aicadium

PLANNING FOR THE AGE OF AI TOKENS

The Industrial CFO's Guide to Planning and Budgeting for AI

A perspective for C-suite, CFOs, and finance leaders in industrial companies

FINANCE & AI STRATEGY

SEP 2026 · AICADIUM

Executive Summary

Industrial companies are approaching a familiar-looking decision: how much to spend on a new, rapidly evolving technology with a cost structure that behaves nothing like the ones their planning processes were built for. AI spend, and token-based inference spend in particular, is usage-driven rather than seat-driven, moves in orders of magnitude rather than percentages, is scattered across cloud bills, is embedded in SaaS features, and is sometimes masked behind shadow purchases rather than sitting in one clean line item. FinOps Foundation data suggests mature teams forecast cloud spend within 1–3% and miss AI spend by a factor of two to three; a majority of enterprises report AI costs have already exceeded original projections. For a finance function used to forecasting capex on a five-year depreciation schedule, this is a different discipline, not a bigger version of the old one.

This paper sets out practical ways for CFOs and finance leaders at industrial companies to think about four questions: why are AI costs structurally harder to control than traditional IT spend (which we call "the Four V's"); how to plan and budget for AI investments given that difficulty; how to monitor whether the spend is generating a return; and how to keep effective use of AI growing. For most industrial companies today, under-adoption, not overspend is the bigger risk, without letting AI costs run wild.

Our central argument is that industrial companies should treat AI spend as two distinct portfolios that need two different governance models. The distinction that matters is not whether an investment is "productivity" or "strategic" as most strategic AI earns its return through productivity but whether there is a measurable "before" state to judge it against. Investments that improve something you are already doing have a baseline, a unit of value and a payback period that can be budgeted much like any other IT investment. However, investments that build a capability you do not yet have, for example, a computer-vision system that lets you deploy a new inspection use case or a new agentic system that changes a plant's operating model, cannot be evaluated in the same way because there was no "before" state to compare against. Conflating the two is the single most common planning mistake we see, and this is why boards are told an AI initiative "didn't pay back" this quarter, when it was never meant to.

We have four recommendations to address these challenges:

RECOMMENDATION 01

Name and govern the Four V's explicitly rather than treating AI cost as an extension of the cloud budget.

RECOMMENDATION 02

Fund new-capability AI like a venture portfolio and usage-based AI like a metered utility, not both like enterprise software.

RECOMMENDATION 03

Measure early-stage bets on leading indicators, and reserve financial ROI for the validation gate and not at the funding gate.

RECOMMENDATION 04

Make the economics of usage visible to the people generating it, so purposeful adoption grows without requiring cost caps to do the governing.

A. The Four V's Problem With AI Costs

Three major forces are converging to accelerate AI spend for industrial companies:

FORCE 01

Frontier foundation model providers have begun moving away from bundled, subsidised token allowances toward committed-consumption pricing.

FORCE 02

Workloads driving the fastest growth in consumption are agentic AI, for example, systems that plan, call tools, and retry, consume five to thirty times more tokens than an equivalent chat interaction.

FORCE 03

A growing share of AI cost is embedded inside operational technology with safety-related implications, for example, control-system, Manufacturing Execution System ("MES"), and industrial-SaaS vendors are adding AI features into renewal contracts which masks cost visibility.

However, none of these reasons justify reducing AI budgets. On the contrary, use cases in industrial companies, for example, predictive maintenance, computer-vision quality inspection and field-service copilots, have proven that AI delivers real operating leverage: fewer unplanned outages, lower scrap rates, faster technician ramp-up. The risk is not in funding too much AI, it is that AI is funded based on the wrong tools and expectations. At the end, finance leaders either get surprised by the bill, or become so wary of being surprised that they slow-walk the AI initiatives which could have the best returns.

CFOs already know how to budget for software that is licensed per seat: predictable, easy to model a year out, and simple to true up at renewal. However, AI cost does not behave this way because of the four structural reasons which we call the Four V's — Visibility, Variability, Volatility, and Value elaborated below. Together, they are why a growing number of industrial companies are trying to build a dedicated AI cost governance capability, often described as "FinOps for AI" rather than folding AI spend into the existing IT budget process.

EXHIBIT 1: EMPIRICAL REPORTS ON AI COST CONTROL ISSUES

73%

of enterprises say AI costs have exceeded original projections

2–3x

the margin by which AI forecasts miss, against 1–3% for mature cloud forecasting

31% → 98%

share of FinOps practitioners managing AI spend, 2025 to 2026

Source: Community reporting from FinOps Foundation and 2026 industry surveys

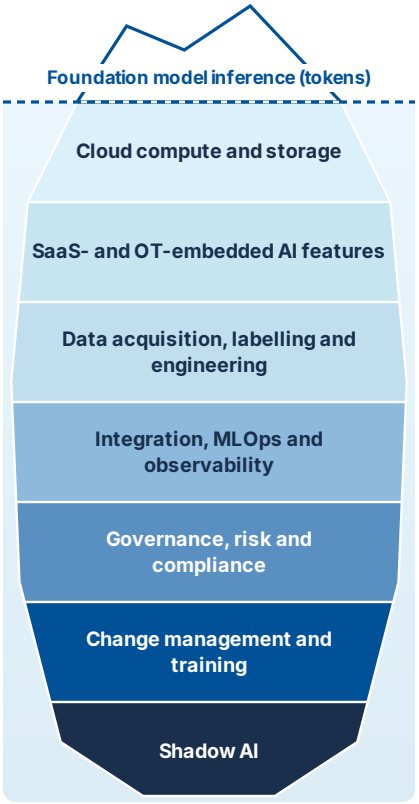
1. Visibility: you cannot budget for what you cannot see

