48.8744° N
2.2965° E

The AI Compass

ISSUE 02

Stop buying AI by price.
Buy it by value.

Capgemini

Table of contents

Executive Summary — 04

Part I – Why this matters now — 06

The cost of conversartion everyone is having — 06

Part II – Our perspective — 10

An attribution bias: a billing metric is not a unit of value — 10

Part III – What it means for CEOs and CAIOs — 14

Price vs value: What the evidence shows — 14

Part IV – Making it real — 21

From FinOps to ValueOps — 21

Conclusion — 28

Buy the value, not the token — 28

Sources & notes — 32

Executive summary

Enterprise leaders are confronting a strange contradiction: AI is getting cheaper by the token, yet their overall AI costs keep rising. So, how do we cut the token bill? It sounds like the obvious question to ask, but it's the wrong one.

The price of a token is observable, falling fast, and converging across vendors within a capability tier. But price tells us nothing about value. This value depends entirely on the outcome and whether the workflow it is used for successfully completes the task and contributes at the standard the business expects. Reliability, therefore, matters enormously, and confusing a token's cost and its value can undermine the economics of AI.

If an AI model isn't reliable enough to perform a task well and consistently, then a cheaper token may not be just creating less value, it may also be destroying it. This happens through errors that propagate throughout multi-step tasks, demand intensive human oversight to contain, and can trigger such a significant amount of loss that an entire year of token savings is erased in a single moment.

The metric that matters most is a correctly completed workflow. And a token's worth depends on whether the model is reliable enough to make errors a negligible risk. Reliability is precisely what the market does not put a price on, because it is context-specific and because some models tend to overthink and generate useless tokens.[1]

In our second issue of “The AI Compass”, we share our research which shows that price does not stand in for value, and within a capability tier, the two are uncorrelated. At realistic volumes, the token bill is 2 to 3 orders of magnitude smaller than both the value created by a correctly completed workflow, and the cost of the errors triggered by a faulty one. Paying more does not necessarily buy more value, but paying less can destroy it.

For a CEO, the implication is direct. An AI cost program that chases price per token while ignoring reliability will cut spending, destroy value, and still report the saving as a win. Instead, set up an operating model in which:

- You define what a correctly completed workflow is worth, and what an incorrect one costs.
- You set reliability targets based on the P&L stakes of each workflow, not on the price of the tokens it consumes, and hold each engineering team accountable for meeting them.

Part I – Why this matters now

The cost conversation everyone is having

The cost narrative is persuasive because it is directionally correct, but it is incomplete. Inference has collapsed in price. For example, the cost of GPT-3.5-level output fell from about $20 to roughly $0.07 per million tokens between late 2022 and late 2024.[2]

However, demand is elastic across providers and every 10% reduction in price buys an estimated 12–18% increase in usage, so total spend keeps rising even as unit prices fall.[3] Agents multiply consumption further, as a single conversational query runs on the order of a thousand tokens; an agentic workflow, a thousand times more; a continuously running agent, a thousandfold again. One orchestrated interaction that cost about four cents in 2023 now costs over a dollar (Figure 1). Budgets are straining, and the reflex is to meter AI usage, impose limits, and route workloads to the cheapest model that clears the bar.

