48.8744° N
2.2965° E

# The AI Compass

ISSUE 01

The era of AI Yield.
Mastering tokenomics
to maximize enterprise value.

Capgemini

# Table of contents

Executive Summary — 08

Part I — Why AI economics has changed: Eight convictions — 10

Conviction 01:
The end of the free lunch — 10

Conviction 02:
A token price aggregates far more than it
appears to – architecture the most — 13

Conviction 03:
Very few can see clearly yet,
so build for flexibility — 16

Conviction 04:
Seven engineering levers that can more
than halve your token costs — 18

Conviction 05:
UnTokenized Intelligence is now
here and is a good bet — 22

Conviction 06:
As the token market spans luxury
goods to commodities, routers should be part of your stack — 26

Conviction 07:
Our primer study shows that value
is not connected to costs — 28

Conviction 08:
In the end, it is all about finding the right yield — 30

Part II — What AI Yield means for the CEO — 34

Consider the scenario in which token costs reach
10% of cost base — 34

Protecting margins requires E2E transformation — 35

Monitor yield with the AI Economics Stress Test — 38

Aligning your C-suite is of strategic importance — 40

Part III — What AI Yield means for the CAIO, CIO and CTO — 44

The best strategy today is a dual mode — 44

Build your own portfolio of small language models (SLMs) — 47

Conclusion — 52

# Executive summary

Ultimately, there is one metric for the profitability of enterprise AI:

**AI Yield = Verified Business Value — Full Cost of Safe Delivery**

This is the verifiable value of AI-assisted enterprise outcomes, minus the cost of achieving it safely, including the human oversight required to catch errors. Cost and adoption are inputs to this yield, not substitutes for it.

This paper, exploring what contributes to AI Yield, is organized in four sections:

Firstly, we set out our eight convictions for "Tokenomics", the name given to the specific economics of enterprise AI.[1] We then describe the Capgemini AI Economics Framework that we use to model it. We articulate what this means for a CEO, who is expected to present an AI deployment strategy to a management board or executive committee. We conclude with what Tokenomics means for technology leaders, that is, CAIOs, CIOs or CTOs – the owners of the AI mechanism that should deliver returns.

Three significant factors changed in enterprise AI in early 2026, and their effects have compounded. Firstly, the effective subsidy that funded the first three years of enterprise AI is being withdrawn. Secondly, the cost of a token depends far less on the price list than on how the workflow around it is engineered. And finally, the value a token creates is not correlated with what it costs.

[1] Federal Reserve Bank of Atlanta, Policy Hub Macroblog, "How Much Are Firms Spending on AI (and What Will Happen to Headcounts)?", 6 May 2026 — Survey of Business Uncertainty, fielded March 2026, N = 855

AI has become one of the fastest-growing and least-governed enterprise inputs. Most leadership teams still cannot show, use case by use case, what that spend is buying. Scaling token consumption without accurate measurement of related outcomes creates volume but delivers productivity in name only.

> The winners of the AI era will not be the enterprises that consume the most or least tokens. They will be the ones that generate the highest AI Yield.

Capgemini Invent's analysis suggests that token cost is not, by itself, a reliable indicator of business value. In the firm's work on token economics, the value derived from AI systems is presented as being influenced more heavily by factors such as model performance, workflow design, and human oversight than by token price alone. This implies that organizations should be cautious about managing AI purely through token consumption or spend metrics and should instead evaluate AI investments against measurable business outcomes.[2,3]

A Capgemini Research Institute study sought to quantify cost reductions that could be achieved by running open-weight models on enterprise-owned infrastructure. It identified cost reductions approaching 90% for qualifying workloads and introduced our concept of “UnTokenized Intelligence”, illustrating why it is a bet worth placing.[4]

Those findings land at a moment when open-weight models are no longer a fringe option. Models in the 25 to 35 billion parameter class now execute well-scoped enterprise tasks with acceptable quality, are licensed permissively from providers, and whether hosted or self-hosted, cost one to two orders of magnitude less per token than frontier reasoning closed-weight model tiers.

The practical consequence of this progress for most enterprises is a **dual mode** business model: frontier closed models for the minority of genuinely hard work, open-weight models for the routine majority, with a routing layer that assigns each task based on measured evidence rather than practice. This is no longer experimental. For operating at scale, it is the default architecture.