AI cost is rarely a one-line item. It is buried inside cloud bills across AWS, Azure, and GCP; inside vendor SaaS subscriptions that have quietly added AI features at renewal; and inside dozens of team-level tool subscriptions that were never centrally procured. Cost allocation data from the FinOps community suggests that actual metered AI spend can run at a small fraction of what organisations report as AI

budget intent, much of the rest is embedded in SaaS contracts invisible to the buyer. For industrial companies, this problem has a second layer: operational-technology vendors (control systems, MES, asset-performance-management platforms) are beginning to bundle AI-based analytics into plant-floor software the same way productivity vendors have, so a portion of "AI spend" shows up as an Operational Technology ("OT") capital or in the maintenance line.

The consequence is shadow AI and decentralised procurement: individual teams and employees sign up for AI tools outside centralised procurement, creating budget leakage, duplicate tooling, and potentially more seriously, a data governance gap when sensitive plant, customer, or engineering data is uploaded to personal or unauthorised accounts. In regulated industrial segments (aerospace, pharma, energy, defence-adjacent manufacturing), this is not only a cost problem: use of AI output in decision-making needs to be traceable, and an unapproved tool with no audit trail can create a compliance exposure that costs far more to remediate than the subscription itself.

EXHIBIT 2: THE AI COST STACK "ICEBERG"



THE COST STACK	WHERE THE COST LANDS	WHY IS IT EASY TO MISS?
Foundation model inference (tokens)	**Cloud and AI vendor invoices** *IT infrastructure*	The one layer your dashboard shows — but reasoning and agentic work consume 5–30x the tokens of an equivalent chat, for the same visible output.
Cloud compute and storage	**Cloud and AI vendor invoices** *IT infrastructure*	Bundled into a broader infrastructure line, not tagged to the AI initiative that generated it.
SaaS- and OT-embedded AI features	**SaaS and OT software renewals** *Procurement, plant systems*	Priced as a flat subscription increase at renewal; the underlying token meter is invisible to the buyer.
Data acquisition, labelling and engineering	**Project and capital budgets** *Engineering, project sponsor*	Plant and machine data is messier than SaaS data; cleaning and labelling is frequently the largest single line, and the least budgeted.
Integration, MLOps and observability		Ongoing, not one-time — model monitoring, drift detection and re-tuning continue for the life of the deployment.
Governance, risk and compliance	**Operating cost and headcount** *Business unit, quality, HR*	Security review, bias and safety audits, legal review of vendor contracts, and validation and change control for safety-relevant use cases.
Change management and training		Coaching, workshops and champion programmes; frequently cut first when budgets tighten, which undermines the ROI case for everything above it.
Shadow AI	**Expense claims — or nowhere** *Nobody*	Unbudgeted and untagged by definition; typically discovered at renewal or during a security review, not during planning.

2. Variability: usage-based cost defeats seat-based forecasting

Traditional software is licensed per seat and easy to forecast a year ahead. AI spend however, for example, LLM API usage, inference costs, compute for training or fine-tuning, scales with tokens processed, API calls, and GPU-hours, which fluctuate with adoption, feature usage, and even how a prompt is written. A single query routed through a retrieval-augmented pipeline with a reasoning model can consume one to two orders of magnitude more tokens than a direct prompt to a smaller model with

no visible change in what the end user experiences. This makes it genuinely difficult to translate month-to-month cost into an annual budget number and makes adoption-curve forecasting a real challenge. Tools can go viral internally or fail to gain traction entirely.

3. Volatility: the ground moves under the budget you just built

Model providers change pricing, deprecate models, and release new ones with different cost-performance trade-offs on a cadence measured in months, not budget cycles. A model that is falling in list price can still produce a rising bill because consumption, especially from agentic and reasoning workloads is growing faster than unit prices are falling. The build-versus-buy-versus-rent decision is unusually dynamic for the same reason: building in-house capability, buying a fine-tuned or licensed model, and renting API access carry different cost structures and risk profiles, and the "right" answer for a given use case can flip within a budget year as pricing and open-source model quality both move.

4. Value: attribution and ROI resist the old formula

Unlike headcount or a marketing campaign, AI's productivity gains, for example, time saved, quality improvements, revenue lift are often soft, delayed, or hard to isolate from other changes happening at the same time. Boards are asking for AI ROI with increasing urgency and industry surveys consistently show a wide gap between the share of executives who perceive productivity gains from AI and the much smaller share who can confidently measure the return in financial terms. We address this gap directly in Part C, because it is where most industrial AI programmes actually stall i.e. not at the funding stage, but at the "prove it" stage six months later.

A common thread runs through all four V's: CFOs are being asked to budget for a cost category that behaves like a metered utility using processes built for fixed, predictable IT costs. That mismatch is the main reason why AI budgets overrun so consistently across industries today.

