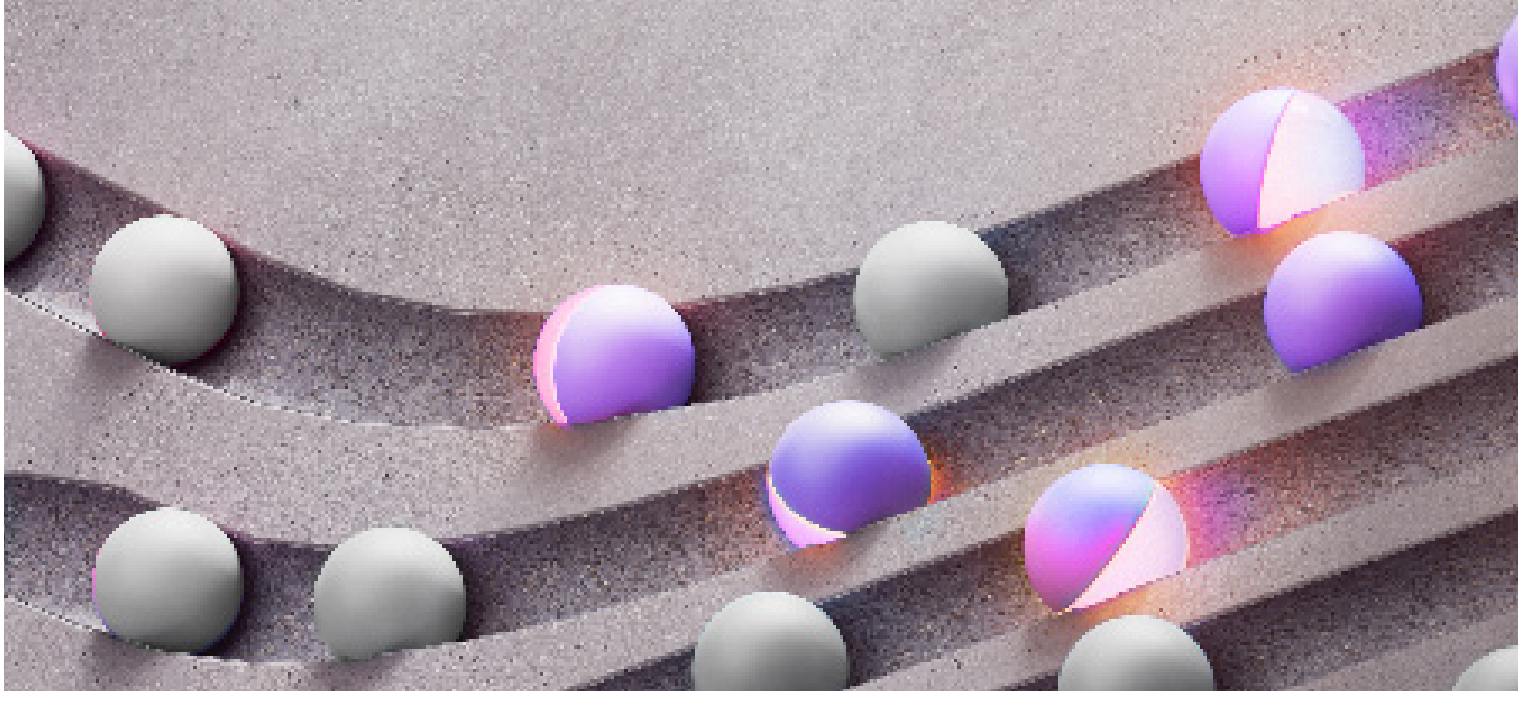
Accenture Research Report

The CIO's guide to AI tokenomics

How to see, control and account for AI token spend at scale



Authors

Ajoy Menon

Senior Managing Director
Digital Core, Lead

in

Lan Guan

Senior Managing Director
Chief AI and Data Officer

in

James Burrows

Senior Managing Director
Digital Core,
Tech Strategy Lead

in

Surya Mukherjee

Principal Director -
Accenture Research

in

Enterprise AI is creating a new cost of doing business: tokens. As consumption scales, token economics is becoming a material risk and opportunity, drawing increasing attention across the C-suite and the board.

Boards are pressing CEOs to explain the exposure and the return, CEOs are looking to CFOs for control, and CFOs are turning to CIOs to make AI consumption predictable, economically disciplined and demonstrably valuable. The companies that master these economics will gain an advantage over those that simply pay the bill.

Our global survey and analysis, coupled with advisory work, revealed that AI's token challenge is both a cost problem and a management problem. The data showed five disciplines that help scale AI usage without breaking the bank. Companies pulling ahead consistently **build observability** into every workload before it reaches production, establish **financial accountability** through chargeback, **route** each task to the lowest-cost model capable of delivering the required outcome, require a **measurable value case** before new workloads go live, and **train employees and developers** with the skills to choose the right models and use AI efficiently. They apply the same rigor to AI tokenomics that they apply to other technological resources such as capital, cloud infrastructure and software assets.

In sum, they manage token spend the way they manage any other cost of production. That puts the CIO at the center of the response, but not alone. Controlling token economics requires the CIO to enlist finance, business leaders, HR and the rest of the C-suite in changing how AI is funded, deployed and used.

What is tokenomics?

The Tokenomics Foundation defines AI tokenomics as the discipline of converting energy and capital into AI capabilities and efficiently consuming that intelligence across the organization, to realize measurable business value.[1] Tokens are the units of data that AI models process. Every prompt, response, retrieval step and agent interaction consumes them. This report and the disciplines it outlines focus on token cost and value.

The payoff is significant. Over a 24-month horizon, companies adopting these disciplines can avoid a projected 44% increase in token costs[2], approximately $4.4 million in avoided cost growth for every $10 million of annual spend (or $44 million at $100 million of spend). The same discipline closes a second gap simultaneously: today only 20 cents of every dollar spent on tokens can be expressed as a quantified financial outcome to a board or CFO. Organizations that implement chargeback and workload-level measurement raise that to 32 cents, making approximately $1.15 million per $10 million of spend provable for the first time. Together, that is **$5.6 million of combined dividend in cost avoided and value proved, per $10 million of annual spend (or $55 million at $100 million),** all without deploying new models or increasing AI investment.

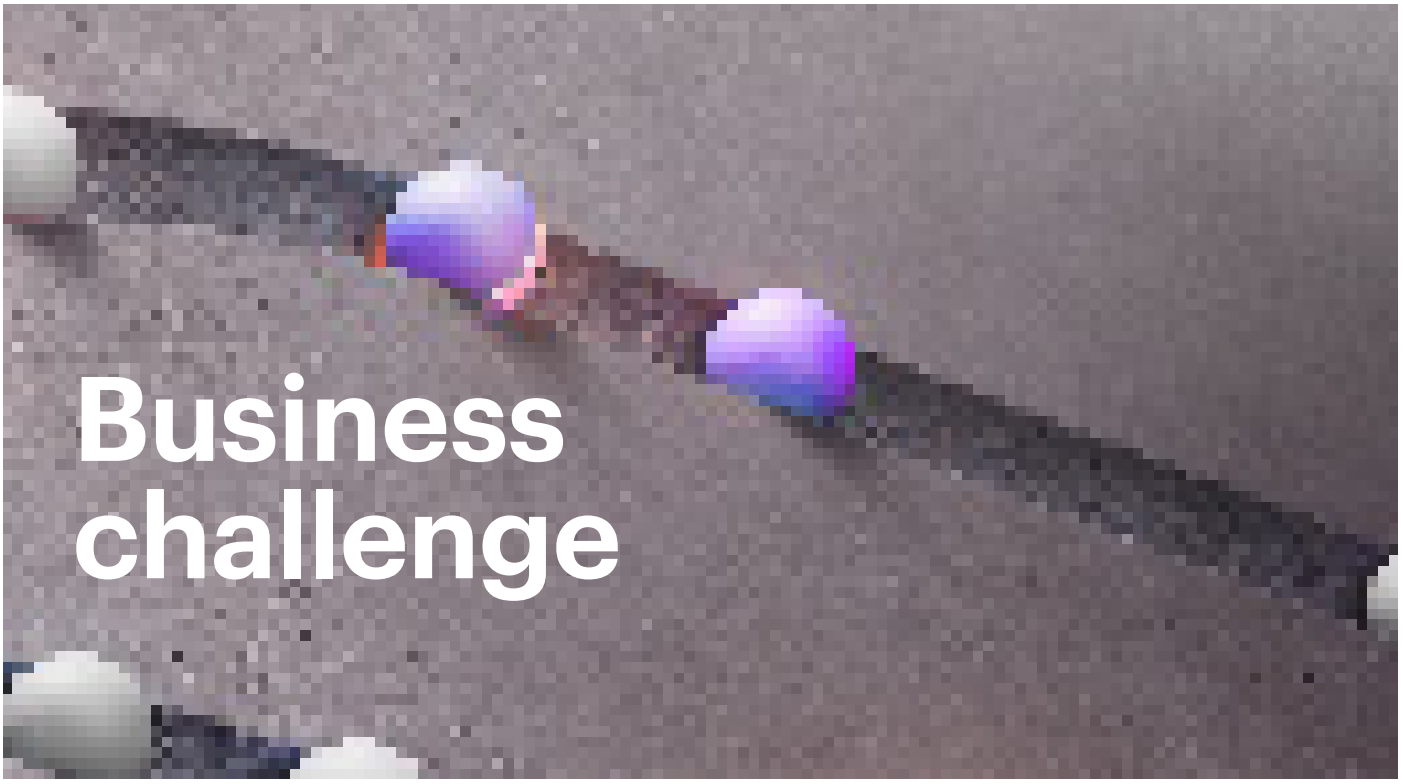
Other companies still manage AI the way they managed enterprise software a decade ago: They wait for the bill. That breaks down fast now that spending has shifted from fixed software licensing to usage-based AI, and consumption is climbing across employees, agents and workflows.