[2] Stanford Institute for Human-Centered AI, AI Index Report 2026; Stanford Digital Economy Lab, "How Do AI Agents Spend Your Money?" (2026)

[3, 4] Capgemini Research Institute, “The multi-year AI advantage: Building the enterprise of tomorrow” (2026)

# Tokenomics and the five value pools of the Capgemini Group

Token cost affects the five value pools defined by the Capgemini Group at its Capital Markets Day held on 27 May 2026.

Modernization of technology debt:

1. Becomes a cost-per-token argument: a weak data and API estate makes every agentic transaction structurally token-expensive, and that penalty recurs on each run rather than once at project time.
2. The new agentic technology stack is where the unit economics are set - model routing, context caching, retrieval design.
3. The agentic control plane is where tokens stop being an inference bill and become a metered, attributed, budgeted P&L line with a named owner.
4. Agentification of products and services raises a margin question: an offer priced per seat with a cost base scaling per token carries unhedged exposure to consumption growth, and the reasoning-heavy models that make it differentiating are the ones that expand volume fastest.
5. Agentification of enterprise operations restates the same arithmetic internally - the case becomes tokens per outcome, not FTE saved.

# Part I — Eight convictions

Conviction 01

## The end of the free lunch

American saloons invented the technique of offering a free, heavily salted lunch to sell more drink. "There ain't no such thing as a free lunch" is the proposition Milton Friedman used to underline the role of the price signal in markets. Data science has its own no-free-lunch theorem: any performance gain on one class of problem is paid for by an equivalent loss on another.

In November 2022, ChatGPT took us back to the saloon. OpenAI built a base of roughly 900 million weekly users largely on free or below-cost access and posted an operating loss of close to $21 billion against $13 billion of revenue in 2025. This equates to roughly $2.60 spent for every $1 earned, funded by capital markets rather than customers.

Enterprise agentic AI ends the arrangement. Across the market, flat-rate plans have given way to usage-based billing, weekly caps, peak-hour limits and quota cuts. This is a continuous sequence of restrictions from early 2025 through 2026 most notably across OpenAI, Anthropic and Google.

The behavior on the other side has a name: "tokenmaxxing" – treating raw consumption as a proxy for productivity. Uber exhausted its entire 2026 AI budget in four months as Claude Code and Cursor usage spread across roughly

5,000 engineers faster than finance had modelled, compounded by an internal leaderboard ranking teams on AI usage. Meta's employees consumed 60 trillion tokens in a single month, some $9 billion at public list prices.

The macroeconomic signal now comes from a central bank rather than a vendor. The Federal Reserve Bank of Atlanta's Survey of Business Uncertainty, fielded March 2026 across 855 firms, has found that US firms spent $1,358 per employee on AI in 2025 and expected to spend $2,068 in 2026.[5] This represents a 50% rise in a single year, employment-weighted. Aggregated across private non-farm payrolls that is roughly $280 billion in 2026, strikingly close to Stanford HAI's independent $285 billion estimate. At 2025 levels this already represented some 2.5% of the US private wage bill and approximately 1% of GDP.

Two features of the distribution matter more than the average. It is extraordinarily skewed: more than half of respondents expected to spend no more than $200 per employee in 2026, while the top decile planned at least $2,800, a fourteenfold gap between leading adopters and the median firm. Further, the sector spread is widening: professional and business services expect $3,470 per employee in 2026, up 74%; manufacturing expects roughly $900.

Token spend has reached the scale at which it behaves like an industrial input cost. On current trajectories, it is today an issue of the same order as gas prices. Within three years it may be of the same order as oil.

[5] Federal Reserve Bank of Atlanta, Policy Hub Macroblog, "How Much Are Firms Spending on AI (and What Will Happen to Headcounts)?", 6 May 2026 — Survey of Business Uncertainty, fielded March 2026, N = 855

CEO perspective

AI and related costs will become a material part of the enterprise cost base. Rather than a problem to be delegated to a technology division, it is a business-wide issue requiring collaboration and synchronized execution. CIOs, CAIOs and CTOs will be responsible for managing AI, but final accountability stays with the executive committee.