EXHIBIT 3: SUMMARY OF THE 4 V's PROBLEM WITH AI COSTS

COST SIDE
Three ways AI spend escapes control

VALUE SIDE
One way it escapes proof

Visibility	Variability	Volatility	Value
You can't see it.	**You can't predict it.**	**You can't lock it down.**	**You can't prove it.**
Fragmented across cloud bills, SaaS renewals and plant software. In industrials, part of it is filed as an OT line and never reaches the CFO's AI view.	Cost scales with tokens, not seats. The same question can cost 100x more depending on model and context — invisible to the person asking.	Pricing, model deprecation and build-vs-rent economics all move inside one budget year. Unit prices fall while bills rise.	Benefit is soft, lagged and entangled with other change. Most executives report gains; far fewer can defend a number to a board.

THE PRACTICAL CONSEQUENCE
Finance is asked to budget a metered utility with a process built for fixed, seat-based IT cost. That mismatch is why AI budgets overrun.

B. How to Plan and Budget for AI Investments and Costs

The starting discipline is to stop treating "AI spend" as one budget category. It is at least two budget portfolios, and they need different owners, different approval processes, and different success measures:

- **Improving something you already do**. For example, copilots, embedded SaaS features, well-scoped automation of existing workflows behaves like a traditional IT investment. It has a "before" state, a measurable unit of value, and a reasonably short payback window. It should be budgeted, approved, and measured the way you already budget, approve, and measure other productivity software, with the addition of usage-based cost controls described below.
- **Building a capability you do not yet have**. For example, a computer-vision model that enables a new inspection use case or a new agentic system, changes how the business competes. Traditional ROI math (cost saved ÷ cost spent) does not really apply, because the payoff is a new market position, product capability, or competitive moat rather than a line-item saving. This category should be funded and governed like an R&D or venture portfolio: staged funding, iterative scope, and evaluation gates rather than one ROI number upfront (see Part C for the full framework).

EXHIBIT 4: "IMPROVING WHAT YOU ALREADY DO" VS. "BUILDING WHAT YOU DON'T HAVE YET" AI BUDGET PORTFOLIOS

The highest-value move: stop running one AI budget. Run two, with different rules.

	IMPROVING WHAT YOU ALREADY DO Copilots · embedded features · scoped automation	BUILDING WHAT YOU DON'T HAVE YET Vision inspection · agentic ops · new offerings
The test it must pass	Does it beat the status quo on a measurable unit?	Does it buy the option to compete differently?
How to fund it	Annual budget, like other software	Staged tranches, like R&D or venture
What gates the money	Business case with a payback period	Evidence gate + pre-agreed kill criteria
What to measure	Cost per outcome, adoption, payback	Technical proof → engagement → P&L
When ROI is fair to ask	At approval	At the scale gate — not before
Who decides	IT / finance business partner	Named owner with authority to kill it
Failure mode to watch	Licences bought, never adopted	Permanent pilot nobody can shut down

Evaluating and funding new-capability AI investments

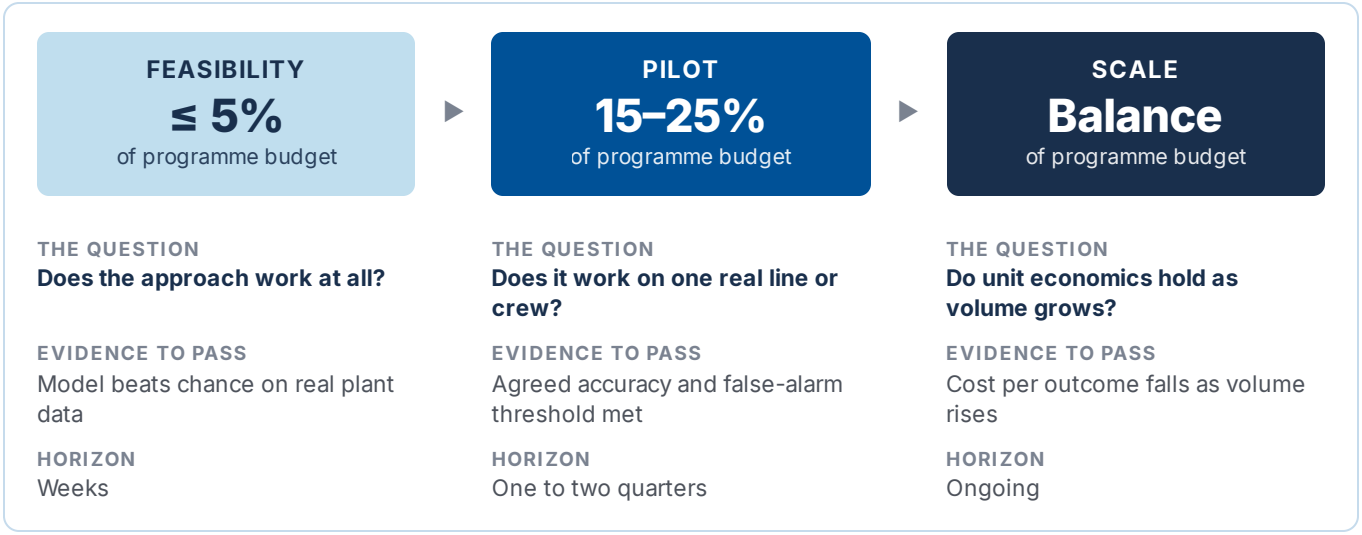
When evaluating potential new-capability AI investments, we recommend that industrial finance teams should require an evaluation of the following dimensions before capital or budget is released, even at pilot scale:

DIMENSION	THE QUESTION A CFO SHOULD BE ABLE TO ANSWER BEFORE APPROVING
Strategic fit	Does this defend or extend a specific competitive position, for example, cost, quality, speed, or a new capability?
Data readiness	Does the plant, process, or system actually produce data of the volume and quality the model needs? This is where industrial pilots most often stall.
Technology and operational readiness	Can the model be integrated with the systems that currently control the plant and is there a named operational owner who will act on its output? For industrial deployments this is frequently the largest cost line and the most common reason a technically successful pilot never scales.
Total cost of ownership	What is the full AI cost stack (See Exhibit 2)?
Kill criteria	What is the fastest test that would tell us this doesn't work, and who has agreed in advance to act on that signal?
Risk and compliance exposure	If this touches a safety-relevant process or regulated data, what validation and audit trail does it need and is that cost already accounted above?

Once a new-capability AI initiative is approved, the funding mechanics matter as much as the initial approval:

- **Fund like R&D, not capex.** Stage the funding and let the scope evolve (See Exhibit 5). A full business case demanded before a proof of concept produces fabricated projections, not better decisions.
- **Ring-fence an experimentation budget**. Typically 2–5% of the technology budget with lighter approval, requiring a measurable hypothesis rather than a full ROI case at pilot stage.
- **Stage-gate the capital**. Cap phase one at roughly 15–25% of total anticipated programme cost, and release later phases only when the initiative meets pre-agreed gate criteria.
- **Run it as a portfolio, not a list of projects**. A dual lens combining an advancement pipeline with defined gates and a dashboard that balances risk and return across horizon and capability. Enterprises that manage AI this way are markedly more likely to reach mature adoption than those approving initiatives one at a time.
- **Pre-agree kill criteria before the pilot starts.** Without an evidence threshold set in advance, weak projects become permanent pilots and zombie platforms that no one has the authority to shut down.
- **Split the funding envelope three ways**. Committed, discretionary, and a mid-year opportunity reserve and rebalance quarterly based on portfolio performance instead of at the end of each annual cycle.
- **Insist on continuous monitoring and quarterly re-forecasting**. Paired with proactive vendor renegotiation as pricing shifts which tracks reality instead of finding out at the end of the annual budget cycle.

EXHIBIT 5: INDICATIVE ALLOCATION OF PROGRAMME BUDGET AND GATES TO NEXT ROUND FINANCING



Budget by how the cost behaves, not by who asked for it

Once a category is approved, budget it according to how the underlying cost behaves, not according to which department requested it. We see four distinct behaviours inside almost every industrial AI portfolio:

SPEND TYPE	BEHAVIOUR	BUDGETING APPROACH
Seat-based subscription (Copilot, ChatGPT Enterprise, Claude, etc.)	Predictable; scales with headcount	Model like any other per-seat software licence; true up quarterly against actual seats used, not seats purchased
Usage-based API and inference (tokens, GPU-hours)	Variable; can spike without warning	Set consumption forecasts with wide scenario bands; use quotas, alerts, and showbacks (See Part D)
AI embedded-in-SaaS and OT	Often invisible until renewal	Require vendors to disclose AI-attributable price increases at renewal; negotiate audit rights into the contract
Build and customisation (engineering, integration, data prep, fine-tuning)	Frequently underestimated	Budget as a project with its own line, not as "implementation" buried inside the licence cost

App-based AI is usually priced per seat or per subscription and scales with headcount, while token-based AI is priced per unit of consumption and scales with usage. An initiative that improves an existing process may well be token-based, and building a new capability may arrive as an app, hence a budget needs to be set against both. The reality however is that the pricing models for the frontier providers will continue to evolve as they have already begun moving from bundled token allowances to committed-consumption pricing, and the SaaS and OT vendors that today absorb token cost behind a flat fee are unlikely to keep doing so indefinitely. Anything budgeted this year as a fixed, app-based cost should be stress-tested for what it looks like if it re-prices toward consumption at renewal.

Specifically for token costs, the following are additional budget measures that can be considered to avoid surprises:

- **Put hard guardrails on usage**. For example, quotas, throttling, and anomaly detection so a spike is caught in days, not at the next invoice.
- **Buy predictability where the workload allows it.** Batch and scheduled jobs that have a calculable cost before they run. Commitment-based pricing can help but model the break-even explicitly first rather than accepting a vendor's default tier.
- **Match the deployment model to how uncertain demand is.** Cloud/API consumption is an operating expense that flexes with usage which usually is the right choice where demand is uncertain or spiky. Self-hosting only pays off at high, steady volume.
- **Negotiate contracts with the volatility of this market in mind**. Price-protection or most-favoured-pricing clauses given how fast list prices move, clear exit and data-portability terms given the pace of model deprecation, and audit rights into any embedded-AI SaaS renewal so token consumption isn't a black box.
- **Avoid multi-year, high-minimum commitments if usage patterns and model pricing are volatile**. Flexibility is worth paying a small premium for at this stage of the market.