As token costs rise, boards are increasingly asking CEOs three questions: What is driving AI costs? How fast will those costs grow? And what business value is AI creating? Many CIOs cannot yet give their CFOs, CEOs and boards answers with confidence, largely because they lack visibility into the workloads generating token consumption, the models driving it and the business outcomes being created. Just as important, CIOs often explain the problem in technical language rather than the financial language CFOs need to equip CEOs for the boardroom. That makes it harder to attribute investment, manage risk and scale AI with confidence.

Companies that adopt the five disciplines treat every token as a production input to manage before they spend it, and not an invoice to review after the fact. Combining the five disciplines gives CIOs the rigor needed to scale enterprise AI while keeping costs under control and proving business value.

Research snapshot

Our July 2026 survey of 750 senior executives from enterprises with annual revenues exceeding $1 billion across 17 countries is coupled with 15 in-depth interviews conducted with Fortune 500 technology and finance leaders between June and July 2026. Respondents evaluated how their organizations manage AI consumption, optimization, routing, accountability and value measurement, providing insight into both the economics of AI deployment and the mechanisms used to govern it at scale. We use generative AI in our research production process. Our research experts review and validate the generative AI outputs with traditional research methods where possible, applying Accenture's Responsible AI standards.



Business challenge

The CIO's accountability gap

The enterprises in our survey spent about $2.5 billion on AI tokens last year in total. Without intervention, this figure climbs toward $3.6 billion in 24 months.[3] The aggregate spans a wide range: a global technology company spending north of $20 million annually, a European insurer with an AI inference budget of $150-200 million, and a government contractor with a global AI spend ceiling of $250 million. Because spending varies significantly by deployment stage and organizational scale, our analysis is expressed per $10 million in annual token spending rather than anchored to any single baseline. If your organization spends $30 million, multiply by three. If it spends $300 million, multiply by 30. AI token costs rank third among all AI cost drivers, behind infrastructure and software development and maintenance.

At today's levels, token spend alone may not command board attention, but the trajectory increasingly does. As consumption accelerates and forecasts become less reliable, CFOs face a cost category that can move from an operating expense to a material financial risk surprisingly quickly. That pressure travels upward as CFOs must explain the exposure to CEOs and they, in turn, must increasingly explain it to their boards.

Their concern ultimately comes down to two gaps: token volume and spend cannot be predicted quarter to quarter, and the return cannot be explained once the money is spent. These two create a double bind. When costs are unpredictable, finance hesitates to approve larger AI budgets. When value is invisible, leaders cannot defend their requests for AI budgets. That leaves the CIO with a new mandate from finance: make the economics of AI visible and manageable before the cost curve becomes a board-level problem.

The trajectory is the real concern. Companies in our survey expect token consumption to grow 78% and per token price to decline by 19% over the next 24 months, on average.

The alarm is far from irrational.

Uber spent its entire annual AI token budget in four months.[4] A Chief Technology Officer overseeing cloud, platform and security architecture at a global financial data and media company put it plainly:

> **"It's a real problem ... growing by a minute, by a day ... keeping track of [token volume and spend] is definitely a challenge, especially if it's distributed across multiple organizations, multiple teams."**

A technology leader at a major US financial institution described a 5-10x increase in six months. At the lower end of that range, a $10 million annual base becomes a $50-100 million annual run-rate within a year before any price decline offsets are realized. Similarly, a large European bank moved from near-zero to €5 million ($5.8 million) in annual token spend in a year, with a single workload running at €200,000 (~$231,000) a week before anyone intervened.

One in three organizations has already exhausted its annual token budget before the year ended. **The token bill will not stabilize on its own, and approaching AI spend like this is no longer sustainable** (Figure 1).

Figure 1: Three growth scenarios show why the token bill will not stabilize on its own

The token bill will not stabilize on its own

Every $10 million spent on tokens today....

Current optimization: 23%
Break-even threshold: 31%

Scenario	Annual run-rate and growth logic	Management implication
1. Managed growth Closes the 8-point optimization gap to 31% break-even threshold	**$10M - held flat** Token volume +78% over 24 months Prices decline 19% Optimization reaches 31%	Growth can be absorbed, but only if routing, caching and prompting discipline improve before consumption scales.
2. Unmanaged growth No additional optimization	**$14.4M within 24 months** Token volume +78% over 24 months Prices decline 19% No additional optimization	Adds approximately $4.4M to the annual token bill despite falling unit prices — volume growth outpaces price declines.
3. Breakout growth 5–10x surge in six months	**$50-100M within a year** Based on rapid adoption pattern reported by a major US financial institution	Traditional annual budgeting fails. Workload-level guardrails and intervention thresholds must operate continuously.

Source: Accenture Tokenomics Survey (N=750)

Note: The cost curve is a management choice. Managed growth holds the annual run-rate near today's level per $10 million of spend; unmanaged growth pushes it toward $14.4 million; breakout adoption can lift it beyond $50 million before price declines provide meaningful relief. Scenarios 1 and 2 use survey-based assumptions. Scenario 3 is an interview-derived stress case, not a forecast.

Token consumption originates across three broad categories of work: employee and developer tools, which account for around 30% of spend; internal process automation and enterprise software, at roughly 29%; and customer-facing applications, analytics and agentic workflows, at around 39%. Agentic workflows, currently the smallest slice at 11%, multiply inference calls with every step and typically grow the fastest, making them the category most likely to drive the next wave of unbudgeted cost growth.

The token problem has two faces

Both problems set off the same alarm. Even though they look like two failures with two fixes, they both come from one blind spot: enterprises can't see token consumption at the workload level. Spend that isn't tied to a workload can't be predicted, and value that isn't tied to a workload can't be proven.

1. The bill nobody can forecast

Organizations are not reacting to the actual dollar value of token spend. They are reacting to the inability to forecast it, govern it or explain it when the bill arrives. Volume is growing faster than prices are falling, and the gap between them is widening. Budget models built before agentic tools became mainstream cannot keep up.

The pattern is already visible at the frontier of adoption. A senior engineering leader at a major Asian bank we spoke to switched on an AI coding assistant for his developers. The logic was sound: productivity gains will more than pay for the tokens. Three months later, the bill arrived, well beyond the forecast. Thousands of engineers had reached for the most powerful model when a less expensive one would have done the same job. Nobody did anything wrong. The budget process had no way to see it coming.

2. The value nobody can prove

The return from AI investment exists. Enterprises are generating real productivity gains, accelerating decisions and improving customer outcomes. What they cannot do is connect those returns to token spend in financial terms a CFO can act on, a CEO can justify and a board can evaluate.

Only one dollar in five of enterprise token spend shows up as a quantified financial outcome, whether that is revenue influenced, cost avoided or productivity converted to a dollar figure. The remaining four dollars sit in a space companies believe is productive but cannot prove. The value is there. What's missing is the built-in tracking that would tie it to a dollar figure.

Even among companies that are ahead of this problem, value attribution remains elusive. A director of AI strategy at a Fortune 500 financial services institution described a situation familiar to many enterprises: AI is clearly creating value across the organization, but that value is not being measured against the tokens consumed to produce it. Productivity gains are visible, outcomes are improving and adoption is expanding, yet the financial link between consumption and impact remains largely unquantified.

