acceligence

Scaling AI without breaking the economics

Insight

May 3, 2026

Category: technology

Tags: adoption, ai, budget, cost, economics

Every executive team is facing the same next question on AI: *how do we move from promising pilots to scaled adoption that performs in the real economics of the business?*

That is a good question. It means AI has moved out of the lab and into operating plans, customer workflows, productivity targets, and board-level expectations. The conversation is no longer about whether AI will matter. It is about how thoughtfully, confidently, and economically organizations can put it to work.

The opportunity is significant. AI can compress cycle times, improve decision quality, remove low-value work, and give teams more capacity for the work that actually requires judgment. But as adoption accelerates, so does the need for discipline. Production AI is not just a technology decision; it is an operating model decision. It has architecture, governance, talent, vendor, and cost implications that compound quickly once usage scales.

That is where the next advantage will be created. The leaders who pair ambition with economic clarity will be able to move faster, not slower. They will know which workflows deserve frontier-model performance, which can be routed to lower-cost options, where abstraction matters, and how to protect margins as usage grows.

In 2026, the mandate is adoption. By 2027, the differentiator will be the ability to scale AI with the economics intact. The companies that prepare now will not be caught defending AI's cost. They will be using AI's cost discipline as a source of strategic advantage.

From experimentation to enterprise scale

The pressure to adopt AI is real, and in many organizations it is producing useful momentum. Leaders are looking across procurement, contracting, customer support, engineering, finance, sales operations, and other core functions with a sharper question than they were asking even a year ago: *where can AI create measurable capacity, speed,*

quality, or insight?

That shift is encouraging and it means AI is no longer sitting at the edge of the business as an innovation exercise. It is being pulled into the core operating model, the very DNA of the company. Business leaders are identifying workflows. Technology teams are building the infrastructure. Governance groups are adapting to a higher volume of use cases. Vendors are competing for enterprise commitments. The work is active, serious, and increasingly tied to business outcomes.

But scale changes the nature of the decision.

A pilot can succeed on enthusiasm, narrow scope, and favorable pricing. A production deployment has to succeed under repeated use, across teams, inside real workflows, and against the economics of the business. The assumptions that felt reasonable during experimentation become much more important once AI is embedded in day-to-day operations.

That is why the next leadership move is not to slow adoption. It is to strengthen it. Executives have an opportunity now to ask which assumptions behind their AI roadmap are durable, which are temporary, and which need to be stress-tested before usage compounds. Model performance, vendor strategy, governance capacity, workflow design, talent readiness, and cost structure are no longer separate conversations. At scale, they become one operating question: can this create value reliably, repeatedly, and economically?

The organizations that answer that question early will not lose momentum. They will build the confidence to move faster, because their AI adoption will be grounded in the realities that determine whether innovation turns into enterprise value.

The economics of scale

Token prices today should be treated as an unusually favorable adoption environment, not a permanent planning baseline.

At current rates, many AI use cases clear the business case quickly. That is useful. It gives executives room to experiment, redesign workflows, build user trust, and move proven capabilities into production. But the conditions making that possible are not entirely structural. Barring a major hardware breakthrough, or a step-change in compression and inference efficiency, today’s pricing is being shaped at least partly by capital-funded growth strategies, enterprise land grabs, and pre-IPO positioning – not by mature, steady-state unit economics.

What makes that important? Three forces are moving in the same direction.

Infrastructure capacity is still scarce

AI infrastructure is not a single line item. It depends on GPUs and accelerators, power availability, cooling, networking, and data center capacity. Each of those inputs is under pressure. Chips will improve. Facilities will expand. Providers will find efficiencies. But those gains may not arrive fast enough for providers to absorb rising demand through cost reduction alone. The executive implication is simple: do not build a production roadmap that assumes AI infrastructure will follow the same downward pricing curve as traditional cloud compute. Deloitte has projected that AI's next phase will require more compute, not less, while recent infrastructure analysis has pointed to power, cooling, and accelerator supply as major constraints on scaling AI data centers.