C. How to Monitor Returns From AI Investments

For a tool that improves something you already do, you can measure the "before" and the "after". For an investment that builds a capability you do not yet have, there is no "before" state to compare against, because the capability did not exist. You are not measuring efficiency; you are measuring option value and market position. Hence by applying a productivity-tool ROI calculation (time saved × hourly rate) to this category will systematically undervalue it and can kill a genuinely good AI initiative.

To comprehensively monitor and value the costs and benefits from new-capability AI investments, we propose the following framework:

Use stage-gated milestones, not one number upfront

Evaluate new-capability AI investments in three gates, the way R&D and venture investment already do:

GATE 1

Feasibility gate

Does the technical approach work at all, on a small budget and short timeline?

GATE 2

Pilot gate

Does it work with a real use case at small scale, against a defined success metric such as an accuracy threshold or cycle-time improvement?

GATE 3

Scale gate

Does the unit economics hold as volume increases?

Each gate carries its own budget and its own kill criteria, so a full year's budget is never sunk into an initiative that fails a fundamental assumption in month two.

Real options framing

Value the investment as buying the right to scale later, not a commitment to scale now. A modest pilot that proves or disproves a new use case has value, even if it "loses money" on its own as you have bought information that changes a much larger future decision. The right question shifts from "did this pay back?" to "did this resolve enough uncertainty to justify the next, bigger decision?".

The competitive-necessity test

Some AI capability investments are defensive and not measurable in ROI. The question is what it costs not to have this capability in eighteen months, when competitors do or a key customer expects this capability as a standard or required offering. That is a judgment call and it should be made explicit at approval, not smuggled into an ROI number to make the decision look more rigorous than it is.

Leading indicators before financial ones

Early-stage new-capability investments will not show financial return for some time. Instead, track proxies: technical performance (model accuracy, latency, false-positive rate etc., whatever the capability's core metric is), real engagement (did an actual plant, line, or customer agree to test it,

extend the pilot, or pay for it), and progress against the capability roadmap. Financial ROI is a lagging validation once the capability reaches commercial or operational scale and for this reason, it should not be the gate that decides whether early-stage work continues.

Reference-class comparison instead of false precision

A full NPV model on a new-capability investment is based on so many assumptions which are inherently difficult to prove or disprove. It is often more honest to ask what comparable investments, yours or industry peers', cost to reach commercial viability, and how this initiative compares. This benchmarking grounds the decision in the real world, without pretending to a precision the inputs cannot support.

ILLUSTRATIVE EXAMPLE: PREDICTIVE MAINTENANCE ON A PRODUCTION LINE

Key questions to ask at each gate:

Feasibility gate
Can the model predict failure mode X from existing sensor data at all, using six weeks of historical data and a small team?

Pilot gate
Does the model achieve an agreed precision/recall threshold and an acceptable false-alarm rate, tracked weekly as leading indicators?

Scale gate
Once the pilot clears its threshold, what are unplanned-downtime hours avoided, multiplied by the fully loaded cost per hour of downtime on that line, compared against the all-in cost of the rollout across additional lines.

The financial ROI number only becomes meaningful at the scale gate. Using ROI to judge the feasibility or pilot gate may have killed the initiative before it had produced the evidence needed to judge it fairly.

Defining "cost per outcome" and linking it all back to financials

"Hours saved" no longer clears the bar with most boards, because it does not show how saved hours reached revenue, margin, or risk reduction. The more durable metric is cost per outcome which can be broadly defined as follows:

$$\text{Cost per outcome} = \frac{\text{Fully loaded cost of the AI}}{\text{Number of business outcomes produced in the period}}$$

Fully loaded cost of the AI
(i.e. licences, inference/token spend, integration, and the change management needed to get real adoption, See the AI cost stack in Exhibit 2)

Number of business outcomes produced in the period
(e.g. inspections completed to spec, maintenance interventions correctly triggered, quotes generated, cases resolved)

Tracking this quarter over quarter shows whether unit economics are improving as the team learns to use the tool and as engineering-side cost levers (described in Part D) bring the denominator down, independent of whether overall volume is rising or falling.

Who owns the measurement, and why that matters

New-capability bets are more prone to sunk-cost continuation than tactical tool spend because fuzzy ROI calculations can justify almost anything in hindsight.

A recurring failure is letting the team that built or bought the AI capability also be the one that reports on its ROI. We recommend that Finance teams should own the measurement framework and the baseline. The baseline should be captured before go-live, not reconstructed after the AI solution is implemented. The business and technical teams should still continue to own the leading indicators.

Separately, give the two portfolios distinct budget lines and distinct approval committees because by mixing them means new-capability bets get judged by the wrong yardstick, or improved-capability tools get over-scrutinised because they are bundled in with riskier strategic spend. Set explicit time horizons and re-evaluation checkpoints at approval and name a single person accountable not just the fund decision, but also for the kill decision.