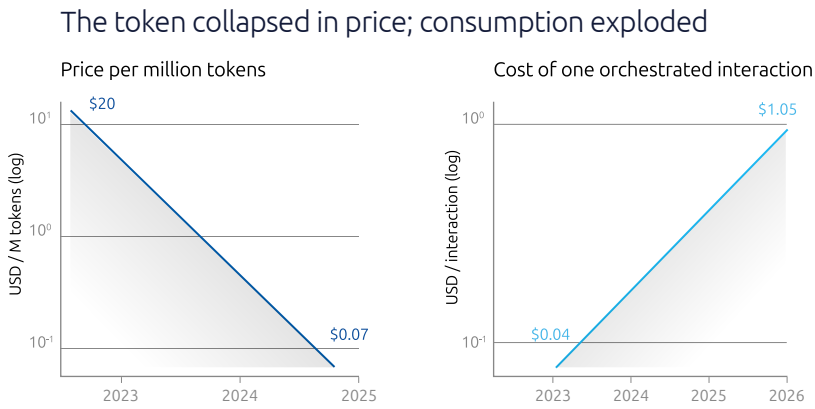
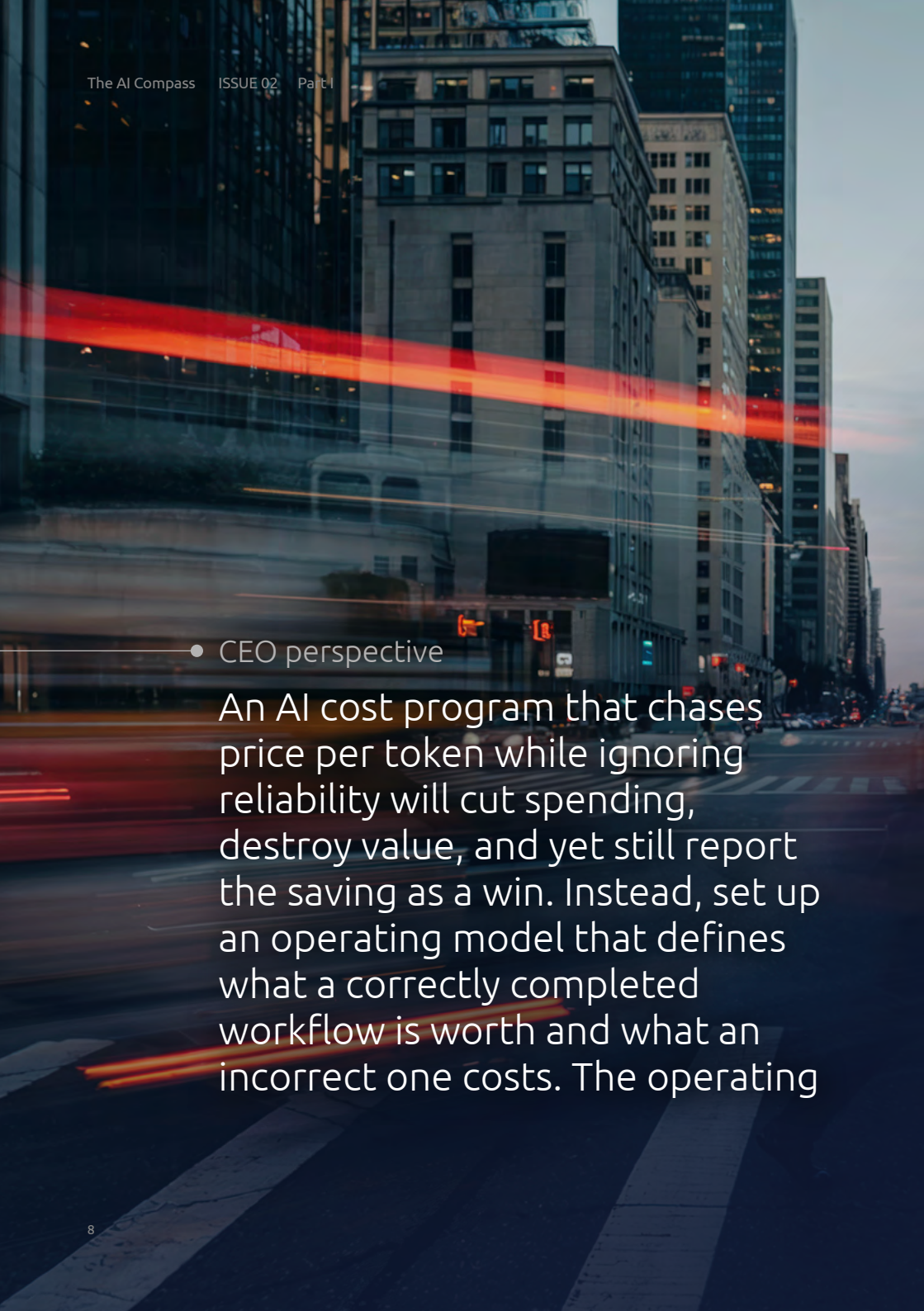


Figure 1. The token collapsed in price while consumption exploded. Inference fell roughly 99% between 2022 and 2024; a single orchestrated interaction now costs about 25 times its 2023 level, and demand elasticity (1.2–1.8) keeps total spend rising. [2,3]

None of that is wrong. However, it is incomplete in the way that matters most. Cost discipline answers, "How do we spend less?", but it does not answer, "What is each token producing?" Yet the second question is where the money is truly won or lost.

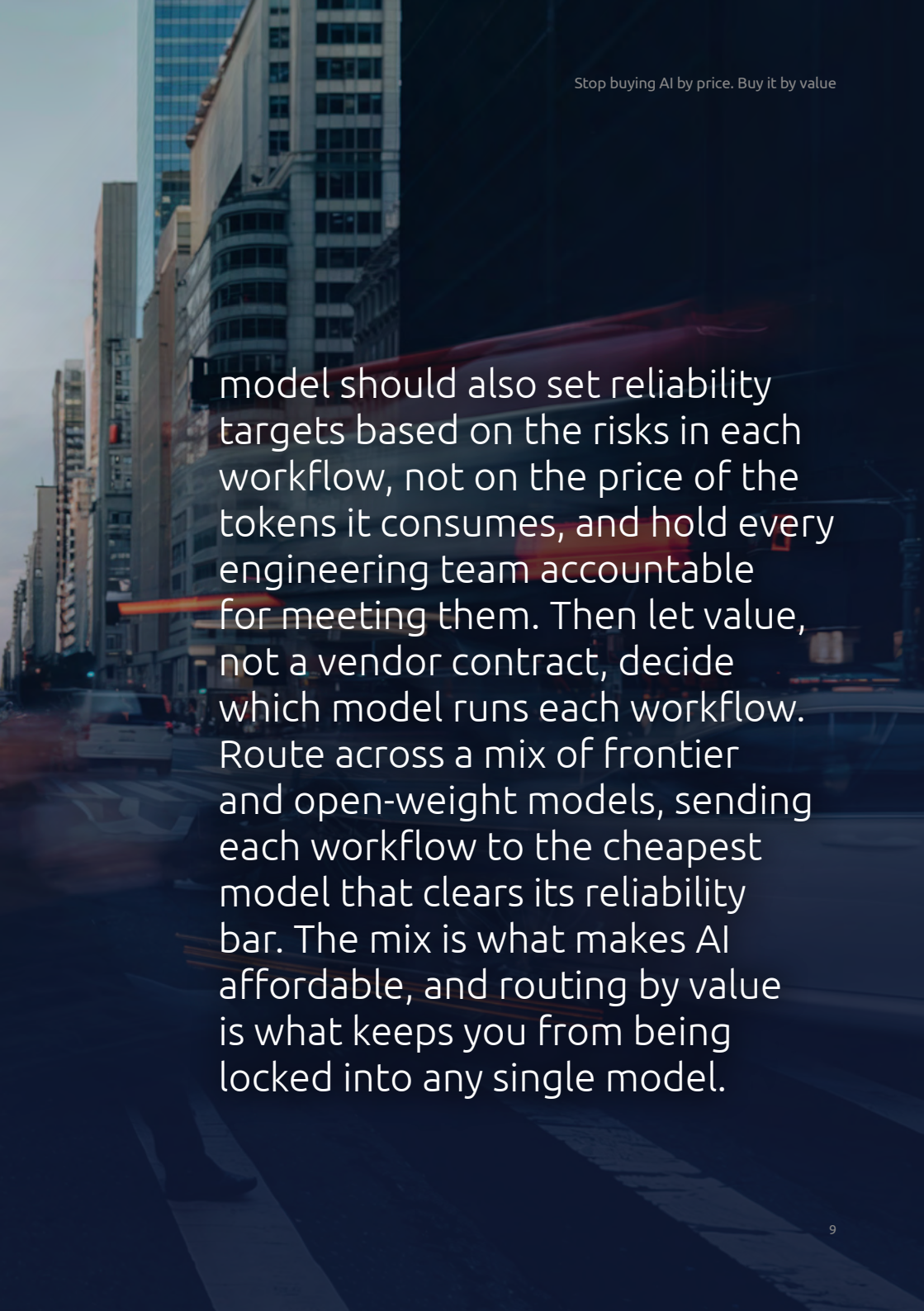
To illustrate further, the Capgemini Research Institute found in their latest AI economics snapshot that 59% of executives saw AI expenditure significantly exceed expectations, while most organizations still struggle to connect AI spending directly to a successful business outcome.[4]