The way out of these problems is to fix the root cause with workload-level visibility, ownership and financial accountability.

Falling AI prices are fueling a spending boom

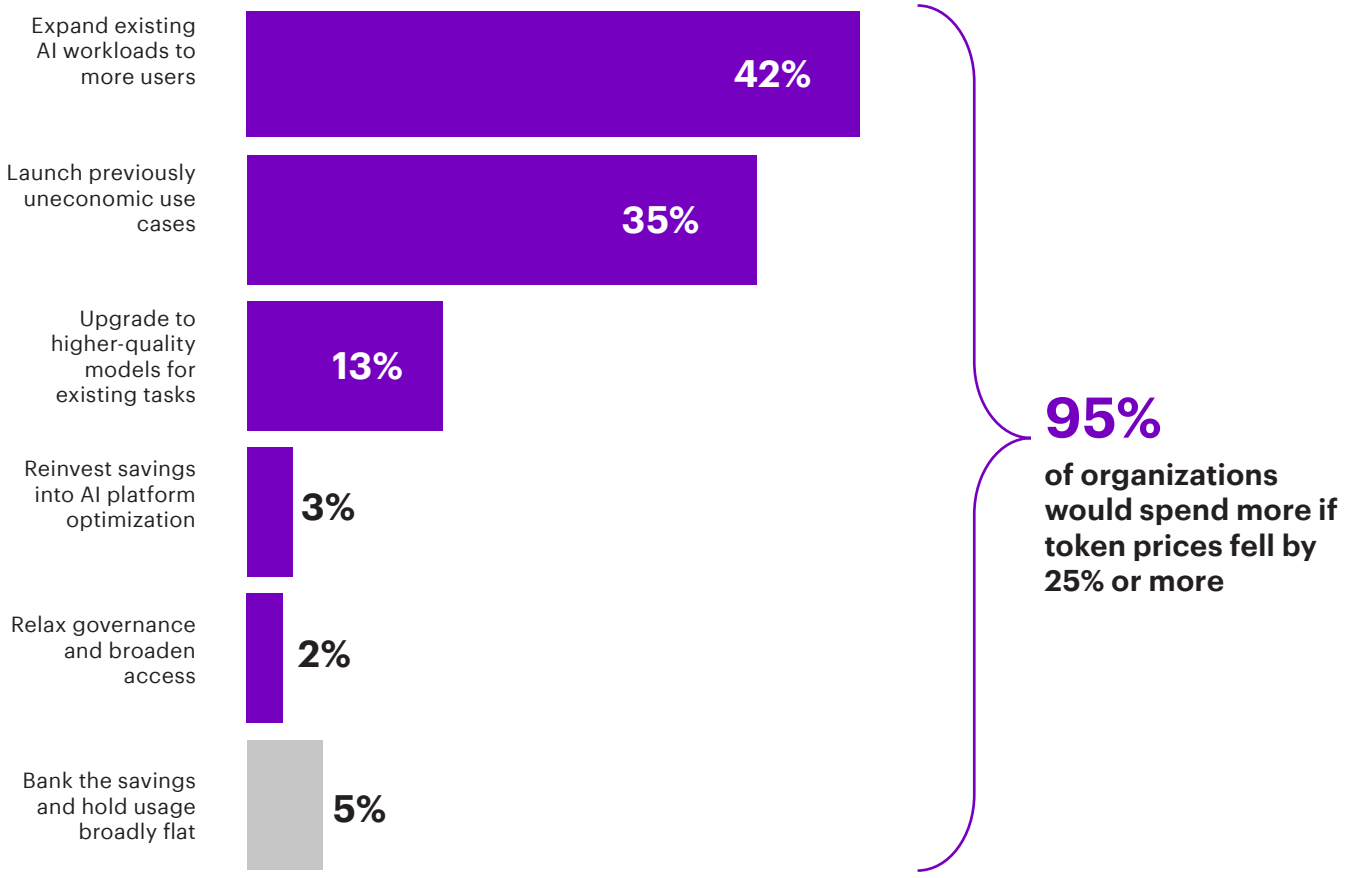
The paradox: every saving funds more consumption

Per-token prices are falling. A model that cost a dollar per million tokens 18 months ago costs a fraction of that today. Logic suggests enterprise AI bills should be shrinking. They are not.

This is the tokenomics logic based on the Jevons paradox: just as more efficient steam engines expanded Britain's coal use rather than cut it, falling token prices push companies to run more AI, not less (Figure 2). When asked what they would do if prices fell a further 25%, only one in 20 executives said they would bank the savings. The rest would expand: more users, more use cases, better models.

Figure 2: Response to falling token prices

Q : If token prices fell by 25% or more, what would your organization most likely do?



Source: Accenture Tokenomics Survey (N=750)

A Field CTO for AI, cybersecurity and data at a major global technology infrastructure company described this dynamic precisely:

> **"Even if the models become cheaper, teams just use them more. It's like ... you got your first gigabyte hard drive back in the day? And it wasn't like, 'Oh my God, we're going to stop using ... this hard drive space.' No, it just grew exponentially because it's the same exact thing."**

Work that was once too expensive to justify suddenly makes financial sense. New use cases open up. Agents that would have been prohibited at last year's prices get deployed at this year's rate. The volume response swamps the price decline, and the bill climbs.

Harvey, a San Francisco-based legal AI company serving more than 100,000 lawyers across more than 1,300 law firms and corporate legal teams, saw monthly token consumption rise from approximately one trillion tokens in January 2026 to a pace of 12 to 13 trillion by May.[5]

The increase shows how quickly consumption can scale once AI moves deeper into production, and why even substantial declines in unit cost can be overwhelmed by volume growth.

Six blind spots that break the budget

The true cost of tokens is hidden in six places your finance team can't see (Figure 3).

Figure 3: Six blind spots behind AI budget overruns

1. Unknowable task cost

Tokens have no intuitive human equivalent; cost emerges only after tasks run.

2. Frontier by default

People choose the most capable model even when lower-cost models are sufficient.

3. Agentic multiplier

Chained calls, expanding context and repeated processing multiply consumption.

4. Budgets built for yesterday

Plans predate mainstream agentic tools and usage growth.

5. Fragmented pricing

Multiple providers, models and pricing structures prevent a consolidated view.

6. Visibility without accountability

Showback dominates; limited chargeback weakens cost ownership.

Unpredictable demand + weak governance = recurring budget overruns

Unknowable task cost

Nearly every other cost you manage announces itself before you commit. A software license, a server, a new hire: you know the number upfront. AI token consumption doesn't work that way. One sentence of instruction can trigger a chain of agent actions that burns millions of tokens, and there is no preview. You have no gut sense of whether a task will cost one dollar or a hundred.

Frontier by default

Your people reach for the most capable model available, partly because they can't see the cost difference and partly because no one has given them a reason or incentive to choose differently.

Accenture classified 9,368 occupational tasks from the O*NET database and found that fewer than 10% of these workloads genuinely require frontier model capability. The remaining 90% are being processed by models costing 10 to 20 times more per token than the task demands. As a VP of Enterprise AI Strategy at a major global technology company observed:

> **"10% of people are pounding the table ... [saying] 'Why don't I have the latest model?' And then you have basically [the] majority of an organization using this as search."**

Agentic multiplier

Agentic workflows increase token consumption. A single request can trigger multiple model calls, larger context windows and repeated processing of the same information. As workflows become more complex, the cost curve can outpace usage growth, making traditional budgets increasingly unreliable.

One company rolled out an AI coding assistant at scale and watched coding costs escalate faster than expected. When they investigated, the fix turned out to lie in architecture rather than pricing. The biggest savings came from moving developers from a premium model to a lower-cost model that was sufficient for most coding tasks, and from enabling prompt caching so large code repositories no longer had to be reread on every interaction.

The model change alone reduced token consumption by roughly 90% while caching delivered further savings. AI costs are often determined less by model pricing than by model selection, context management and architectural design. User education and architectural guidance are essential for cost management.

Budgets built for yesterday

Most enterprise AI budgets were constructed before agentic tools became mainstream. As a result, one in three organizations hit their annual token budget before year-end. A Chief Technology & Information Officer at a global quick-service restaurant chain described the experience:

> **"Early in, we didn't really understand what the budget would be. ... Some groups were definitely blowing through the budget. It was going two to three times over budget."**

Fragmented pricing

Most large enterprises run models from different providers at once, each with its own pricing, context window mechanics, caching policies and output-token rates. No one has a consolidated view. Each new model release, often with little advance notice, breaks the cost assumptions set for the previously available models. Pricing is fragmenting as fast as the models themselves.