Models are consuming more tokens per task

Frontier models are improving, but much of that improvement comes from doing more work inside the task: reasoning longer, calling tools, retrieving context, verifying outputs, executing code, and coordinating multi-step agentic workflows. That is exactly what makes the technology more valuable. It also changes the cost profile. A workflow that once looked like a single prompt and response can now become a sequence of model calls, tool calls, retries, and verification steps. Some agentic workflows already consume 10x to 100x the tokens of a comparable single-shot completion. Cost models built around 2024 usage patterns were not designed for that kind of consumption curve.

Providers are beginning to expose the strain

The signals are showing up less in vision statements than in product and pricing decisions. In March 2026, OpenAI announced it would shut down Sora, its consumer AI video app, roughly six months after release. Reporting tied the move to the economics of video generation: Sora's worldwide user count reportedly peaked around one million and fell below 500,000, while the app was burning roughly $1 million a day in compute. The timing also mattered. OpenAI was redirecting resources toward more commercially sustainable products ahead of a possible public-market debut, and Disney, which had been tied to a reported $1 billion Sora partnership, learned of the shutdown less than an hour before the public announcement. When a product with that level of strategic visibility gets pulled that quickly, the lesson is not that AI demand is weak. It is that compute allocation has become a top management-level strategic and economic decision.

A quieter signal pointing in the same direction

In April 2026, Anthropic briefly tested removing Claude Code from its $20 Pro plan for a small subset of new subscribers, then restored the listing after developer backlash and confusion. The reason matters. Amol Avasare, Anthropic's head of growth, explained that engagement per subscriber was up, long-running agents had become everyday workflows, and "usage has changed a lot and our current plans weren't built for this." Industry coverage was more direct: The Register reported that Anthropic's subscription plans were charging far less than the book value of tokens consumed, sometimes by a factor of ten or more. Around the same period, Anthropic more than doubled its public estimate of average Claude Code token spend for enterprise developers from $6 to roughly $13 per active day.

Taken together, these are not signs to pull back from AI. They are signs that the economic models around it are growing up.

The economics that powered early pilots will not automatically power production deployments. Pilots can succeed on narrow scope, favorable pricing, and enthusiastic usage. Production systems have to work repeatedly, across teams, inside real workflows, and at volumes large enough for small cost assumptions to become material.

That is where the opportunity sits. Leaders who understand unit economics early can scale with more confidence. They can decide which workflows truly require frontier-model performance, which can be routed to lower-cost models, where vendor optionality matters, and where architecture needs to be redesigned before cost pressure limits choice.

The CFO question worth modeling now: *what happens when token prices increase 3x to 5x?*

A workflow that costs $0.02 per execution and runs 10 million times a month costs $200,000 a month today. At 3x, it becomes $600,000. At 5x, it becomes $1 million. If the value created by that workflow still clears the hurdle, scale it. If the economics break, redesign it now, while the organization still has leverage, optionality, and time.

The goal is not to slow AI adoption. It is to make adoption stronger. Cost discipline is not a constraint on ambition; it is how ambition becomes durable.

The resilient AI playbook

This is where resilience becomes more than a principle; it becomes a strategic advantage.

We have written before about rethinking operating models for the age of AI and the irreducible core of humanity in adopting AI; building organizations and architectures that can absorb change without losing momentum – and their humanity – are positioned to win. The next phase of AI adoption will test that discipline. As usage scales, pricing shifts, models evolve, and vendor strategies change, organizations will need more than enthusiasm for AI. They will need systems that can adapt.

That is the opportunity in front of executive teams now. AI cost pressure does not have to become a constraint on adoption. Handled early, it can become the forcing function that makes AI programs stronger, more flexible, and more durable.

Four pillars are worth pressure-testing before repricing forces the issue. Each belongs to a different leader on the executive team, yet they are all related and should not be handled in isolation.

1. Architectural resilience

Owned by your CTO or CIO

Are your AI workflows model-agnostic, or are they hard-wired to a specific provider's prompt format, context window, tool-calling syntax, and orchestration pattern?

A useful test is simple: how long would it take your team to swap your primary inference provider this quarter?

If the answer is measured in months, there is architectural debt hiding inside the roadmap. That debt may not matter much during pilots, when the priority is speed. But in production, lock-in limits negotiating leverage, slows adaptation, and makes cost pressure harder to manage. The goal is not to avoid strategic vendor relationships. It is to make sure those relationships are chosen, not trapped.