Only 14% of CFOs report clear, measurable AI ROI, and the most common reason is that they are measuring inputs, not outputs. IDC research states that 88% of AI agent pilots never graduate to production deployment, citing unclear business objectives and value, insufficient data readiness, a lack of in-house expertise, and escalating costs.[4] The cost problem and the value problem are not separate. They are the same problem, viewed through different lenses.



CEO perspective

An AI cost program that chases price per token while ignoring reliability will cut spending, destroy value, and yet still report the saving as a win. Instead, set up an operating model that defines what a correctly completed workflow is worth and what an incorrect one costs. The operating



model should also set reliability targets based on the risks in each workflow, not on the price of the tokens it consumes, and hold every engineering team accountable for meeting them. Then let value, not a vendor contract, decide which model runs each workflow. Route across a mix of frontier and open-weight models, sending each workflow to the cheapest model that clears its reliability bar. The mix is what makes AI affordable, and routing by value is what keeps you from being locked into any single model.

Part II – Our perspective

An attribution bias: a billing metric is not a unit of value

Every general-purpose technology eventually settles on a unit of value, and it is rarely the unit on the invoice (Figure 2). Electricity was billed per installed light bulb before it was metered in kilowatt-hours. Online advertising moved from untracked banners to cost-per-click to cost-per-acquisition. The billing metric measures what the supplier delivers; the value metric measures what the buyer gets. For AI, the supplier delivers tokens and the token price is model-dependent. Correctly completed work is what buyers value. And determining the likelihood of having a correctly completed workflow from a model card is far from being straightforward.

Exponential View's 2026 State of the AI Economy, built from the ground up using data from more than a thousand firms, reaches the same conclusion and states it plainly: tokens are "AI's billing metric, but not yet a unit of value."[5] The report's own candidate for a usable unit, quality-adjusted output tokens, exists precisely because raw token count says nothing about whether the output was worth having.