D. Promoting AI Usage Without Letting Costs Run Wild

For most industrial companies the larger risk over the next two years is under-adoption, not overspend. A finance function that treats every AI request as a cost threat will slow the initiatives with the best potential returns alongside the ones that were never going to work. The goal is to make good usage easy and visible, and to make waste visible and correctable, without defaulting to a blanket restriction as the primary control.

The following are some suggestions on how to promote AI use in your organisation without letting costs go out of control:

Design the access model in tiers

Make low-cost, low-risk tools, for example, general chat assistants for drafting, summarising, and research broadly available with minimal approval friction. This is where most of the workforce-wide productivity gain sits, and gating it heavily just pushes usage into unmanaged personal accounts. Gate higher-cost or higher-risk tools like API access, agentic systems, anything connected to plant, customer, or safety data behind a lightweight business case or team-lead sign-off. Controls should be risk adjusted proportionate to the actual risk rather than uniform across every tool.

Tie budget to value, not headcount

Allocate AI budget for departments against a metric that grows with expected benefit, for example, share of throughput, transaction volume, or revenue rather than an open-ended line-item allocation. This keeps the budget conversation anchored to what the spend is supposed to produce and gives a department justification to ask for more when it can show the growth in usage is earning its keep.

Make cost visible to the people generating it: showbacks and chargebacks

Showback reports usage and cost back to teams for visibility while keeping the expense on a central budget. It is low-friction, requires no ERP integration, and is the right starting point while cost-allocation data is still being trusted.

Chargeback formally assigns the cost to a department's own budget, which creates a direct feedback loop but it requires mature usage-tagging (most FinOps practitioners look for at least 80% coverage) and finance alignment on how shared costs get split.

Many industrial companies land on a hybrid: showbacks broadly across the organisation, with chargebacks applied only to the largest consumers of usage-based spend, where the incentive effect matters most. Whichever model you choose, the single highest-leverage nudge is showing people the cost of what they are doing while they are doing it. A per-query or per-report cost estimate inside the tool itself changes behaviour far more effectively than a monthly report that arrives after the spend is already committed.

Pull the engineering-side cost levers before reaching for a cap

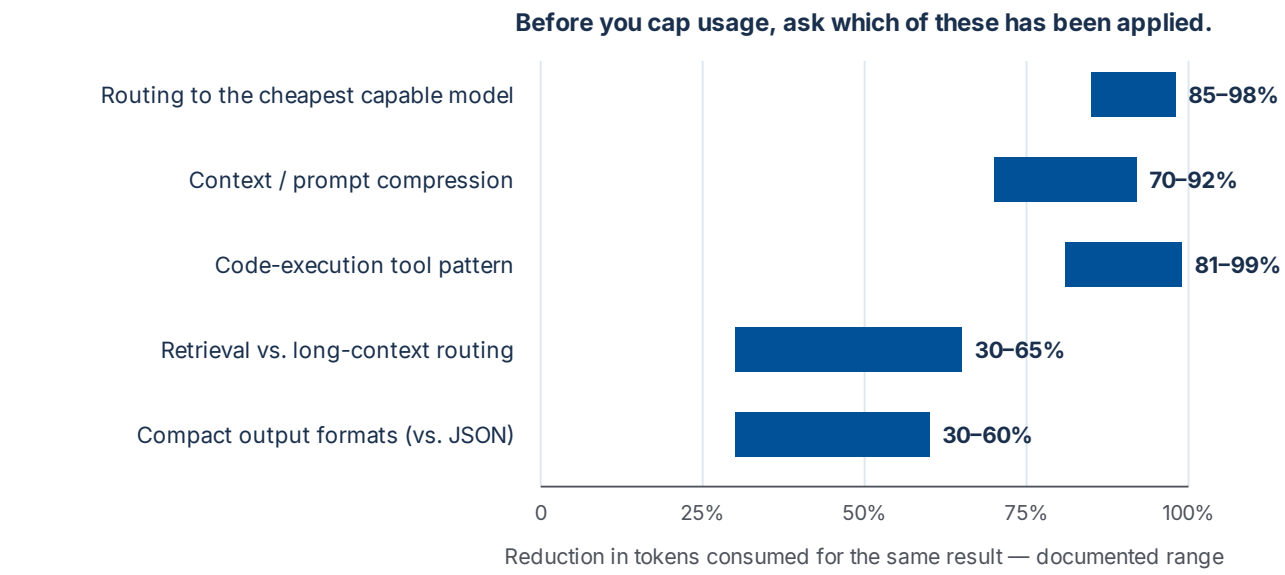
A meaningful share of token cost is addressable by the AI or engineering team without touching usage volume at all, and CFOs should use these levers before placing usage caps to control a rising AI bill:

- **Model routing and cascading**. Sending each query to the cheapest model capable of answering it well, rather than defaulting every request to the most expensive model. Published approaches (FrugalGPT, RouteLLM) have demonstrated cost reductions in the 85–98% range against a most-expensive-model baseline, for equivalent output quality.
- **Prompt and context compression.** Trimming retrieved documents, conversation history, and system-prompt overhead down to what the query actually needs; documented approaches report 70–90% token reductions on affected workloads with no quality loss.
- **Semantic caching.** Recognising when a new query is asking substantially the same thing as a recent one and reusing the prior computation rather than re-running inference.
- **Structured output formats.** Requesting compact formats for tabular or structured data instead of verbose formats, which alone can cut relevant token volume 30–60%.
- **Batching and off-peak scheduling.** For workloads that do not need a real-time answer, taking advantage of lower-cost batch pricing where providers offer it.

