ARTIFICIAL INTELLIGENCE

Return on AI: How CFOs and CIOs Can Manage the Token Meter

By Joppe Bijlsma, Djon Kleine, and Filippo Scognamiglio

ARTICLE JULY 01, 2026 8 MIN READ

This is the second of two articles on the costs of AI tokens and how companies can manage them. The first article examined the true costs of AI . This article details the challenges of measuring them.

The first wave of AI governance was about access. The next wave will be about economics.

As AI moves into production, CFOs, CIOs, and CTOs inherit a big new cost to manage: the tokens consumed to produce business outcomes. To date, the costs associated with tokens, the basic units of data that AI models use, have been largely buried in the IT budget and mostly managed using the principles and procedures of FinOps. But as AI usage scales through the organization—from software development, where it is most advanced, to sales, marketing, operations and other functions—AI token bills are exploding. The rising use of AI agents, which consume tokens in quantity, pushes the meter into overdrive. FinOps cannot keep pace.

CEOs will expect their C-suite colleagues to rise to the challenge. Here's how they can control AI costs and focus the technology on driving outcomes rather than just activity.

FinOps Is a Starting Point, Not an Answer

FinOps has real strengths. It is good at exposing infrastructure cost, assigning accountability, and optimizing usage. But AI works differently than traditional software as a service, and it changes the unit of management. A token bill depends on a host of factors, including the length of prompts, the context that needs to be retrieved, the quantity of output, choice of model, reasoning effort, choice of tools, cache behavior, and the number of loops an agent runs before it stops.

Token cost is not automatically exponential, and it can vary widely across platforms and models. The key metric to track is not simply cost but cost per outcome. A ratio that we call return on AI (RoAI) is a good way to capture the full cost of the AI being applied:

$$RoAI = \frac{\textit{Economic return}}{\textit{Cost of human Intelligence} + \textit{Cost of tokens}}$$

In unmanaged agentic workflows, tokens consumed per useful outcome can compound sharply and invisibly, the result of four forces that interact in ways that traditional software and infrastructure forecasts miss.

The first force is breadth and depth of adoption. Token cost grows as users move from simple chats to multistep research, code generation, agent deployment, and workflow orchestration. The second is task intensity: a short answer, a document review, and a long-running agent session may all look like one request, but each has very different economics.

The third force is context and loops. Agents carry large volumes of information forward, including instructions, history, retrieved documents, and tool outputs, and often replicate them in each loop, which consumes lots of tokens. Fourth is model mix: defaulting to the newest frontier model, the highest reasoning setting, or the largest context window is a design choice with a big billing impact. The inverse can also be true. A weaker model that needs multiple retries to reach an answer can cost more per successful outcome than a stronger model that gets to the same place quickly.

Since unit costs rise linearly with tokens consumed, the right optimization target is cost per successful outcome at the required quality, latency, risk, and additional burden of human review. The challenge is to make this denominator visible and controllable.

A New Cost Management Operating Model

Token costs do not belong in the IT budget, nor should they be assigned to a single line anywhere in the P&L. Token costs come in three types, and each should be accounted for differently. Tokens that build reusable capability are an investment, similar to capex. When AI is used to design agents, redesign workflows across functions, and—for those that run their own AI inference—provide execution at scale, AI infrastructure is capital expense. Tokens that run internal work are an operating expense. CFOs can set a sensible budget per function and per workflow and hold mangers accountable for the output. Tokens inside a product or customer interaction are cost of goods sold (COGS) and must be managed as part of gross margin, where they can have a significant impact and that impact is fully transparent.

Token cost transparency is especially important for AI-enabled products. The classic economics of software as a service benefited from scale because once the software was licensed, each new customer was added at virtually no cost. AI inference alters the cost curve: each customer interaction carries an associated cost from running the model. Management needs to be able to see the impact on gross margins of absorbing inference at scale, including price, usage, model architecture, caching, and routing.

To accurately account for costs and assess RoAI, companies need a workflow-level operating model that enables management to do three things: see what is happening, shape the cost, and either prove value or stop (or minimize) the activity.

See: Build Attribution Before Optimization

The quickest path is to start with budget-owner attribution and move toward outcome-level attribution. Every material workflow should be assigned to an owner, function, product-versus-internal use, P&L line, model, provider, and intended outcome. Design a dashboard to track what is happening. The first version can be simple: separate workspaces or projects for each initiative, disciplined API keys, clear product-versus-internal tags, and provider identification.

The dashboard should feature questions that leaders can act on, such as the following:

- Which workflows consume the most tokens, and what outcomes do they produce?
- What is the cost per successful outcome?
- What share of spending has an assigned owner?
- What is internal opex versus product COGS?
- What percentage of input tokens is cached?
- How much work defaults to frontier models?

Spending without an owner is ungoverned. Spending without a defined outcome is unproven. Spending without an assigned P&L line invites mismanagement.

Shape: Four Technical and Behavioral Levers

Organizations need to learn how to manage AI and its associated costs, which is probably a joint responsibility of the CIO or CTO and functional and business unit leaders. They should focus on four areas.

Eliminate unnecessary model use. Many early agents exist because they were fast to build, not because a model was the right tool. Structured lookups, rules-based routing, arithmetic, and deterministic policy checks are tasks for software, APIs, or reusable skills, not AI. Routing work to traditional software is itself an effective cost lever, and it often improves reliability as well.