The billed unit is rarely the one that captures value

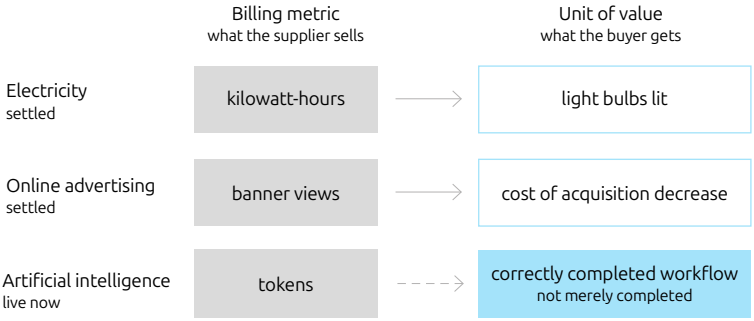
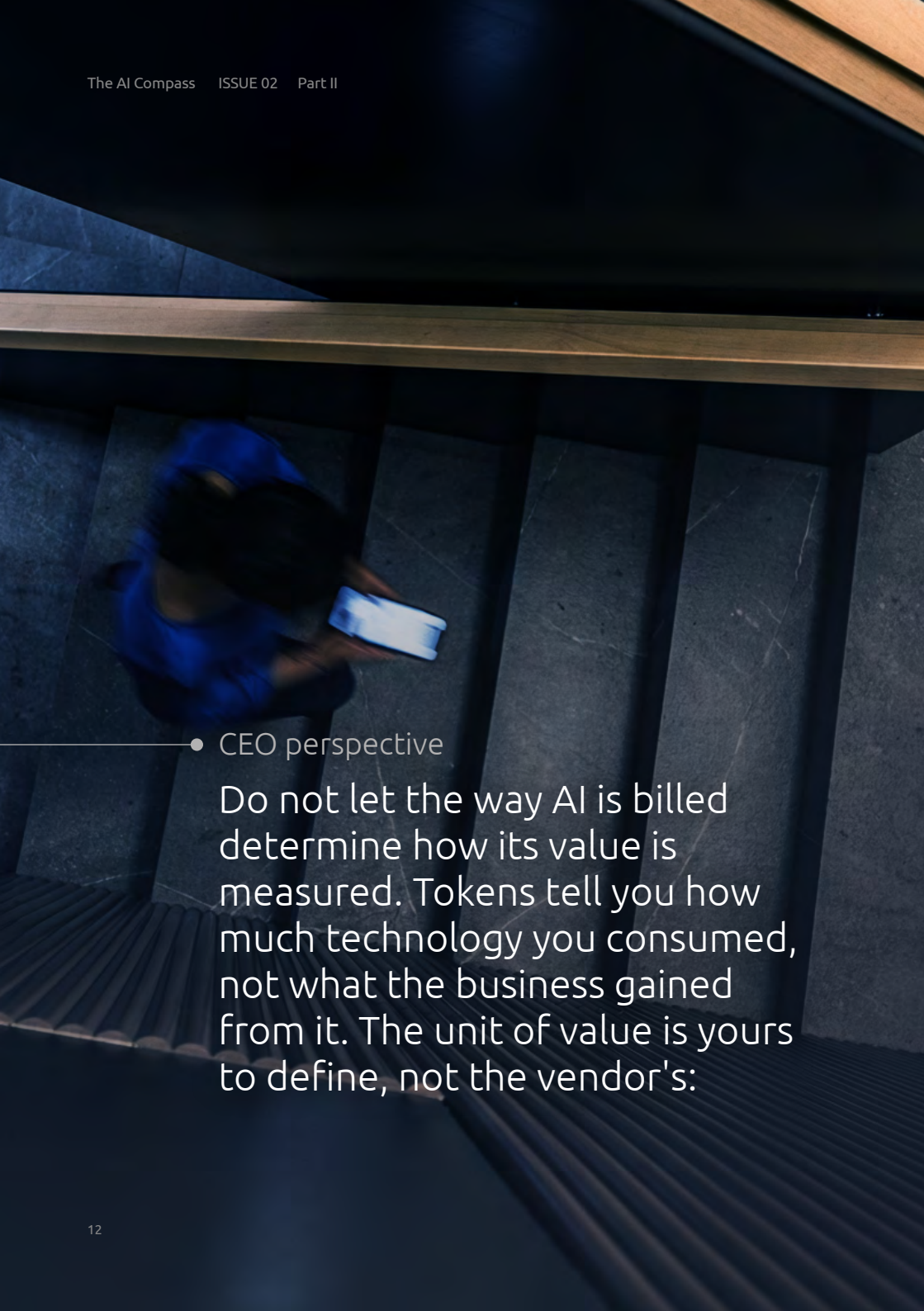


Figure 2. Every general-purpose technology settles on a unit of value that is not its billing metric. AI is mid-transition: the unit is not the completed workflow, but the correctly completed workflow — finished reliably enough to ship without a human redoing it (the more demanding version: a trusted workflow, completed correctly, and unsupervised).

The market has begun voting through its choice of vocabulary. Salesforce now reports "agentic work units," the work its agents complete, rather than the tokens they burn.[5] Yet even these measures stop short of capturing what ultimately matters: the business value of a completed workflow. A token cannot be a unit of value because value is not a property of the token itself. It is a property of the match between a model's reliability and a workflow's stakes. A token that completes a high-stakes workflow correctly is worth a great deal. A token that falls just short can be worth less than nothing, because it leaves errors for a human to catch or, worse, lets them reach a customer.



CEO perspective

Do not let the way AI is billed determine how its value is measured. Tokens tell you how much technology you consumed, not what the business gained from it. The unit of value is yours to define, not the vendor's:



a correctly completed workflow, priced in your own P&L, finished to the standard the business expects. Treat the proxies that vendors now publish, whether quality-adjusted tokens or agentic work units, as better billing metrics, not as measures of what you earned. A model card will not tell you whether that unit is being delivered; only repeated runs on your own workflows will. Then put the two numbers side by side in every AI report: what was spent on tokens, and what was correctly completed.

Part III – What it means for CEOs and CAIOs

Price vs. value: What the evidence shows

Our research formalizes the value of a token within a workflow-based model of production and calibrates that model using a public agentic benchmark, namely an airline customer-service workflow, across the mid-2026 frontier-model lineup.[6] The results are stated as signs and orderings, not precise dollar levels, and they hold across a wide range of assumptions.

Net value = value when it succeeds – cost of errors that escape – cost of fixing errors you catch – cost of human oversight – cost of tokens

The value a token produces within a workflow is equal to the value of the work when the model succeeds, after deducting four expenses: errors that slip through unchecked, errors that a reviewer catches and must fix, the review process itself, and the tokens consumed. Reliability determines how often value is realized and how often the errors incur costs, while oversight decides how many escaped errors become cheap fixes rather than expensive failures.

The token bill is the final term in the equation, and at enterprise volumes it is the smallest by two to three orders of magnitude. That captures the whole argument algebraically: a cost program focused on reducing token prices is targeting the part of the equation with the least influence. Price and value can even rank in opposite directions: the cheapest token is the most value-destroying, while a more expensive alternate may deliver almost no value or deliver the most.

