model called, and every dollar spent — broken down by team, workflow, and SDLC phase. The data foundation for ongoing forecast refinement.	A financial dashboard giving executives the same visibility into AI spend they have into headcount and software licensing.

The Implementation Sequence Matters

Build the observability and measurement layer first. You cannot govern what you cannot see.
Implement model routing immediately — before onboarding scales. Lowest effort, highest return.
Deploy quality guardrails in audit mode initially, then enforce.
Add workflow budget controls as automated pipelines mature.

06 THE BUSINESS CASE

The ROI of Getting This Right

Executive conversations about AI governance often get framed as a trade-off: invest in controls and slow down, or move fast and accept the risk. This framing is incorrect.

The governance infrastructure described in this paper runs transparently in the background and adds no friction to how engineers work. The question is not whether to build it — it is when.

The Investment

Implementing the full governance architecture requires a one-time implementation investment of approximately $175,000–$200,000 in engineering time and tooling, plus approximately $20,000–$25,000 per month in ongoing operational costs. These figures are consistent with documented AI cost governance implementations.

The Return

At full program scale, the difference between a governed and ungoverned architecture is approximately $1.75–$1.85 million per month in avoided AI infrastructure costs.

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

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

≈ 3 days
payback period
at full program volume

No conventional IT infrastructure investment has a payback period measured in days. As Ismanov et al.[3] document, systematic AI cost governance is not a cost — it is the highest-return investment in the program.

The Risk of Waiting

Every month of governance delay is a month of avoidable overspend. At full scale, a single month without governance costs more than the entire governance implementation investment.

The Governance Timing Rule

Governance infrastructure must be in place before user onboarding scales. The programs that get this wrong are not those that try to govern costs and fail — they are those that defer governance until costs are already out of control, then find themselves imposing restrictions that hurt productivity and trust in the program.

What Good Looks Like: A Phased Approach

Phase	Timeline	Governance Focus	Monthly Cost cted Monthly Cost
Foundation	Months 1–3	Deploy the AI gateway and observability stack. Establish baseline cost measurements. Begin understanding where tokens are being spent before imposing controls.	$420K–$580K (baseline, pre-optimization)
Optimization	Months 4–6	Activate model routing and retrieval governance. Deploy quality guardrails in enforcement mode. First significant cost reductions appear.	$780K–$1.1M (growing user base, first optimizations active)
Full Scale	Months 7–18	All governance controls active. Automated workflows operating under budget constraints. Monthly forecasts accurate to within 10–15%.	$1.5M–$2.0M (fully governed) vs. $3.2M–$3.6M (ungoverned)

07 FIVE DECISIONS EVERY EXECUTIVE MUST OWN

Your Role in AI Cost Governance

AI token economics is a technical discipline — but the decisions that determine success or failure are executive decisions. Research on managing AI at enterprise scale[8,9] consistently identifies executive ownership of governance decisions as the single most important predictor of program cost outcomes. Here are the five that matter most.

The Investment

Implementing the full governance architecture requires a one-time implementation investment of approximately $175,000–$200,000 in engineering time and tooling, plus approximately $20,000–$25,000 per month in ongoing operational costs. These figures are consistent with documented AI cost governance implementations.

#	The Decision	Why It Matters	The Risk of Getting It Wrong
1	Fund the governance infrastructure before the program scales.	The payback period is days, not months. Every month of delay is a month of avoidable overspend.	Programs that defer governance routinely overspend by 2–4× in months 3–6, then impose emergency restrictions that damage adoption and team trust.
2	Require cost attribution from day one.	You cannot govern what you cannot see. Knowing which teams and workflows are spending what is the prerequisite for every other governance action.	Without attribution data, cost reduction efforts are guesswork. Budget conversations with engineering leaders become adversarial rather than data-driven.
3	Set token budgets as ranges, not hard caps.	Budgets set too aggressively block legitimate work. Budgets at 150% of baseline during ramp-up, tightened quarterly as data improves, protect both the budget and the program.	Over-restriction reduces developer productivity, degrades output quality, and causes adoption abandonment — shifting costs from IT to the business unit line.
4	Gate scaling decisions on governance maturity, not just adoption metrics.	Fast user onboarding without mature governance is the single most common cause of AI budget overrun.	Approving 200+ user onboarding before routing, caching, and observability are operational turns a controllable trajectory into an emergency.
5	Treat the forecast as a living document, not a one-time estimate.	AI cost models improve significantly as real data accumulates. A quarterly forecast vs. actual reconciliation is a strategic asset.	Static forecasts diverge from reality quickly. Programs without regular reconciliation lose the ability to make proactive decisions and are permanently reactive.

08 CONCLUSION

The Programs That Win

Enterprise AI has moved past the proof-of-concept stage. The question is no longer whether AI belongs in enterprise software development — it does. The question is which organizations will build their AI programs on a foundation that allows them to scale confidently, and which will spend the next two years reacting to budget surprises.

The organizations that will win are those that treat AI token economics as a first-class strategic concern from the start of the program — not as a technical detail to be addressed after costs become a crisis. This conclusion is consistent with a decade of research on AI business value: Enholm et al.[10] found that AI value realization is mediated by governance capability, not technology selection. Davenport and Ronanki[11] established that operationalization — not innovation — is the central challenge of enterprise AI.

The frameworks and governance architecture described in this paper are not experimental — they are operational, validated approaches applicable to any enterprise multi-agent AI deployment, regardless of industry or technology stack.

The Summary in Three Sentences

AI token costs are architecture-driven, not user-driven. Two decisions — how you route AI model calls and how you govern information retrieval — account for 78% of all achievable savings.

The governance infrastructure that controls these costs pays back in days at full scale. There is no credible reason to defer it.

The programs that build governance before they scale spend 40–50% less, with no capability reduction. That is the standard worth building toward.

Selected References

1. Alvertos, K., et al. (2025). A decision-making framework for integrating generative AI in enterprise workflows. 33rd European Regional ITS Conference.
2. Vial, G., et al. (2023). Managing artificial intelligence projects. Information Systems Journal, 33(3), 669–691.
3. Ismanov, I., et al. (2024). AI and cost management. 2024 International Conference on Knowledge Engineering and Communication Systems. IEEE.
4. Peretz-Andersson, E., et al. (2024). AI implementation in manufacturing SMEs: A resource orchestration approach. International Journal of Information Management, 77, 102781.
5. Hassan, S. Z., et al. (2025). From tokens to tactics: Operationalizing generative AI in enterprise workflows. IEEE ICITEICS 2025.
6. Sorensen, S. (2026). The token industrialist. Human–AI Collaborative Research Series Working Paper.
7. Jannat, S. F., et al. (2024). AI-powered project management. International Journal of Applied Engineering & Technology, 6(1), 1810–1820.
8. Berente, N., et al. (2021). Managing artificial intelligence. MIS Quarterly, 45(3), 1433–1450.
9. Przegalinska, A., et al. (2025). Collaborative AI in the workplace. International Journal of Information Management, 81, 102853.
10. Enholm, I. M., et al. (2022). Artificial intelligence and business value: A literature review. Information Systems Frontiers, 24(5), 1709–1734.
11. Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.

Authors

Scot Whigham
Principal / Infosys Consulting

Karl D. Pfeiffer, Ph.D.
Independent Researcher | JANUS Research Group

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