None of this requires the CFO to become a prompt engineer, but it requires asking the AI/engineering team as a standing question in every cost review, "which of the known efficiency levers have we applied to this workload, and which haven't we?" before approving a usage cap as the fix.

One caveat worth stating plainly: these levers are a capability, not a one-off exercise. They require engineers who know the current techniques, and these techniques keep changing; what saves 80% of tokens this year may be superseded in the next. If that skill is not in-house, the honest options are to buy it in or to accept a higher cost per token. The economics also only work above a threshold: where a 50% reduction on a workload is worth less than the fully loaded engineering time needed to achieve it, then the lever is not worth pulling. Name an owner for AI unit economics, and revisit that threshold as spend grows.

EXHIBIT 6: ANECDOTAL SAVINGS FROM USING ENGINEERING-SIDE COST LEVERS



Published best-case results on specific workloads (FrugalGPT, RouteLLM, Cloudflare, Zilliz, Flexpa). They are NOT additive and will not all transfer to yours.

Invest in adoption deliberately, not just permission

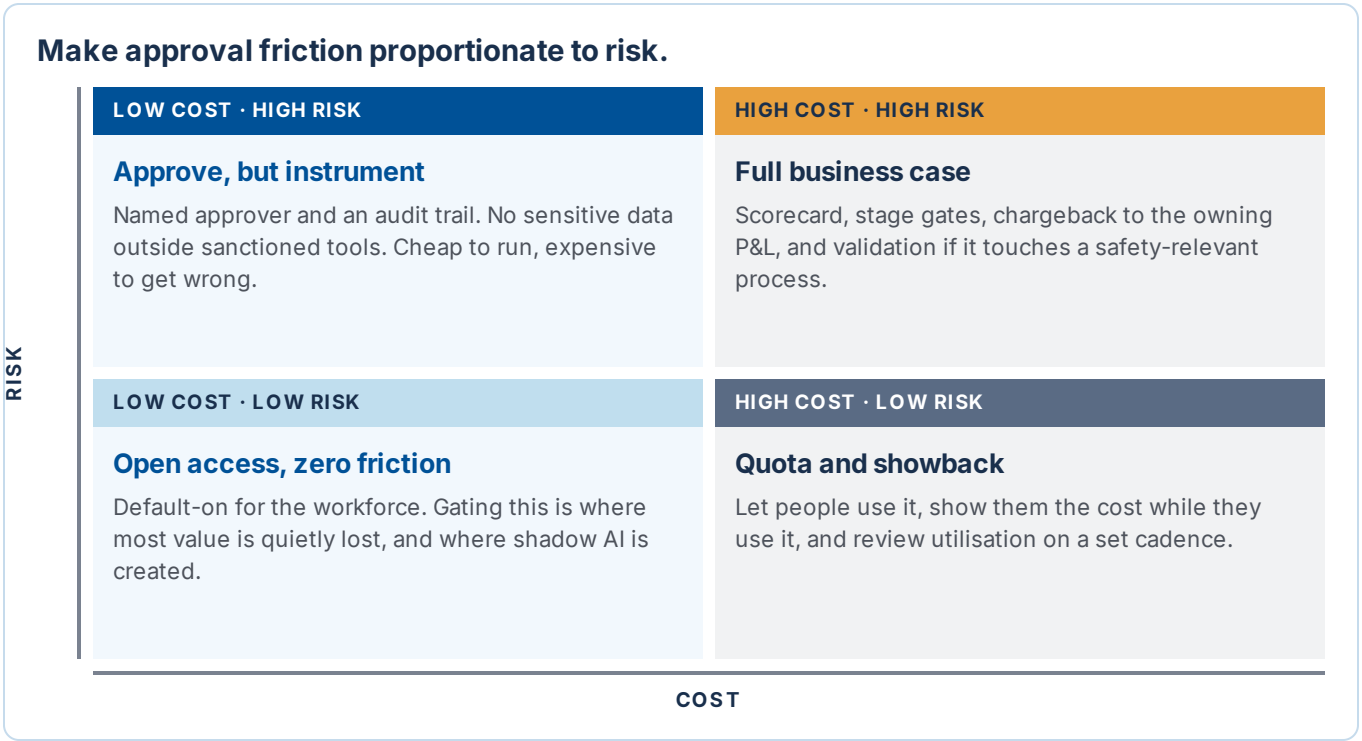
Identify a small number of high-visibility pilots per function, paired with internal advocates who drive adoption peer-to-peer rather than top-down. In most cases, usage grows faster and more sustainably through a colleague showing a genuinely useful workflow than through a mandate. Set aside real training hours during working hours rather than treating AI literacy as something employees absorb in their own time. Using tools the company has already paid for is usually the cheapest efficiency gain available.