A senior technology leader at a European bank told us vendors increasingly price AI by the business value of the use case, not the volume of tokens consumed.

A contact center solution could cost roughly 10 times more than a general-purpose AI platform, and specialist legal tools were priced many times higher because providers benchmark against the salaries of the professionals they replace. As pricing shifts from token consumption to use-case economics, similar workloads can produce radically different bills, potentially making forecasting simpler at the level of an individual provider, but more complex across the portfolio and more expensive overall.

Commercial pricing design can create sudden step changes in cost that budgets cannot anticipate. Pylon, a B2B customer support and agentic AI company, discovered that its annual token bill was set to rise from roughly $400,000 to $1.4 million as the company crossed a 150-seat threshold and moved to enterprise pricing. Finance teams need to model vendor pricing tiers and thresholds in advance, rather than discover them at renewal.[6]

Visibility without accountability

Token balances are typically buried in provider dashboards that finance teams and developers can’t access and admins do not check in real time. What we call showback means consumption data is visible but there is no financial consequence. Chargeback means consuming teams are actually billed for what they use, creating a direct financial incentive to manage consumption. Today, just over half of organizations have showback, and only 7% have any form of chargeback. Showback creates awareness that is useful when coupled with metrics like KPIs. Chargeback creates direct accountability but may still need to be tied to additional incentives to drive behavioral change.

Four of every five dollars lack a quantified link to business outcomes

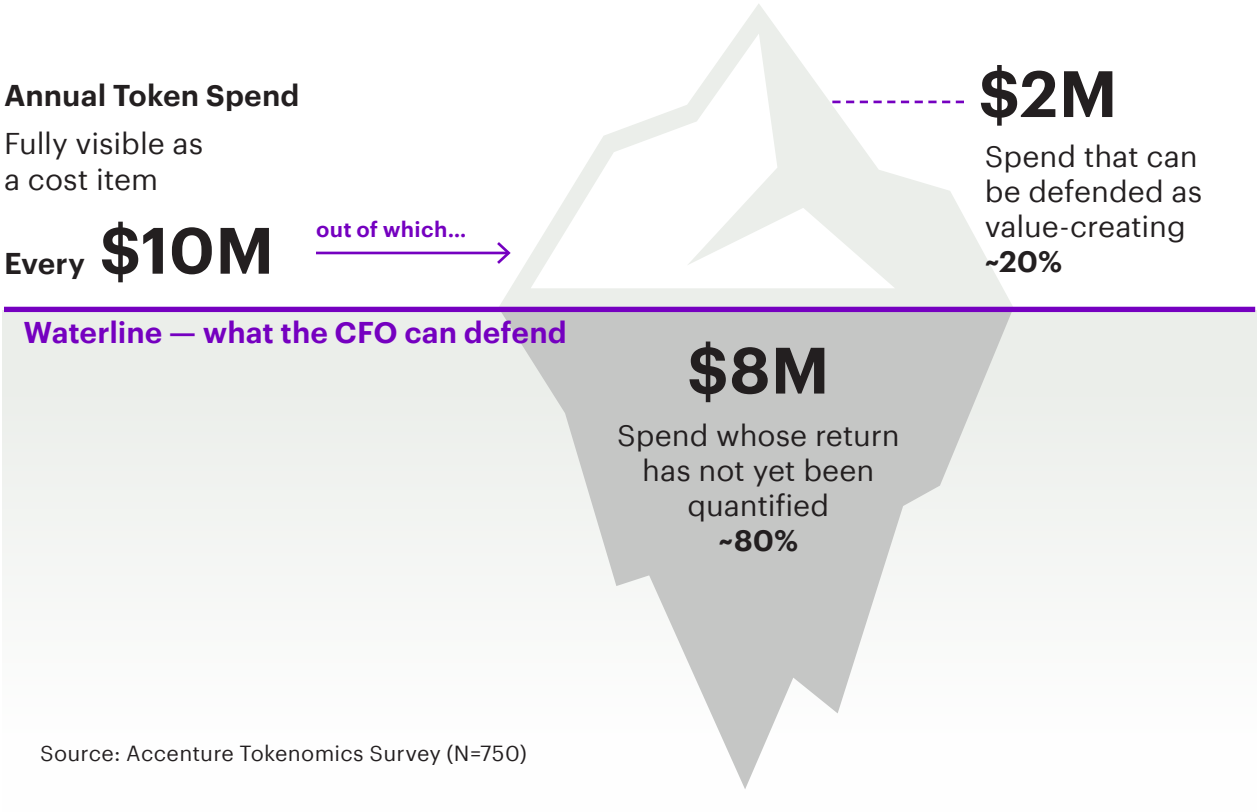
Finance can quantify the return on fewer than one in five dollars spent

Ask executives to trace their token spending to a quantified financial outcome—whether revenue influenced, cost avoided or productivity expressed in dollars—and fewer than one in five dollars clears that bar. For every $10 million spent, organizations cannot translate the return on $8 million into the financial language CFOs need to manage the investment, CEOs need to justify it and boards increasingly expect to see (Figure 4).

Figure 4: Only a fraction of AI's return survives the journey to the boardroom

CFOs can link fewer than $2 in every $10 spent to a business outcome

The CFO sees every dollar spent but can defend the return on under one in five



Source: Accenture Tokenomics Survey (N=750)

AI creates value where finance cannot see it

The cost side of AI is cleanly decomposable: tokens are 25% or less of total AI spend per our survey—identifiable, traceable and attributable to specific workloads and teams.

The value side compounds across a whole workflow: a productivity gain created by an AI-assisted workflow reflects the model, the data it retrieved, the infrastructure running it, the engineer who designed the workflow and the employee who used it. No amount of careful measurement will isolate the share of the realized value that belongs to tokens alone.

A Field CTO for AI, cybersecurity and data at a major global technology infrastructure company articulated the core question:

> **"The economic question is not ... how many [tokens] were consumed, but what did that token actually do?"**

Forty-eight percent of organizations rely on shared IT-finance accountability, with no single individual responsible for AI cost and return. Technology teams can see which function or person is consuming tokens, and at what cost. What they cannot determine is whether that consumption produced anything worth paying for. That judgment belongs to the business unit leader who understands the workflow, the customer interaction or the business decision the AI supported. Most have never been formally asked to make that judgment.

Only 35% of organizations can calculate cost per business outcome even for their single largest use case. Fifty-nine percent are working from directional estimates. A director of AI strategy at a Fortune 500 financial services institution said:

> **"I would say the majority of what we are doing now is actually value creating. But I wouldn't say we are looking at tokens to derive that."**

The data on what drives financially attributable value is unambiguous and actionable (Figure 5).

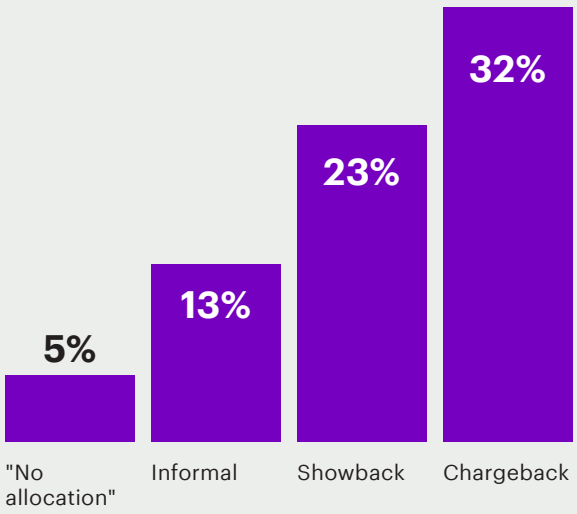
Figure 5: Accountability, routing and visibility make AI value provable

Three signals drive financially attributable value

Financial accountability is the strongest signal; attribution climbs at every step.

Financial accountability

Strongest signal



Financially attributed value by accountability level

Routing quality

Same investment

The same telemetry routes intelligently and attributes outcomes.

1 build, not 2

Workload visibility

The foundation

~2x higher attribution when consumption is traced to a workload. You cannot route or attribute what you cannot see.