When deciding which model is economically viable, two results matter most. First, for each enterprise process, there is a reliability floor, below which no amount of human review makes the economics positive. In other words, you cannot supervise your way out of an unreliable model. Second, even the top-tier models that create value on average carry a dangerous tail of catastrophic decisions. They can be profitable in most cases and yet lose heavily in the worst 5%, because one wrongful denied boarding decision costs far more than a routine success earns. Our analysis shows whether value is created or lost and how the options compare, rather than estimating precise dollar amounts. These conclusions hold across a wide range of assumptions.

Price and value can rank in opposite directions. In the calibration, one of the most expensive tokens to buy returns the most value per token, while the single most expensive token in the lineup returns almost none, and the cheapest token returns the most negative value. Paying more does not buy value, and paying less can destroy it. In reality, no one model is best, and having a mix of options is what makes the best match for a particular need.

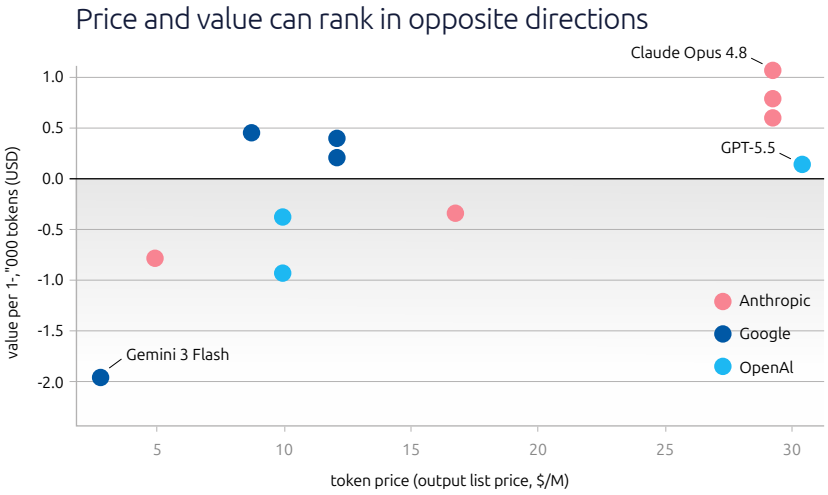


Figure 3. Across the mid-2026 lineup, token price and value per token rank in opposite directions. The cheapest token (Gemini 3 Flash) is the most value-destroying; the most expensive (GPT-5.5) returns almost nothing; a mid-priced reliable model tops the ranking. τ²-bench airline calibration.[6]

Price is almost irrelevant to value. One would expect price and value to be correlated. Our study shows that none of the variation in token value is explained by its price. Model performance and human oversight costs are the dominant factors. At realistic agentic volumes, the price of the tokens sits two to three orders of magnitude below the positive outcome of the workflow, the cost of the errors those tokens cause, and the oversight needed to catch them. This holds until an agent consumes the order of a million tokens per workflow, at which point price re-enters the calculation. That is a real but distinct regime.

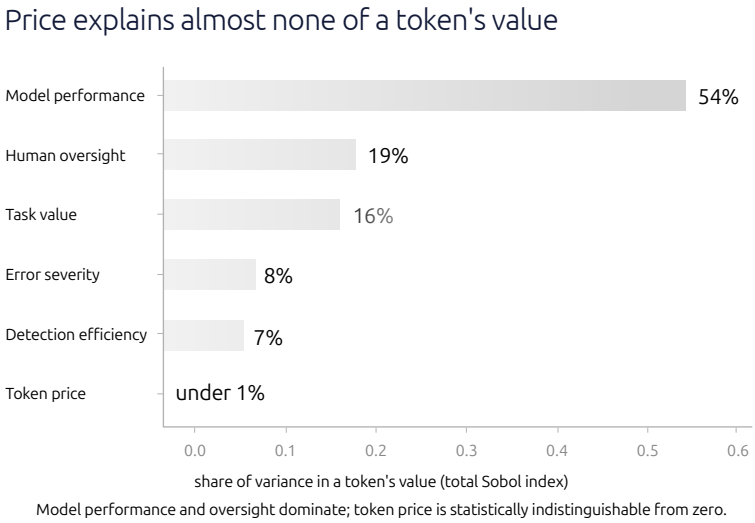


Figure 4. A global variance decomposition (total Sobol indices) attributes about 54% of a token's value to model performance and 19% to oversight, and essentially none to token price. The token bill is 2 to 3 orders of magnitude below the cost of errors and oversight. [7]

Value has a threshold and a tail. Below a certain reliability threshold (measured as the probability of executing correctly the worfklowworkflow) of roughly 0.89 on this workflow, no amount of human oversight rescues the economics. You cannot supervise your way out of an unreliable model. Even models that create value on average carry a negative tail. Typically, they are profitable and yet lose heavily in their worst 5% of workflows. One wrongful denied-boarding or one lost high-value customer costs an order of magnitude more than a routine success earns. Based on average performance, these models would appear economically sound. The risk,

however, lies in the tail, where high-stakes automation fails and outweighs any value created even by many successful decisions.