Put a governance structure around the whole system

Where the size of the AI budget for the organisation is a significant, having a proper AI cost governance structure will be more effective in identifying and controlling AI spend. Some ways to achieve this include:

- Having a cross-functional AI steering body comprising of finance, IT/engineering, and the business-unit owners to evaluate, allocate and prioritise AI budgets instead of leaving usage-cap decisions solely to procurement or IT.
- Embed a finance business partner within the AI/data team so that cost discipline is continuously monitored instead of being reviewed on a quarterly basis from the outside.
- Periodic utilisation review across the whole tool portfolio: downsize or retire subscriptions with low utilisation on a set cadence, rather than letting renewals auto-continue by default.
- Fast-track procurement for tools IT has pre-vetted as low-cost and low-risk, so the approval friction is proportionate to actual exposure instead of having a flat process applied to everything.

EXHIBIT 7: MAKE APPROVAL FRICTION PROPORTIONATE TO THE RISK



Key takeaways: Ten Actions for CFOs

AI spend will keep behaving like a metered utility for the foreseeable future: usage-driven, price-volatile, and hard to attribute cleanly to a business outcome. Hence there is an urgent need to build a lighter, faster, more continuous governance model for AI investments or risk falling behind competitors that supercharge their processes with AI successfully. To ensure your organisation purposefully adopts AI into its workflows without letting AI costs run wild, consider adopting these ten measures:

1. Split the AI budget into two portfolios: improving what you already do, and building what you don't have yet, with separate owners, gates and success measures.
2. Build one consolidated view of AI spend, including embedded and shadow spend, even if it starts imprecise.
3. Require the six-question scorecard (strategic fit, data readiness, technology and operational readiness, total cost of ownership, kill criteria, risk exposure) before funding any AI proposal.
4. Stage-gate new-capability AI funding, with pre-agreed kill criteria and naming who owns the kill decision.
5. Put quotas, alerts, anomaly detection on every usage-based AI cost and re-forecast quarterly.
6. Track leading indicators (technical performance, real engagement, roadmap progress) for new-capability bets, and reserve financial ROI as the validation gate for scale decisions.
7. Adopt cost per outcome as the standard unit metric, with a baseline captured before go-live.
8. Ask which engineering cost levers have been applied before approving a usage cap as the fix.
9. Make the cost of usage visible to the people generating it because visibility, not restriction, is what sustainably keeps adoption growing and cost under control at the same time.
10. Stress-test every app-based and embedded-AI contract for what it costs if it re-prices toward consumption at renewal, and negotiate price protection and audit rights before that happens.

References

This paper draws on the following primary sources, alongside Aicadium's own client work with industrial companies.

- FinOps Foundation — Token Economics: The Atomic Unit of AI Value — https://www.finops.org/insights/token-economics-the-atomic-unit-of-ai-value/
- FinOps Foundation — FinOps for AI Working Group — https://www.finops.org/wg/finops-for-ai-overview/
- FinOps Foundation — Token Economics for SaaS — https://www.finops.org/wg/token-economics-saas/
- Harvard Business Review — Manage Your AI Investments Like a Portfolio (2026) — https://hbr.org/2026/01/manage-your-ai-investments-like-a-portfolio
- Forbes — Why Finance Transformation Topped the CFO Agenda in 2026 — https://www.forbes.com/sites/anders-liu-lindberg/2026/07/09/why-finance-transformation-topped-the-cfo-agenda-in-2026/
- CFO.com — CFOs Targeting Both Business Growth and Cost Reductions in 2026 (Gartner) — https://www.cfo.com/news/cfo-targeting-both-business-growth-cost-reductions-2026-gartner-dennis-gannon-risk-ai-expenses/808037/
- CFO Brew — CFOs Look to Optimize Costs in 2026 But Also Want to Invest in Growth — https://www.cfobrew.com/stories/2025/12/23/cfos-look-to-optimize-costs-in-2026-but-also-want-to-invest-in-growth-survey-says
- The CFO — Beyond the Pilot: How CFOs Can Master the AI Value Chain — https://the-cfo.io/2026/04/07/beyond-the-pilot-how-cfos-can-master-the-ai-value-chain/
- TechTarget — FinOps for AI: How CIOs Are Navigating Tokenomics — https://www.techtarget.com/searchcio/feature/finops-for-ai-how-cios-are-navigating-tokenomics
- CloudZero — Chargeback vs. Showback: Cloud Cost Allocation Models Explained — https://www.cloudzero.com/blog/chargeback-vs-showback/
- Evanta — Top 3 Priorities for CFOs in 2026 — https://www.evanta.com/resources/cfo/survey-report/top-3-priorities-for-cfos-in-2026

Additional directional data points on cost overruns, forecast accuracy, and FinOps scope expansion are drawn from FinOps Foundation community reporting and industry surveys.

Disclaimer

This document provides general information and is not intended as professional advice. It reflects the views of the Aicadium team and does not represent the position of any client, portfolio company, or individual practitioners involved in its preparation. Aicadium does not act as a professional adviser in any of these areas.

The paper adopts an outside-in perspective, based on published research, community reports, industry surveys, and Aicadium's own client work. Figures, allocations, and examples are labelled as such and should not be interpreted as benchmarks or forecasts for any organisation.

Nothing here should be relied on for any specific business, financial, legal, or personnel decision. Regulations, vendor terms, and tool capabilities change quickly, so read anything time-sensitive against current sources and take advice from your own qualified advisers before acting on it.

aicadium

aicadium.ai