Source: Accenture Tokenomics Survey (N=750)

Financial accountability

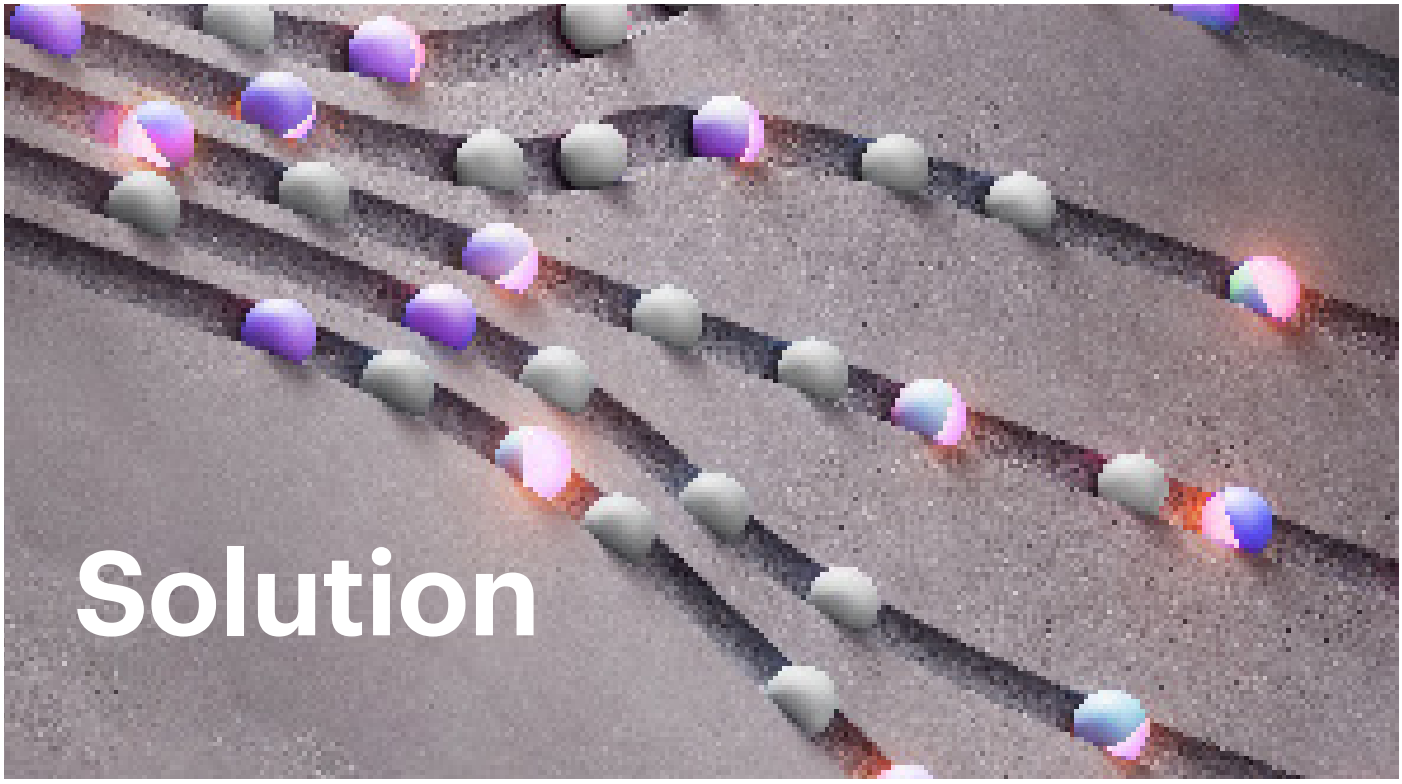
This is one of the strongest signals. Each step up the accountability ladder, from no-cost allocation to full chargeback, adds approximately six to seven percentage points to financially attributed spend, regardless of maturity, size or industry. Financially attributed value runs from about 5% under no-cost allocation to 13% with informal reporting, 23% with showback and 32% with chargeback.

Routing quality

The same observability layer that lets you route intelligently—knowing which workload ran, on which model, at what cost and to what result—is essential for attributing outcomes. Routing discipline and value attribution are parts of the same investment. You don't build one first and the other later.

Workload visibility

Organizations that can trace token consumption to a specific user, team or workload score nearly twice as high on compound attribution as those that cannot. Without that visibility, neither routing nor financial accountability can work. You cannot route what you cannot see, and you will not be able to attribute value to a workload you don't have the observability layer for.



Solution

One build, $5.6 million in dividends per $10 million of spend

Our research identifies four independent predictors of both cost optimization and value attribution: named ownership at the workload level, chargeback, measurement precision and model routing discipline. Each predictor moves cost and value independently, but all four depend on the same foundation: workload-level ownership, observability and accountability. Build that once, and you improve both sides of the equation.

Named ownership at the workload level enables cost predictability and value measurement. Without it, costs are anonymous and value remains unattributed. Chargeback creates the incentive to manage cost precisely and to prove value. The financial accountability that drives cost consciousness also drives the motivation to demonstrate return. Measurement precision reduces forecast error, surfaces the workloads where AI is generating financial return and builds the evidence base for the next investment cycle. Model routing discipline is both a cost optimization lever and a prerequisite for workload-level attribution.

A CTO overseeing cloud, platform and security architecture at a global financial data and media company described what this architecture produces in practice:

"We just [agreed on] attribution down to the org level, down to the team, and in most cases, we can attribute to an individual. And we designed it that way because essentially, we wanted to make sure that the chargeback model was set from the beginning."

The size of the prize if you get it right

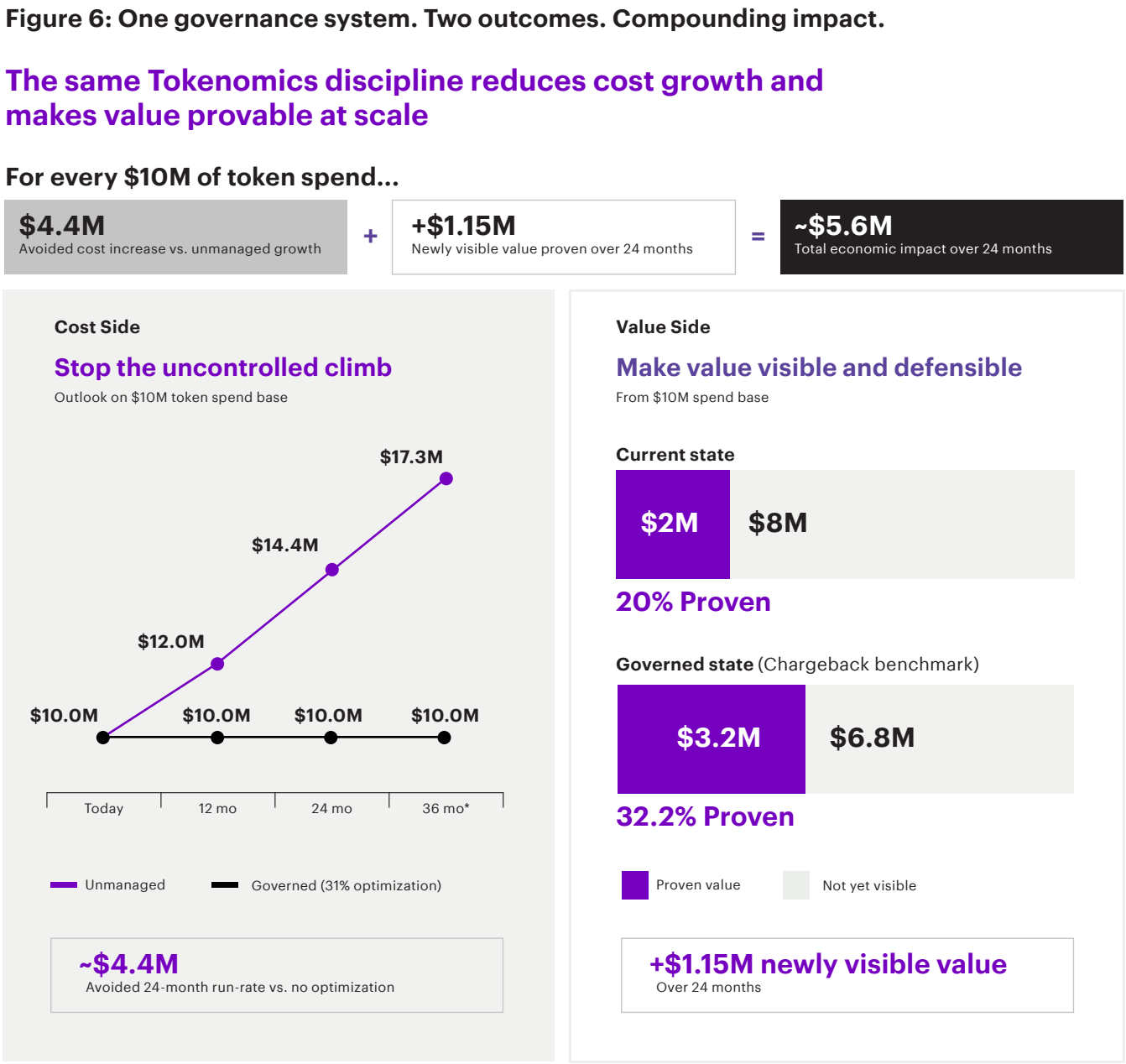
Most organizations have optimized 23% of token consumption volume today. Our model shows they would need roughly 31% optimization to hold spend flat as volume grows nearly 80% over the next 24 months, leaving an eight-point gap. Do nothing and every $10 million of annual spend climbs toward $14.4 million over 24 months, and toward $17.3 million over 36 months.