Conviction 02

# A token price aggregates far more than it appears to – architecture the most

The instinct is to treat a token as a purchased commodity with a published price, and therefore as a procurement problem. That instinct misdirects the cost program.

A token price does indeed aggregate the physics of production. Arthur Mensch's description of AI as the discipline that "turns electrons into tokens" grounds it in the constraints of the physical world: energy cost, hardware efficiency, amortization of training capital, provider margin. Those inputs have improved fast. Tokens produced per gigawatt have risen roughly fourfold in eighteen months, driven by generational hardware efficiency (NVIDIA's GPUs progressing from H100 to Blackwell to Rubin, for example) and by GPU farm utilization, i.e., GPUs per server rack, passing 50%. Global token production has multiplied by 28 in two years.[6]

But the physical layer is no longer where the variance sits. Since generative models arrived there are four pillars, as AI research firm, Epoch AI, has documented:

- Latency, tied to parallelization
- Data, as the reservoir shifts from the open internet to enterprise data
- Compute, tied to semiconductor supply and data-center construction
- Electricity, tied to generation and grid

[6] Stanford Institute for Human-Centered AI, AI Index Report 2026; Stanford Digital Economy Lab, "How Do AI Agents Spend Your Money?" (2026)

All four are no longer valid. The right mental model is a single end-to-end production chain in which any link can become the bottleneck, the current high-bandwidth memory shortage being the most current illustration.[7]

What determines an enterprise's actual cost per unit of work sits downstream of all of this: namely, the agentic architecture, the harness or operational layer, and context management. A conversational query runs on the order of a thousand tokens. An agentic task, a thousand times that. A continuously running agent, a thousandfold again.

One orchestrated interaction that cost four cents in 2023 costs over a dollar today. A Stanford Digital Economy Lab study found agentic coding tasks consume up to 1,000 times more tokens than a chat exchange, driven almost entirely by the model re-reading context it has already seen.[8]

This is why the bill is not governable at the price list. Cost depends on multiple factors: prompt length, retrieved context, output length, model choice, reasoning effort, tool calls, caching behavior and agent loop count. None behaves like a seat count; each is set anew by every task rather than fixed by procurement. The token is an engineering outcome before it is a purchase item.

[7] Artificial Analysis, State of AI Q1 2025 Highlights Report and 2026 intelligence-versus-cost index

[8] FinOps Foundation, State of FinOps 2026 (1,192 practitioners, >$83bn annual cloud spend); Harness, 2026 State of AI in FinOps (conducted with 700 engineering leaders in five countries)

CEO perspective

AI cost is not a procurement issue, but an enterprise financial control issue. Falling token prices are irrelevant if poor architecture allows consumption to grow faster than unit costs decline. The mandate is clear for CIOs, CAIOs and CTOs: maximize measurable business value per token.

Conviction 03

# Very few can see clearly yet, so build for flexibility

The market is operating without the price transparency that economics normally assumes.

The token usage bill is like fog: opaque and shifting. Providers rarely expose a comparable breakdown of input against output tokens, and the split matters because the two are priced very differently and consumed in very different ratios by workflow. Model versions, context windows and caching behavior can change the unit economics of the same use case month to month, with no line-item explaining why.

The variance is not only unexplained, but also intrinsic. Research from MIT and Stanford shows that for an identical task on an identical model, the inherent randomness of generative models produces **cost variation by a factor of 30**. Such volatility makes any unit-economics claim built on a single observation meaningless.[9]

Governance data confirms organizations know this but have not solved it. The FinOps Foundation's State of FinOps 2026 (conducted with 1,192 practitioners with more than $83 billion in annual cloud spend) finds 98% now actively manage AI costs, up from 31% two years earlier. Harness's 2026 State of AI in FinOps (700 engineering leaders, five countries) puts numbers on the surprise: **72%** hit an unexpected AI bill in the past year and a third were caught more than once; organizations estimate **26% of AI spend is wasted**; **52%** say there is no clear owner of AI cost; and if spend doubled overnight, only **20%** could identify the reason within hours.