You cannot supervise your way out of an unreliable model

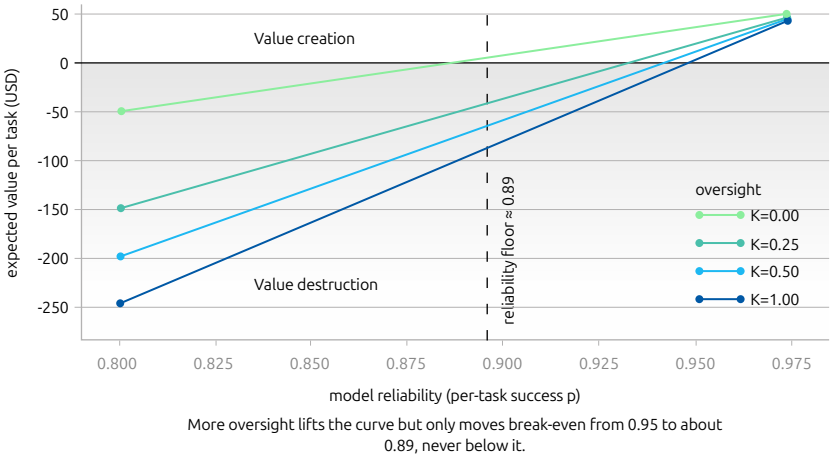


Figure 5. The reliability floor. More human oversight lifts expected value but only moves the break-even reliability from 0.95 to about 0.89; below that floor the workflow destroys value at every oversight level. [6]

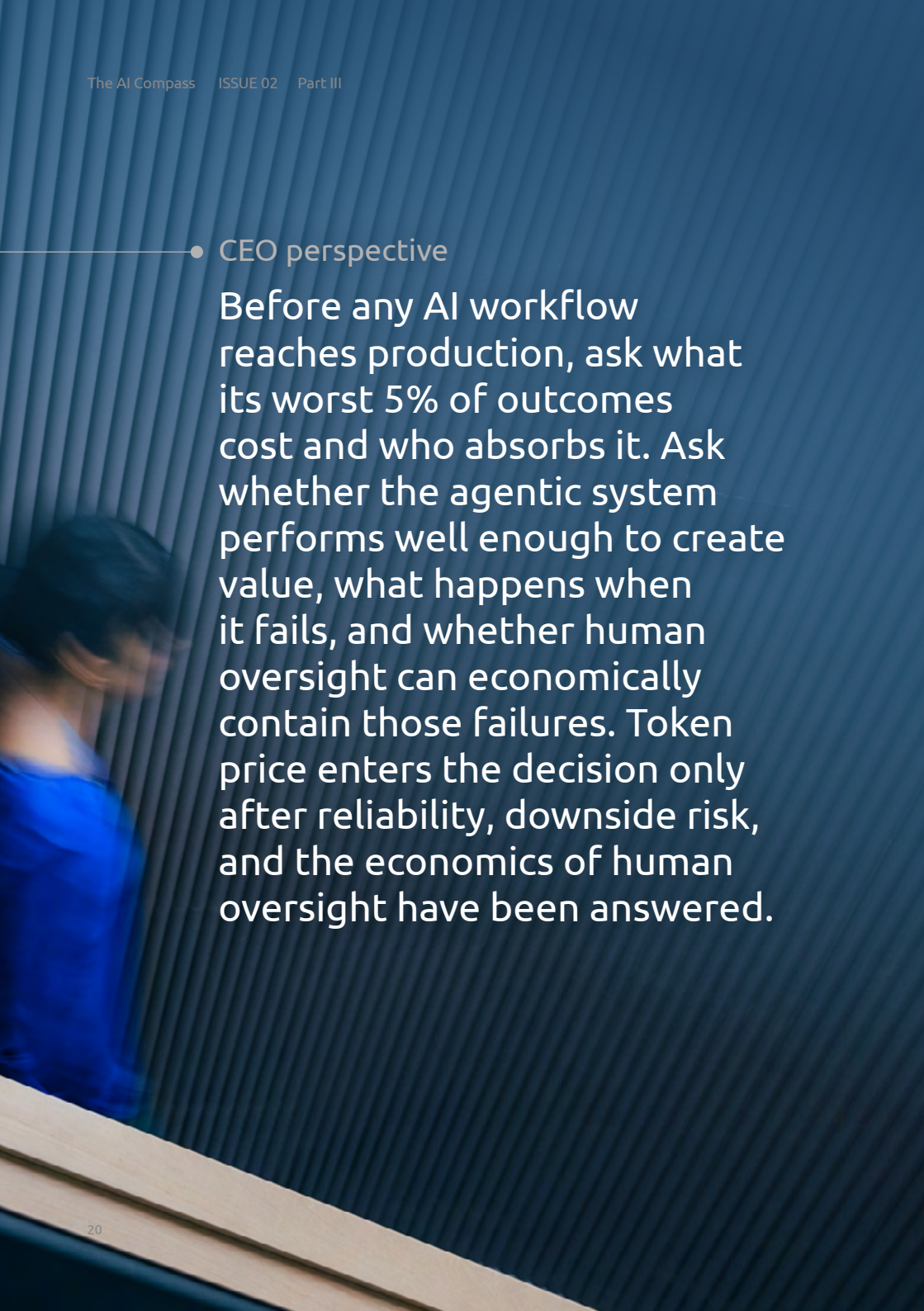
To sum it up:

The evidence leaves CEOs and CAIOs with a short list. The token bill is the smallest term in the equation. Reliability, and what happens in the worst 5% of outcomes, decide whether a workflow pays. That splits the work in two, with one decision the two roles must make together.

CEO: Own the two numbers no vendor can supply: what a correctly completed workflow is worth, and what a failed one costs. Have finance, operations, and the business sign off on both before any model is chosen, because every router and every oversight rule will optimize against whatever definition they are given. Then ask one question of every AI business case that reaches your desk: what does the worst 5% of outcomes cost us, and who absorbs it? A case that cannot answer is measuring inputs, not outputs, the same gap that leaves only 14% of CFOs able to show a measurable AI return.