The value side comes from the same build. Today, only one dollar in five of token spend can be translated into a financial return. With full chargeback and precise measurement, provable value climbs to around 32 cents on every dollar spent. On every $10 million of spend, that shift makes approximately $1.15 million of value provable for the first time. For the CFO, that is the problem the CIO must solve: not simply reducing the token bill but creating the instrumentation and operating discipline that makes both cost and value financially legible.

One governance build, two results: approximately $4.4 million in avoided cost growth and $1.15 million in newly visible value, per $10 million of annual spend. Both come from instrumenting the work once, with named ownership and financial accountability at the workload level, and applying it to cost and value at the same time.

The organizations that build it start speaking in terms of returns, not just costs (Figure 6).

Figure 6: One governance system. Two outcomes. Compounding impact.



Source: Accenture Tokenomics Survey (N=750)

Assumptions: Token consumption growth: +78.41%; Token price change: -18.87% (rounded to 19%); Break-even optimization required: ~31%; Current optimization level: ~23% ~8 pt gap to break-even; 36-month unmanaged spend extrapolated beyond survey horizon; N=750 global senior executives with revenue >$1B. *Illustrative extrapolation using same growth and price assumptions beyond 24 months.

Combining open-weight and proprietary models can control costs for some

So far, we have focused on the model most enterprises use today: paying per token through a provider API or cloud platform. That pattern covers 70% of enterprises. But deployment architecture changes the bill more than most CIOs expect. Options range from open-weight models on CPU or self-hosted GPU to open-weight models on public cloud and proprietary models via cloud platform or direct API. Each has a different cost structure and different trade-offs in speed, control, quality and infrastructure burden (Figure 7).

Figure 7: Different workloads need different cost structures

Deployment model	Cost structure	Best suited to	Key trade-offs
Open-weight on CPU	Very low inference cost; higher infrastructure, talent and operations costs	High-volume, low-complexity tasks: classification, summarization, extraction	Slower; significant MLOps overhead
Open-weight on GPU-self-hosted	Low inference cost at scale; higher infrastructure, talent and operations costs	High-volume commodity workloads at scale	High upfront capital; full infrastructure responsibility
Open-weight on public cloud	Low inference cost; additional cloud infrastructure and management costs	Organizations wanting open-weight cost structure without infrastructure burden	Fastest-growing pattern; less control than on-premises
Proprietary via cloud platform	Per-token pricing; infrastructure and operations included	Most production workloads	Dominant enterprise pattern; frontier capability; enterprise SLAs
Proprietary direct from model provider	Per-token pricing; infrastructure and operations included	Direct API access	Functionally equivalent to cloud platform for most purposes

Source: Accenture analysis

For the roughly 90% of tasks that don't require frontier reasoning, open-weight deployment looks compelling on cost. The data reinforces this selectively: organizations running a roughly equal mix of proprietary and open-weight models show higher routing quality (55% of requests correctly routed, versus 47%) and greater deployment maturity (61% running AI in production at scale, versus 32%) than predominantly proprietary organizations.

Beyond cost, open-weight models offer three further advantages: data sovereignty, resilience against export controls on frontier models, and stronger performance on non-English, domain-specific tasks when fine-tuned.

The limits matter equally. Token spending is 25% or less of total AI expenditure; shifting to self-hosted open-weight models often raises infrastructure, talent and operational costs elsewhere. Commercial support requirements and in-house capability gaps have kept enterprise adoption selective—though both are changing. Emerging open-weight models and entry-level infrastructure can now handle most commodity enterprise workloads capably.

Overall, open-weight models are a valuable option in the routing toolkit, but rarely the dominant architecture at enterprise scale per our survey. For organizations without a dedicated team and the operational maturity to support them, the gains on the token line can be more than offset by other costs mentioned above.

The right position is a deliberate continuum: route each workload to the model tier that best balances cost, capability, data governance and operational complexity.

Next steps

Five disciplines for the CIO

Our research and client work identified five disciplines: workload-level visibility, financial accountability, smarter routing, proof of value and AI-skilled people. Together, they add up to a single operating model for running AI at scale. These disciplines are not sequential steps; they work together as one operating model. Visibility comes first because the other four disciplines depend on it, while the remaining four can be implemented in parallel or sequenced to match the organization's AI maturity.

- **Discipline 1:**
  Make every workload observable before it scales

  Deploy an AI gateway or observability layer that captures, for every AI interaction: the workload name, the owning team, the model used, the token cost and the output. At scale, this requires policy-as-code and automated economic guardrails, not manual review of individual workloads. It is the prerequisite for both cost management and value attribution.

Tools including LangSmith, Arize and Helicone provide workload-level cost attribution at modest overhead.[7]

The pattern of building this too late is now well-documented. Uber encouraged engineers to use AI "as much as possible" and ran internal leaderboards for token consumption. By April 2026 the company had burned through its entire annual AI budget in four months. The company then introduced limits of $1,500 per employee per month.[8]

Shutterstock shows why early instrumentation matters. As the company examined its AI cost base, it found that output tokens accounted for 75% of total token consumption, while ChatGPT licenses had expanded from 50 to 750 in a single month. It also identified $250,000 of committed AI spend at risk of going unused. The response was not simply to restrict AI use. Instead, Shutterstock expanded its FinOps remit to cover AI and cloud costs, bringing the data into a common process with finance before adoption expanded further. The result was earlier visibility into AI consumption and vendor commitments, allowing potential waste to be addressed before adoption scaled further.[9]

The technology executive from a global financial data and media company interviewed for this research described the payoff of building it early: attribution traceable to the organization, the team, and in most cases the individual, with the chargeback model set from the outset rather than retrofitted.

▪ Discipline 2:

Make teams pay for what they use

Bill consuming teams for their token costs. Start with engineering and the top three or four consuming functions. Showback without financial consequence changes awareness but only changes behavior when coupled with other incentives. In our research, chargeback is one of the strongest predictors of cost optimization and value attribution, but fewer than one in 12 organizations use it today.

The mechanism is direct. When a team pays for its own consumption, managers ask whether the spend is justified, and engineers start selecting models with cost in mind. Organizations that implement chargeback tend to be more advanced in governance and observability overall. The mechanism works best when it accompanies workload-level visibility and clear ownership rather than preceding them. In a less mature organization, chargeback alone is unlikely to produce the same results. The field CTO at a major global technology infrastructure company described an operating model of centralized standards and federated accountability: each production workload is expected to have a named business-unit sponsor responsible for explaining the intended outcome, success measures and economics as it scales.

Discipline 3:

Send each task to the model it actually needs

Identify the approximately 10% of workloads or process steps that genuinely need frontier capability—the complex reasoning, planning, multi-source synthesis and high-stakes generation—and route everything else to capable mid-tier or open-weight models. Enforce this as a routing policy at the gateway, not as guidance left to each developer's judgment.

The price spread is significant: a frontier model can cost 15 times or more as much as a capable lower-tier model. RouteLLM, a UC Berkeley-led open-source framework published at ICLR 2025, demonstrated up to 85% cost reduction on benchmark workloads while preserving 95% of flagship quality.[10]

The VP of enterprise AI strategy at a major global technology company put the problem plainly: most enterprise AI usage is still search and retrieval, while a smaller group of power users demands advanced models for more complex work. Sending both patterns through the same frontier model is how avoidable cost piles up.

The CTIO at a global quick-service restaurant chain described the same issue operationally: users initially defaulted to the most expensive frontier model for even menial tasks because they had not yet learned when lower-cost models were sufficient. A central gateway with dynamic model routing is what prevents that default behavior from turning into runaway consumption.