[9] Exponential View, The State of the AI Economy, June 2026; BCG, AI Radar 2026 (2,360 executives, 16 markets, 640 CEOs)
FinOps Foundation State of Finops 2026 Report, https://data.finops.org/
Harness, 2026 State of AI in FinOps, https://www.harness.io/state-of-ai-in-finops-2026

**Three structural features keep the fog in place:**

1. Consumption is decentralized and behavior-driven: every employee can now consume tokens with limited guardrails.
2. Volumes are extremely volatile: inference costs can spike 10 to 100 times in a single day.
3. The catalogue is dispersing: the gap between cheapest and most expensive model now exceeds 200 times.

CEO perspective

Gartner expects more than 40% of agentic AI projects to be cancelled by end-2027, citing escalating cost and unclear business value in the same breath. They are the same problem, measured from the wrong end. We call this the **AI Accountability Gap**. It is an opportunity as much as a risk: organizations that close it first will scale AI with confidence while their peers are still guessing.

Conviction 04

# Seven engineering levers that can more than halve your token costs

The organizations creating a sustainable AI advantage are not winning through model choice alone. They are winning through engineering discipline that converts lower cost into higher AI yield. This is the most directly actionable material in the paper, and the reason we treat token containment as engineering rather than procurement. Seven levers now known to have a documented effect, ordered below by maturity. The savings figures are documented upper bounds and are not additive.

1 **Reuse previous work:**
Caching can reduce API calls by up to **68.8%**, particularly where users submit similar or repeated requests.

2 **Use the lowest-cost model that can complete the task:**
Intelligent routing can reduce costs substantially, with typical savings of **37% to 46%** and reported upper bounds of up to 98%. Decisions should be based on actual operating costs, not published prices.

3 **Reserve advanced reasoning for tasks that need it:**
Selective use of deeper reasoning reduced token consumption by **48.5%** while improving accuracy by **10.2%**. Applying it to every task adds cost and can reduce quality.

4 **Control how much information the AI processes:** Removing duplication, summarizing prior interactions and limiting unnecessary context can reduce consumption **by two to five times** at comparable quality.

5 **Set limits on external actions:** Quotas on tool calls prevent AI agents from pursuing long and expensive chains of activity, without measured loss of accuracy.

6 **Select models using real business workloads:** Tests found a difference of **1.5 million tokens per task** between models completing work of equivalent difficulty. Model choice should follow a controlled pilot using 50 to 200 representative tasks.

7 **Make systems easier for AI to understand:** Clear code, structured data and good documentation reduce ongoing AI consumption. This has the greatest long-term potential, but benefits are realized over **12 to 24 months**.

Two conclusions follow. The levers stack but the discipline does not: the 60–80% efficiency gap between mature and naïve FinOps organizations compounds across the above seven levers and is reset every time a new agentic workflow ships without observability. Further, sequencing matters as levers one to five are operational and come first; levers six and seven are structural.

Applied together the levers are not marginal. We can document **40-60% reductions** in token spend at broadly equivalent output quality. One telecoms operator cut monthly API costs by 42% through an LLM gateway with dynamic routing; another cut cloud token spend by 60% and improved response speed by 40% by tiering edge against cloud models; a retailer cut costs by 41% through a lightweight proxy, routing and caching while holding quality within 98% of baseline.

A reference case is AT&T, which cut AI run-cost by **90%** while throughput rose from 8 to 27 billion tokens per day. At 8 billion tokens a day, routing everything through frontier models was neither feasible nor economical. Orchestration was rebuilt so large "super-agents" direct small, domain-tuned worker models. As AT&T's Chief Data and AI Officer was quoted in media, "The future of agentic AI is many, many, many small language models."[10]

[10] Fortune, "After blowing through AI budget in a matter of months, Uber CTO says tokenmaxxing era is over" (August 2026); SemiAnalysis, "TokenBudgeting" (June 2026)

## CEO perspective

AI cost efficiency must be designed into every deployment from day one. Every team should be mandated to optimize token consumption through intelligent routing, caching, context management and disciplined use of advanced reasoning before any solution reaches production. Every AI workload must have cost, quality and business value measured together, making token economics a core governance metric rather than a technical afterthought. Frontier models should be reserved for high-value tasks, with lower-cost models handling routine work wherever possible. The goal is not to minimize AI usage but to maximize business value per token, creating a scalable, profitable and sustainable AI operating model.