Architectural resilience gives leaders options. It allows teams to route workloads across models, respond to pricing changes, take advantage of new capabilities, and avoid rebuilding critical workflows every time the market shifts.

2. Economic resilience

Owned by your CFO

Most organizations know their aggregate AI spend by vendor. Far fewer know their true unit economics at the level that matters: cost per execution, per workflow, per business outcome.

That distinction becomes critical as AI moves into production. Contract-level visibility tells you what you spent. Workflow-level visibility tells you why you spent it, where value was created, and which costs are worth scaling.

Without that granularity, repricing becomes a blunt-force event. When token prices move 3x, you do not want to discover that 80% of your AI spend is concentrated in three workflows for which nobody has current cost numbers. You want to know which workflows are margin-accretive, which need redesign, which should move to lower-cost models, and which should be retired.

Economic resilience gives the CFO a more useful conversation with the business. Instead of asking whether AI spend is too high, the better question becomes: where is AI creating measurable value, and are we scaling those economics deliberately?

3. Operational resilience

Owned by your Chief AI Officer or Chief Data Officer

Are frontier models being reserved for the work that genuinely requires them, or are they becoming the default for every use case because they are convenient, impressive, and already approved?

This is one of the most important operating questions in enterprise AI. Every team wants the best available model for its workflow. That instinct is understandable. It is also expensive. As AI usage grows, the ability to distinguish between tasks that need frontier performance and tasks that can be handled by smaller, cheaper, specialized, or open models becomes a source of real advantage.

That shift requires more than policy. It requires routing logic, evaluation coverage, performance benchmarks, exception handling, and organizational permission to move mature workloads down the cost curve. Teams need confidence that lower cost does not mean lower quality where quality matters.

This is a 12-to-18-month capability build, not a procurement adjustment. The organizations that started in 2025 are already creating separation. The organizations that start now can still turn model tiering, evaluations, and workload routing into a durable operating strength before cost pressure narrows their choices.

4. Strategic resilience

Owned by your Chief Strategy Officer/Corporate Development

Every AI roadmap contains pricing assumptions, whether they are explicit or not. Some are embedded in business cases. Some are hidden in adoption targets. Some sit inside vendor projections. Many are borrowed from the cloud pricing curve of the last decade: the expectation that infrastructure costs will trend down steadily, and often faster than planned.

That assumption should not be carried unexamined into AI inference.

Inference economics are shaped by constrained hardware, growing power and data center requirements, rising consumption per task, and providers that are still balancing enterprise growth with their own margin realities. If your three-year AI plan assumes today’s token economics or better, that plan contains a hidden bet on continued subsidy.

That bet may be acceptable. But it should be made consciously, modeled clearly, and hedged appropriately.

Strategic resilience means asking a better set of planning questions: Which initiatives still work if token prices increase? Which vendor commitments protect flexibility? Which use cases depend on frontier models staying cheap? Which workflows create enough value to withstand repricing? Which parts of the roadmap should be accelerated now while the economics are favorable?

The point is not to make AI strategy more cautious; it is to make it more credible.

The companies that build resilience across architecture, economics, operations, and strategy will be better positioned

to scale AI with confidence. They will have more options, stronger governance, clearer ROI, and greater negotiating leverage. Most importantly, they will be able to keep moving when the market shifts because they will have built their AI programs for change from the beginning.

What changes in 2027

By 2027, AI cost discipline will be a board-level conversation. The questions will change. In 2026, leadership teams are being asked how many AI use cases they have launched, how quickly adoption is moving, and where productivity gains are showing up. Those questions will not disappear. But they will be joined by a more mature set of questions: how intelligently are you routing workloads? How disciplined is your model tiering? Which vendors are accountable for the total cost of ownership, not just capability demonstrations? Which workflows are creating lasting value after infrastructure, inference, integration, governance, and operating costs are fully counted?

That shift is healthy and is what happens when an important technology moves from experimentation into the operating model.

A new capability will emerge inside serious enterprises. The title may vary – AI economics engineer, inference FinOps lead, AI cost architect, or something the market has not settled on yet. The work will be the same: making the unit economics of AI visible, manageable, and aligned to business value.