Accenture's own production platform processes approximately 8.7 trillion tokens per week at roughly one-sixth of frontier API cost through a combination of self-hosted open weight inference and intelligent routing.[11] The market is moving in the same direction. According to IDC's 2026 AI and Automation FutureScape, by 2028 70% of top AI-driven enterprises will use advanced multi-tool architectures to dynamically and autonomously manage model routing across diverse models.[12] Vercel's 2026 production data shows the same pattern already emerging at scale, with production teams running an average of 35 models as a routing graph. What was once an advanced practice is quickly becoming the operating baseline.[13]

How to match the right model tier to the right task

As mentioned earlier, frontier models shouldn't be the default. Accenture ran 9,368 occupational tasks from the O*NET database through five tests of whether a task needs frontier-class AI, after stripping out physical work and anything requiring a human review and sign-off. Only 10% made the cut. The remaining 90% run just fine on a capable mid-tier model or need so little judgment the model barely matters.

The 90% is not simple work. It includes evaluating network designs, drafting contracts from established templates, troubleshooting security incidents and analyzing data against known criteria.

These tasks take real expertise. What they lack is irreducibility, the characteristic that separates frontier from the rest.

A task is irreducible when no established method resolves it, when one early error silently corrupts everything downstream, when the picture breaks if you divide the work, when it forces together two distinct professional domains or when its value lives in the coherence of the whole. Absent all five, a capable mid-tier model suffices, however demanding the work looks.

The cost consequence is direct. Frontier models cost 10 to 20 times more per token than capable mid-tier alternatives. Already, 54% of AI requests are being routed to a tier higher than the task needs. Nearly one in four executives names model routing as the single biggest efficiency lever available, and they are right. Production teams report 40% to 60% reductions in model-serving bills when they route requests to the lowest-cost model capable of doing the job.

Discipline 4:

Demand a value case and a management standard before any workload reaches production

Define three things before you commit capital to any new AI workload: what the process costs today, in money or time; what specific financial outcome AI will improve; and how that outcome will be measured in dollars. Then hold the gate. Approve the business case only when observability tooling is in place to measure consumption at the workload level, optimization is on a credible path toward break-even, and the expected value can be expressed as a quantified financial outcome, whether revenue, cost avoidance or productivity in dollars.

This does not need to be complex. It needs to exist before the spend, not while trying to justify it afterward.

Match Group provides a clear example of the shift from experimentation to a value gate. After previously allowing AI pilots a "more or less unlimited budget," CFO Steve Bailey said the company raised the bar for 2026. Material AI spending now requires a business case showing a clear impact through cost savings, efficiency, revenue growth or user impact. Experimentation can start with flexibility; scaling requires a return that finance can evaluate.[14]

ABN AMRO has made this a formal gate. The bank requires every AI use case to contribute to at least one of five defined value outcomes before it reaches production. As of May 2026, the bank had 40 use cases in production, each pre-mapped to a value outcome. The results are specific: its ICS Voice Bot handles 80% of roughly 60,000 monthly calls with a documented 500-hours-per-week efficiency gain; its Delphi risk assessment tool saves approximately 800 hours per month in financial crime operations. Those numbers were defined before deployment and measured after.[15]

Klarna took a similar approach in its customer support deployment. It built outcome measurement into the architecture from the start, tracking resolution time, repeat contacts and deflection rate, and reported $40 million in avoided hiring costs at launch in 2024, growing to approximately $60 million by Q3 2025.[16] The team could measure and attribute usage because that visibility was built in from the start.

The contrast between organizations that set this standard early and those that did not is now visible in public results. At Uber, where AI agents now contribute roughly 10% of committed code, the COO recently acknowledged that it's still hard to show exactly how token use translates into better products.[17] The spend arrived before the measurement infrastructure. Uber is now instituting the discipline that closes this gap—$1,500 per employee per month, internal dashboards and approval for overages. Its experience underscores a broader lesson: The enterprises that fare best put that measurement infrastructure in place alongside access, not after the first budget surprise.

The field CTO at a major global technology infrastructure company described the same sequence playing out in every organization that scales without this discipline in place first. The platform team ends up preparing reports that explain costs no one built the observability for and justify value that was not quantified. A business case that cannot express its expected value in financial terms before launch is not a business case. It is an experiment. Fund it like one, with explicit scope limits and a defined point at which it produces evidence or stops.

The organizations in this research that cannot demonstrate financial return are not, in most cases, failing to generate it. They are failing to measure it. Discipline 4 is the organizational commitment to build those mechanisms before the spend, not while trying to justify it afterward.

What counts as value depends entirely on the function and the industry. In customer operations, it is deflection rate and cost per resolved interaction. In software development, it is cycle time and accepted contributions per developer. In back-office processing, it is hours per transaction converted to a dollar figure. In risk and compliance, it is decisions per analyst and false positive rate. The definition changes. The requirement does not: name the outcome, set the baseline, and build the measurement in before the workload reaches production.

Discipline 5:

Build AI skills across the organization

Most routing mistakes happen because people lack clear guidance, not good judgment. Developers and employees choose the most capable model because they can’t see the cost difference—and often don’t have a simple rule for when a less expensive model is enough. Training helps, but only to a point: 65% of companies have introduced some form of user training, and the routing gap remains. Awareness doesn't change decisions unless it's reinforced at the moment of choice.

The VP of Enterprise AI Strategy at a major global technology company put it bluntly:

"Your Toyota Camry will do just fine. You don't need to upgrade to a Ferrari anytime soon."

The same executive described where he expects the discipline to land: token efficiency built into developer KPIs, budget allocated by persona and use case and quality of AI usage measured alongside volume. Getting there requires that people understand what is expected of them and have what they need to meet that expectation.

The goal is not to make people responsible for costs they cannot fully see. Total cost is difficult to internalize: a less expensive model may take longer, require more prompt iterations, or produce output that needs more human review. The aim is behavioral guidance. Give people a simple framework for which model tier fits which task. Make cost visibility available to those who want it. Set the default through gateway routing policy or desktop policy so that the right choice is also the easy choice, and reserve human judgment for the cases where it truly matters.

The organizations doing this well are setting expectations, making information available and letting the governance structure carry the enforcement weight. People follow clear guidance when it is easy to follow. The CIO's job is to make it easy.

This goes beyond user training. Tokenomics crosses the boundaries of technology, finance, workforce and business ownership. The CIO needs the CFO to make measurable value and cost accountability conditions of AI investment; the CHRO to embed model selection and cost awareness into workforce expectations; and business leaders to give every AI workload a named owner, an economic case and accountability for the outcome from day one.

The five disciplines may form the technology operating model, but making them work is a C-suite responsibility. The CIO's job is to orchestrate it.

Conclusion

The discipline is the differentiator

AI leadership comes down to control and management discipline. Companies win when they manage each token before it becomes a cost to justify, then connect it to a value case they can explain.

In many companies, access came first. Governance came when the bill arrived. That sequence is reversible but reversing it under financial pressure is harder and more expensive than getting it right at deployment. The companies getting this right are not outliers. A technology executive built attribution in from the start. A field CTO asked, “What did that token actually do?” before the tools scaled. ABN AMRO required every use case to map to a defined value outcome before production. The survey shows a similar pattern everywhere: organizations running early pilots optimize about 15% of their token spend and route 30% of requests correctly, while those running AI in production at scale reach roughly 28% and 58%. All of this comes from having visibility, a routing layer, an operating model and the discipline to run AI infrastructure the way a manufacturer runs a production line: managing unit cost, cutting waste and measuring continuously.

The organizations that build this infrastructure now will compound the advantage for years. AI capabilities will keep improving. Model prices will keep shifting. The consumption curve will keep rising. None of that changes the underlying logic. Value comes from what the token does, not from spending money on it.

That is the gap this discipline closes, and the window to close it is before the next model generation, the next agentic expansion or the moment an unpredictable AI bill becomes a material management issue. When that happens, the questions will travel fast from the board to the CEO, from the CEO to the CFO and the CIO. The CIOs who act now will have an answer. Tokenomics is becoming a standard discipline for running enterprise AI. Build it now and the CIO can give the CFO confidence in the economics, give the CEO a return they can quantify and give the enterprise permission to scale AI on evidence rather than hope.

How Accenture can help

A company's technology foundation is no longer a support function. It is the mechanism through which advantage is created and compounded, and it is the CIO's responsibility to build, fund and defend.