Conviction 05

# UnTokenized Intelligence is now here and is a good bet

The levers above optimize consumption on metered infrastructure. A second, more structural move changes the meter itself.

## What the concept means

The underlying transformer still tokenizes text. That mechanism does not change. "UnTokenized" describes what changes commercially: For workloads moved onto owned or dedicated infrastructure, the link between per-token metering and cost is severed. This is a billing and infrastructure shift executed with efficient inference engineering. It is not a change to model internals, and any claim that it is should be treated skeptically.

## Why it is emerging now

Four conditions have converged. Open-weight models in the 25–35 billion parameter class have become genuinely capable on well-scoped enterprise tasks and are sparse mixture-of-experts architectures activating only 3–4 billion parameters per token. Serving software has matured to where the same hardware delivers roughly six times the throughput it did under naïve serving. Regulatory and contractual pressure on data residency has hardened. In addition, agentic workloads have made the metered bill volatile enough that fixed-cost capacity carries value independent of unit price.

## The architecture rests on four pillars

1. Local small models, task-specific: compact, fine-tuned models for regulatory audit, confidential extraction, legal and financial reporting. These are not an oversized generalist.
2. Decoupled billing and compute: once deployed, the marginal cost of inference falls towards infrastructure and maintenance rather than tracking volume.
3. Hybrid by design: routine, high-volume work runs locally, rare complex reasoning routes to frontier APIs through a gateway classifying on intent, context, complexity and semantics rather than cost alone.
4. Evaluated, not assumed: no workload moves locally until it passes an explicit accuracy gate. Architecture without an evaluation gate is a cost decision disguised as an engineering one.

## What our study found

Our modelling works a concrete deployment shape rather than quoting a percentage at an unstated volume. Let's take an example of 1,000 active users across 50 agentic workflows, each running four to five collaborating agents.

This scenario generates roughly 63.8 billion tokens a month, of which a 50-50 routing split places 31.9 billion locally. That translates to 1.669 trillion tokens over three years at 40% annual growth. This increase is stark and should not be ignored.

As a result, we found that three key findings deserve care, because two contradicted the headline finding.

## The cost reduction is real and large

Running this workload on your own infrastructure costs 88 to 94% less than buying the same volume from frontier model providers. The hardware involved is modest – five to nine servers, bought progressively so that as the volume materializes, the economics work.

## How you run the hardware matters more than what you buy

The same servers can be operated well or badly, and the gap is not marginal. Operated well, nine machines carry the workload for a cost equivalence of $594,000 over three years. Operated without the software techniques that keep them busy, you need fifty-three machines and $3.43 million for the same output. Every euro of hardware is worth roughly six times more in the first case than in the second. That gap is a skills question before it is a procurement one.