This role will sit at the intersection of finance, technology, operations, and strategy. It will translate token consumption into workflow economics. It will help teams understand when frontier models are worth the premium, when lower-cost models are sufficient, and when the workflow itself needs redesign. It will give executives a clearer view of where AI is creating margin, where it is consuming margin, and where better architecture can improve both performance and cost.

The most successful companies in 2027 will not necessarily be the ones with the largest number of AI deployments. They will be the ones with deployments that can scale under real operating conditions, with flexible architecture, transparent unit economics, disciplined workload routing, and ROI that remains defensible as pricing changes.

That is an optimistic future, and it means AI value does not depend on cheap tokens forever. It depends on leadership teams building the discipline to use AI well.

What to do this week

Four moves, mapped to the four resilience pillars. None requires a full budget cycle to begin and set an organization up for long-term success and scale.

1. For architecture

Build provider abstraction into anything moving to production

Provider relationships matter, but lock-in should be a choice, not an accident. Any workflow moving from pilot to production should be designed with enough abstraction to preserve optionality across models, inference providers, context formats, and tool-calling patterns.

Lock-in is one of the most expensive forms of technical debt in enterprise AI because it often stays invisible until the moment flexibility matters. By the time pricing shifts, a model underperforms, or a provider changes terms, the cost of switching may already be measured in months. Abstraction gives the CTO room to adapt without forcing the business to slow down.

2. For economics

Audit token economics at the workflow level

Contract-level spend is not enough. It tells you what was paid, but not whether the spend created value. Start with the three highest-volume AI workflows in the organization. Establish a baseline cost per execution, cost per completed workflow, and cost per measurable business outcome. Then identify the drivers: model choice, prompt length, context retrieval, retries, tool calls, human review, and orchestration overhead.

The goal is not perfect accounting. The goal is decision-grade visibility. If your team cannot produce those numbers within two weeks, that is the finding. It means the organization is scaling AI without the instrumentation needed to manage it.

3. For operations

Tier workloads and treat evaluations as infrastructure

Most enterprises are using frontier models more often than they need to. That is understandable during pilots, when the priority is proving value and reducing friction. But in production, defaulting to the strongest model for every task is rarely the most durable operating pattern.

The practical move is to tier workloads by complexity, risk, and value. Reserve frontier models for tasks that require advanced reasoning, high ambiguity, or material business judgment. Move mature, repeatable, verified, lower-risk tasks down the cost curve where performance can be maintained.

Evaluations are what make that possible. Without strong eval coverage, teams cannot prove that a lower-cost model performs well enough to trust in production. With evals in place, model tiering becomes a source of confidence rather

than a compromise.

4. For strategy

Stress-test the 2027 roadmap at 3x, 5x, and 10x current token prices

Every AI roadmap contains a pricing assumption. The question is whether leadership has made that assumption explicit.

Model the 2027 AI budget at 3x current token prices and review which use cases still work. Then test the roadmap again at 5x (or even 10x) for the highest-volume workflows. The point is not to predict the exact increase. The point is to understand sensitivity.

If a workflow still creates attractive value under that scenario, it is a strong candidate for scale. If the math breaks, the organization has time to redesign the workflow, change the model strategy, renegotiate terms, or reconsider the use case before it becomes a board-level surprise. Better to find the weak points in a planning meeting than in a performance review.

For leverage, negotiate rate protection while the enterprise demand is still valuable. CIOs and CFOs have more leverage before repricing than after it. For high-confidence workloads, multi-year agreements with usage bands, rate caps, transparency commitments, and pricing protections can create an important hedge. The goal is not simply to secure a discount. It is to create predictability around the economics of workflows that are becoming part of the operating model. That predictability matters because it gives the business confidence to scale, provides finance with a cleaner planning basis, and gives technology teams more room to optimize without reacting to every market shift.

The companies that act now will enter 2027 with more than AI momentum. They will have AI discipline. They will know where value is being created, where costs are likely to move, and which parts of the roadmap are resilient enough to scale.

The elevator question to ask your teams today

Does our 2027 AI roadmap still work if token prices increase 3x?

If the answer is *yes*, scale with confidence and revisit periodically.

If the answer is *no*, redesign while there is still time.

If the answer is *unclear*, that is the place to start.