CAIO: Set reliability targets based on the financial stakes of each workflow, not on the price of the tokens it consumes, and hold engineering teams accountable for meeting them. Treat the reliability floor as a gate: a workflow whose model sits below it does not reach production, however many reviewers are added. Measure reliability per workflow, where the work happens, and report the worst 5% alongside the average. Token price enters the model decision last, once those numbers are known. Raise reliability through the model mix (frontier, open-weight, and small language models) and through an architecture that supplies structured context to each workflow.

Together: Agree that the AI budget is governed by value per correctly completed workflow, with token spend treated as a constraint rather than the objective. That single rule is what turns a FinOps program into ValueOps, which Part IV describes.



CEO perspective

Before any AI workflow reaches production, ask what its worst 5% of outcomes cost and who absorbs it. Ask whether the agentic system performs well enough to create value, what happens when it fails, and whether human oversight can economically contain those failures. Token price enters the decision only after reliability, downside risk, and the economics of human oversight have been answered.

Part IV – Making it real

From FinOps to ValueOps

The deeper change is from FinOps to what we call ValueOps. FinOps made AI spend visible and drove down cost per token. It remains necessary but optimizes the one term that barely moves value. Even though ValueOps keeps the same cross-functional rigor, it points it at a different number: the value per correctly completed workflow, net of the errors, and oversight the work requires. FinOps asks what a token costs, while ValueOps asks what it produces. In practice, that means managing the two variables that move value, reliability, and stakes, and it changes four disciplines (Figure 5).

Define what a completed workflow is worth before optimizing

The value of any AI model is not a given; it is a choice, and often a contested one. Is a completed booking worth its revenue, its margin after variable cost, its contribution after fixed cost, or the lifetime value of the customer once repeat purchase is counted? In online travel marketing, for instance, teams have spent months agreeing on how to value a booking, and the answer changed which campaigns looked profitable.

How you define the value of a successful outcome influences every downstream decision, from evaluation to optimization. So, a loose or self-serving definition that makes the economics look better than they really are does not just mismeasure value, it biases routing and oversight toward the wrong workflows.

ValueOps starts by putting that definition on the table and getting business, finance, and operations to agree with it, because the first term in the equation, what success is worth, is the one number the model cannot supply for you.

Measure reliability per workflow rather than per benchmark

A single leaderboard score hides how consistently a model finishes your work across repeated attempts, and that consistency is what decides whether it clears the threshold on the workflows you care about.

Route by value, not price

The right rule sends each workflow to the model that creates the most value net of the errors it causes, which is rarely the cheapest one. A router that minimizes price is solving the wrong problem.

Raising reliability is a design choice, not only a purchasing choice. The levers are a model mix that matches frontier, open-weight and small language models to each task, an

architecture that supplies structured, governed context to the model, and routing rules that classify the workflow, apply the reliability floor, rank the remaining models by net value, and escalate to a stronger model or a human when confidence is low. Our calibration covers frontier models on one workflow, so the role of open-weight and small models is a hypothesis to test on each enterprise's own workflows. No single model rules them all: the mix makes the match.

Route by value, not price; size oversight to the tail

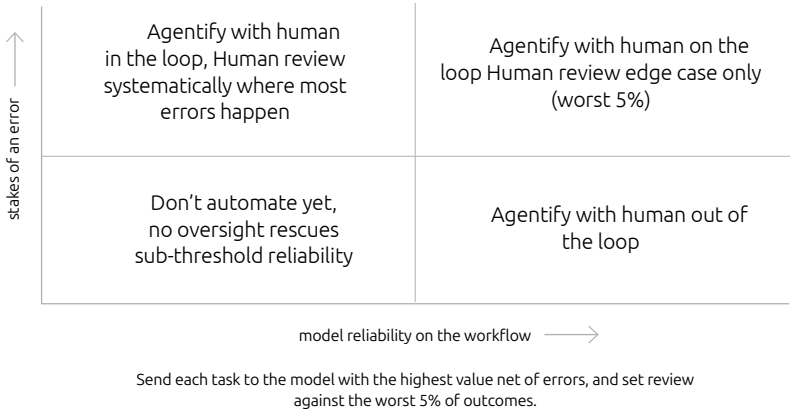


Figure 6. Route by value, not price, and size oversight to the tail. Match each workflow to a model by reliability against stakes; review high-stakes work against its worst 5% of outcomes, not its average.

Size oversight to the tail

In this case, human review works like a detection technology with a cost and a payoff, and how much review to apply is an economic question, not an engineering default. For high-stakes workflows with the potential for severe failures, the economics point to reviewing everything, while for graded, lower-stakes work, they point to reviewing only the fraction most likely to be wrong. For any workflow that can produce a rare, expensive failure, assess the decision against the worst 5% of outcomes, not the average. Average performance can make the model appear suitable while obscuring the risk of significant losses. Operationally, ValueOps lives in the agentic control plane, the central nervous system of the agentic workforce. The four disciplines converge to form a practical approach: define what a completed workflow is worth, measure real-world reliability, size oversight to the tail, and route each workflow to the highest-value model, while tracking decision value and risk. Here, providers charge for the tokens consumed, while enterprises measure whether workflows succeed, how reliably they do so, and the risk of severe failures. The agentic control plane is what makes ValueOps operable and auditable instead of existing only as a principle on a slide. It is the natural home for the calibrated asset an integrator brings: the tested method and reference data used to measure reliability and value on an enterprise's own workflows.