Route by task complexity. Match low-complexity work with lighter or open-weight models, and reserve frontier models for genuinely hard reasoning. Include reasoning effort settings and context window choices in the dashboard, since default settings often overprovide. Let systems autoroute on quality, confidence, latency, and cost rather than sending everything to the strongest model.

Reuse context and components. Prompt caching and batch processing can materially reduce repeated-context and asynchronous workload costs when designed into the workflow. Widely used policies and procedures, such as standard instructions, brand rules, compliance requirements, and common workflows, should become stable, cache-friendly prompt prefixes. They can be included in curated knowledge packs, approved templates, and skill libraries for easy access.

Train for token discipline. Good behavior and discipline lower token costs. Users should learn to ask for the smallest useful answer, constrain output length, avoid pasting unnecessary documents, and distinguish exploration from execution. Interfaces should default to concise outputs, retrieval discipline, and clear stopping rules. Brevity becomes a valuable cost control lever in a metered system.

Prove, Stop, or Minimize: Govern Return, Not Activity

Leaders should establish a standing review process for the highest-token-consuming workflows. Each review should assess the business numerator (outcome) against the full denominator: token cost plus the human time to initiate, review, correct, approve, and operate the workflow. They can then decide whether to scale up, optimize, rescope, or stop the workflow.

Software engineering shows why this type of review matters. An agentic coding workflow can create a large amount of output at a small token cost, but lines of code are an activity metric. The numerator should count code that is actually shipped and survives review. The denominator should include specification, review, correction, testing, and approval time. The same workflow can be a large positive return or a negative return, depending on how much useful, accepted output it produces and how much human effort it consumes.

The same logic applies outside of engineering. A service agent who closes more tickets but drives more escalations is producing activity rather than return. So is a marketing agent who generates assets that no one uses. The key governance question is not whether AI is busy. It is whether the outcome improved after counting the full cost.

Savings Can Start Now and Build Over 12 Months

We estimate that companies can substantially improve the efficiency of their AI spending over 12 months by both reducing token costs and achieving more bang from each token, although the range of savings will vary from firm to firm. (See the exhibit.) The key areas of savings opportunity are:

- Stopping or minimizing avoidable spending
- Routing work to the right workhorse
- Caching content to avoid repetitious AI activity
- Applying the right controls
- Training staff to increase AI literacy

Companies Can Manage the Cost of Each Successful Outcome

Opportunity area	Lever/What changes	Impact (improved efficiency of AI bill)	Pricing mechanic exploited
Stop or minimize Prevent avoidable spending	**Revisit low-ROI usage** Reduce avoidable spending by stopping unnecessary agentic loops	~ 5%	No price change; eliminates unnecessary usage
	Push work out of AI models Use deterministic code, rules, and APIs instead	~ 2%–5%	Rules and lookups replace billed calls
Route Pick the right workhorse	**Rightsize and route the model** Send each task to the cheapest model that meets the intelligence required	~ 3%–8%	Tiered per-token rate cards
	Rightsize GPUs and self-hosting Self-host only when usage is steady and high	~ 2%–3%	GPUs bill hourly; APIs bill per use
	Budget output and reasoning Cap session length; use frontier reasoning only when needed	~ 5%–10%	Output costs 3x–5x input; reasoning bills as output
Cache Ask the right questions	**Cache and compact the context** Reuse stable openings, system prompts, policies, and knowledge bases	~ 3%–12%	Cached text bills at a fraction
Govern Apply the right controls	**Govern to cost per outcome** Assign every workflow an owner, P&L line, and decision threshold	~ 5%–10%	No price change; enforces spending accountability
	Re-architect the agent flow Share a workspace; pass goals, not full logs	~ 5%–10%	Fewer calls and retries, not a lower price
	Buy at the right rate Batch nonurgent work; commit when steady	~ 3%–6%	Batch ~50% off on-demand; volume discounts
Train Increase AI literacy	**Better use of routing and caching** Train employees to understand what they are trying to achieve	Drives rapid adoption of levers above	Levers covered above

Sources: Published cloud and model provider rate cards; Gen AI pricing data; BCG analysis.
Note: Per-token mechanics are based on list prices and do not take into account negotiated discounts; share-of-bill figures are scored estimates that depend on workload mix and are not additive.

It helps to think in terms of three phases: quick wins (months 1 to 3), big rocks (months 3 to 6), and the longer tail (months 6 to 12). The quick wins fund the big rocks, whose impact is compounded by a longer tail of savings. Companies can gain control over the RoAI denominator with financial impact that accumulates over time.

Token costs are real and growing fast. They are attracting CEO and board-level attention. CFOs, CIOs, and CTOs need to be ready with answers when those leaders start asking questions.

The authors are grateful to these BCG colleagues for their ideas and input: Nicolas De Bellefonds, Abhinav Gupta, Steven Kok, Matthew Kropp, Vladimir Lukic, Clark O'Niell, Rohan Panjwani, and Vikram Srikumar.

Authors

Joppe Bijlsma

Managing Director & Partner
New York

Djon Kleine

Managing Director & Partner
San Francisco - Bay Area

Filippo Scognamiglio

Managing Director & Partner; BCG Henderson Institute Functional Leader; Global Cloud Advisory Business Leader
New York

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change — bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.