To execute that responsibility, technology leaders need to bring three things together at once: the right foundation, the right operating model and the right investment logic—done in lockstep with business stakeholders, with a roadmap that prioritizes targeted actions maximizing value to the business. Very few companies are connecting all these facets today, and that is why so much of AI's potential still fails to show up as tangible business results.

We think about rewiring businesses, not just technology stacks. Accenture helps clients develop a reinvention-ready digital core across the three dimensions of the modern estate: the Engine runs the enterprise, Knowledge gives the enterprise a trusted understanding of itself and the Brain interprets that understanding and activates it through AI models, orchestration and agents. All of this is supported with a layer of security and governance.

Accenture's own production platform processes approximately 8.7 trillion tokens per week at roughly one-sixth of frontier API cost, through self-hosted open-weight inference and intelligent routing. Accenture Tokenomics brings that capability to clients through a proven three-stage approach—see it, fix it, keep it—supported by the Accenture AI Token Navigator, a set of seven specialized tools spanning forensic assessment of the AI estate, live model benchmarking, runtime enforcement of budgets-as-code and policy-as-code, real-time monitoring of cost per successful business action, and the engineering and talent practice that makes cost discipline durable.

About the research

Our research combines three inputs: a survey of 750 senior executives across 17 countries, 15 in-depth conversations with Fortune 500 technology and finance leaders, and a simulation of 6,000 senior leaders to stress-test the patterns we saw in the data. Every survey respondent worked at a company with more than $1 billion in annual revenue and had direct responsibility for AI spending decisions.

Primary survey of N=750 senior executives, $1B+ revenue enterprises, 17 countries, July 2026. Synthetic simulation: Aaru, 6,000 agents across six enterprise maturity segments, clearly labelled as simulation data throughout. Qualitative interviews: 15 in-depth conversations with Fortune 500 technology and finance executives, conducted June–July 2026, anonymized and aggregated. Statistical methods: Pearson correlation, multiple regression, ANOVA. All regression and correlation findings are statistically significant at $p<0.05$ or better unless stated otherwise.

Acknowledgments

Research Team:
Amal Sebastian, Aditi Abhijeet, Biswadeep Ghosh Hazra, Alok Ranjan, Ajay Garg, Krishna Kanth Sutraye, Gerry Farkova, Nataliya Sysenko, Ignacio Mamone, Kevin Gallagher, Gargi Chakrabarty, Pradeep Roy

SMEs/Leadership:
Tony Leraris, Steve Courtney, Matt Lagodzinski, Mike Eisenstein

Marketing + Communications:
Brianna Chan, Art Gencarelli, Laurence Mackin, Ed Maney, Janine Stankus

Glossary

Financially attributed value (compound attribution) – The share of total token spend that an organization can express as a quantified financial outcome — revenue influenced, cost avoided, or productivity converted to a dollar figure. It is a compound measure because financial attribution happens in two stages, each captured by a survey question:

- **Linkage:** the share of token consumption that can be connected to a measurable business outcome.
- **Financial expression:** of that linked consumption, the share that can be expressed as a quantified dollar figure.

Because the second stage is conditional on the first, the compound metric is their product:

Compound attribution =
Linkage × Financial expression

For example, an organization linking 60% of its consumption to an outcome and financially expressing 60% of that achieves 0.60 × 0.60 = 36% compound attribution.

Optimization rate – The reduction in token consumption an organization has achieved through efficiency efforts — routing, caching, prompt discipline, context management and similar levers. Reported as a share of consumption removed relative to an unoptimized baseline.

Break-even optimization – The optimization rate required to hold total token spend flat as volume grows and unit prices fall over a 24-month horizon. Derived, not surveyed:

Break-even =
1 − 1 / [(1 + volume growth) × (1 + price change)]

Using the survey's expected 24-month volume growth (+78%) and price change (–19%), the break-even rate is approximately 31%.

Showback vs. chargeback – Showback makes token consumption visible to the consuming team but attaches no financial consequence. Chargeback bills the consuming team for what it uses, creating a direct financial incentive to manage consumption. The report treats the progression from no cost allocation > informal reporting > showback> chargeback as an accountability ladder.

Routing quality – The share of AI requests directed to the lowest-cost model capable of delivering the required outcome. A request sent to a more expensive tier than the task demands is "mis-routed."

Workload-level attribution – The ability to trace token consumption to a specific user, team, application or workload (The observability foundation on which both routing and financial attribution depend).

How to read the figures

A quick reference for the report's core metrics, in plain terms.

Term	In plain terms
Compound attribution	The share of every token dollar you can actually prove to a board as a financial outcome. Today about 20 cents on the dollar.
Optimization rate	How much token consumption volume you've cut through efficiency (routing, caching, prompting). Survey average ~23%.
Break-even optimization	The optimization needed to keep the bill flat as volume grows and prices fall. ~31% over 24 months.
Showback vs. chargeback	Showback = the team sees the cost. Chargeback = the team pays the cost. Only chargeback changes behavior.
Routing quality	The share of requests sent to the least expensive model that can still do the job. Higher is better.

References

1. Tokenomics Foundation, **What is AI tokenomics?**, September 1, 2026.

2. Accenture analysis. Without optimization, token costs are projected to grow approximately 44% over 24 months, driven by 78% volume growth partially offset by 19% price decline. A 31% optimization rate holds spend flat against that trajectory. A small cohort of organizations with full chargeback implementation (N≈53) reported optimization rates approaching 48%; given the sample size this is treated as directionally indicative rather than a population benchmark.

3. Estimated aggregate based on weighted mean annual token spend of $3.37 million across N=741 survey respondents with known spend. The 24-month projection applies survey-derived volume growth (+78%) and price decline (-19%) assumptions to the same base.

4. Natalie Lung, **Uber caps usage of AI tools like Claude Code to manage costs**, Bloomberg, June 2, 2026.

5. Ben Shimkus, **Harvey CEO Says Company's AI Token Use Hit 12 Trillion in May**, Business Insider, June 19, 2026.

6. Henry Chandonnet, **AI token costs are forcing companies to rethink how they hire, budget, and manage usage**, Business Insider, June 17, 2026.

7. Chris Shuptrine, **7 tools for tracking AI token usage across vendors in 2026**, Torii, June 2026.

8. **'We created a monster': companies rein in AI usage as costs strain budgets**, Financial Times, June 19, 2026.

9. Makenzie Holland, **Shutterstock, Prudential Financial share AI cost strategies**, CIO Dive, June 11, 2026.

10. **The AI cost reckoning: Right-sizing model spend 2026**, Digital Applied, June 23, 2026.

11. **AI is on your P&L. Most companies are only reading half of it**, Accenture, July 29, 2026.

12. IDC FutureScape: **Worldwide AI and Automation 2026 Predictions, doc #US53858125**, October 2025.

13. **The AI cost reckoning: Right-sizing model spend 2026**, Digital Applied, June 23, 2026.

14. **Match Group CFO sets "higher bar" for AI spending in 2026**, CFO Dive, December 23, 2025.

15. ABN AMRO Innovation & Technology, Investor Relations, **ABN AMRO and AI: Investor meeting**, May 20, 2026.

16. **Klarna AI customer support efficiency**, Twig, June 10, 2026.

17. Jake Angelo, **Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it,** Fortune, May 26, 2026.

About Accenture

Accenture helps the world's leading enterprises reinvent by building their digital core and unleashing the power of AI to create value at speed for organizations across industries. Our strategy is to be the reinvention partner of choice for our clients and lead in the safe, widespread adoption of AI, and to be the most client-focused, AI-enabled, great place to work in the world. We bring together the talent of our approximately 799,000 people with proprietary assets and platforms, deep process and industry expertise, and leading ecosystem relationships to deliver end-to-end solutions and measurable outcomes at scale. Through our Reinvention Services, we offer broad expertise across Cybersecurity, Digital Core, Finance, Industry and Enterprise, Song, Supply Chain and Engineering, and Talent, with advanced capabilities in AI and Data, Industry and Process, and Technology. We serve approximately 9,000 clients and generated approximately $70 billion in FY25 revenue.

Visit us at accenture.com.

Accenture Research

Accenture Research creates thought leadership about the most pressing business issues organizations face. Combining innovative research techniques, such as data-science-led analysis, with a deep understanding of industry and technology, our team of 300 researchers in 20 countries publish hundreds of reports, articles and points of view every year. We use generative AI in our research production process. Our research experts review and validate the generative AI outputs with traditional research methods where possible, applying Accenture's Responsible AI standards. Our thought-provoking research developed with world-leading organizations helps our clients embrace change, create value and deliver on the power of technology and human ingenuity.

Visit Accenture Research at accenture.com/research.