Global Systems Integrators (GSIs) earn their place here, because they fill a gap that model providers and hyperscalers cannot address alone. Providers sell tokens and compute at

prices and margins they control, but they do not have visibility into your workflow, your stakes, or what an error costs you, and this margin grows with token volume.

Model providers and hyperscalers have neither the visibility nor the incentive to stand behind a delivered outcome. They will underwrite the token, but they will not underwrite the workflow. In contrast, GSIs master the business and workflow context, measure reliability where the work happens, and align oversight with the level of risk. That is what lets them commit to token value rather than token cost, holding a price against a correctly completed workflow while the provider still charges per token. A reusable, calibrated asset that lets an enterprise run this model against its own workflows is what turns ValueOps from a principle into a commitment a CFO can hold a vendor to.



CEO perspective

Ensure the value of AI is a shared business decision, not a technical assumption. Require finance, business, operations, and technology leaders to find consensus on what a successful workflow is worth, then use that definition to guide model selection, oversight, and investment. Optimizing within functional perimeters caps value at each perimeter's sub-optimum; only a holistic definition of value unlocks the full amount.

Conclusion

Buy the value, not the token

The token bill is real, and it is where most CEOs and Chief AI Officers are focused today. But reducing token spend in isolation favors cheaper answers over correct ones. It optimizes only the smallest term in the value equation, while eroding the returns it was meant to protect. What matters is the value of a correctly completed workflow. That value depends on the performance and reliability of the AI system, and it is here that value is created or destroyed. It is also where the organizational chart gets in the way. When each function optimizes AI inside its own perimeter, each lands on its local optimum, and the sum of local optima falls well short of what the enterprise could capture. Value has to be defined and governed across the whole workflow, end to end, or the ceiling is set by the narrowest perimeter in the chain.

What to do next: the Capgemini value mapping quick scan

CEOs and CAIOs do not need a new operating model to start. They need five numbers per workflow, and a few weeks to get them. We run this as a quick scan on an enterprise's ten highest-stakes AI workflows. Each step maps one term of the value equation and produces one output that finance can read.

Step 1.

Map the workflows
List every AI workflow in production or in pilot and rank them by stakes: the value of a correct completion and the damage of a failure that reaches a customer, a regulator, or the books.
Owner: CAIO.
Output: a ranked list; the top ten go through the rest of the scan.

Step 2.

Price success and failure
For each of those workflows, agree with finance and operations on what a correctly completed workflow is worth and what an escaped error costs, including the rare, expensive one.
Owner: CEO sponsors, the CFO signs off.
Output: two numbers per workflow that no vendor can supply.

Step 3.

Measure reliability where the work happens
Run each candidate model repeatedly on your own workflow, not on a leaderboard, and record the share of correct completions and the cost of the worst 5% of runs.
Owner: CAIO.
Output: a reliability score per model per workflow, with the workflow's tail.

Step 4.

Set the floor and size the oversight
From the stakes in step 2, derive the reliability threshold a model must clear for the workflow to create value; any model below that floor is out, whatever its price. For the rest, decide whether to review every output or only the fraction most likely to be wrong.
Owner: CAIO, with operations.
Output: a go/no-go per workflow and an oversight rule.

Step 5.

Route by value and report it
Send each workflow to the model with the highest net value, not the lowest token price, and report value per correctly completed workflow next to the token bill every month.
Owner: CAIO, reported to the CEO and CFO.
Output: a ValueOps view in the agentic control plane that a CFO can hold a vendor to.

Start with the ten workflows where a failure costs the most. If a workflow cannot get through step 2 because nobody can write down what a correct completion is worth or what a failure costs, it is not ready for an agent, whatever the token price.



Sources & notes

[1] Chen et al. https://arxiv.org/abs/2412.21187

[2] Stanford HAI, AI Index 2025 (inference-cost decline for GPT-3.5-level output).

[3] Exponential View, The State of the AI Economy, June 2026.
Token-price elasticity across providers (≈1.2–1.8).

[4] CRI Research Snapshot: The cost of intelligence, 2026.

[5] IDC Research, November 2025.

[6] Exponential View, The State of the AI Economy, June 2026.

[7] Salesforce commentary on "agentic work units" as an output-based metric, 2026.

[8] P. Cordier, "The Value of a Token: Token cost doesn't matter;
Model reliability, human oversight, and tail risk are", 2026.

Authors



Etienne Grass

Global Chief AI Officer
Capgemini Invent



Philippe Cordier

Global Chief AI Scientist
Capgemini Invent

Note:

This POV was created using AI for research and data gathering; all citations and sources have been independently verified by humans. The content has been peer-reviewed and signed off by humans.

About Capgemini

Capgemini is the business transformation partner for enterprises in the age of AI. We help organizations imagine and build an intelligent, sustainable future, combining AI, technology and human ingenuity to transform how they operate, innovate and grow. With unique end-to-end capabilities spanning strategy, technology, engineering and intelligent operations, we bring together deep industry expertise and market-leading capabilities in AI, cloud and data to turn ambition into measurable business outcomes at scale. Supported by a robust ecosystem of partners and nearly 60 years of expertise, Capgemini is a responsible and diverse global organization of over 410,000 team members in more than 50 countries. The Group reported 2025 revenues of €22.5 billion.

Make it real. www.capgemini.com

Capgemini