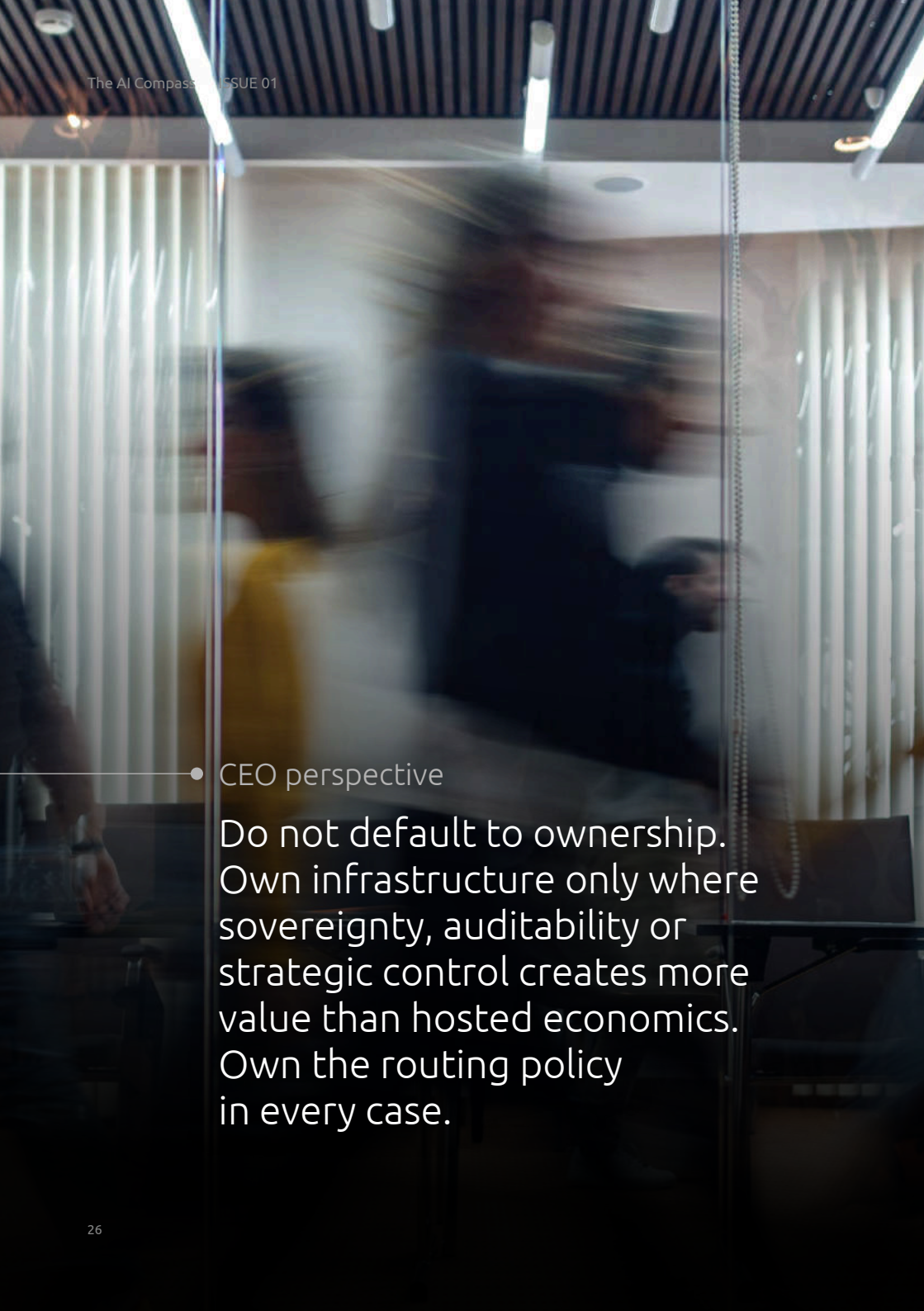
## Owning GPUs is not the only option

The same open-weight models can be rented from specialist providers who host them for you. On price alone, renting beats owning by roughly half again against our optimized cluster, and by three to five times once the full cost of running your own machines is counted. Owning is justified by control: where your data sits, whether a third party can reprice you, whether you may fine-tune on your own archives - not by unit price. Where none of those constraints applies to a workload, rent it.

## What does owning buy?

- Data residency, where localization is a precondition rather than a preference.
- Price certainty and capitalization opportunity. Blended rates moved from $18.40 to $6.07 per million in twelve months, however the market reprices quickly both upwards and downwards.
- Unrestricted fine-tuning on archived enterprise data that cannot be sent to a third party under any commercial terms.
- Independence from a provider's rate limits, capacity allocation and deprecation schedule: a hosted endpoint can be withdrawn, a checkpoint on your own storage cannot.
- Audit and evidence, where inference logs are themselves records you must retain and produce.

Finally, it buys long context without a meter, which changes the design space for document-heavy workloads.



CEO perspective

Do not default to ownership. Own infrastructure only where sovereignty, auditability or strategic control creates more value than hosted economics. Own the routing policy in every case.

Conviction 06

# As the token market spans luxury goods to commodities, routers should be part of your stack

Speaking of "the price of tokens", there is a widely held assumption that a token is a single homogeneous good. That assumption is no longer correct. Segmentation now matters more than any single headline price.

At one end are frontier reasoning models priced as luxury goods, at a platform blended average of roughly $6.07 per million and considerably more for the newest reasoning tiers.

At the other, hosted open-weight 30B-class models at roughly $0.26, represents a spread of more than twenty times between tiers and more than two hundred times across the catalogue. Between them, budget frontier models at around $2.94.

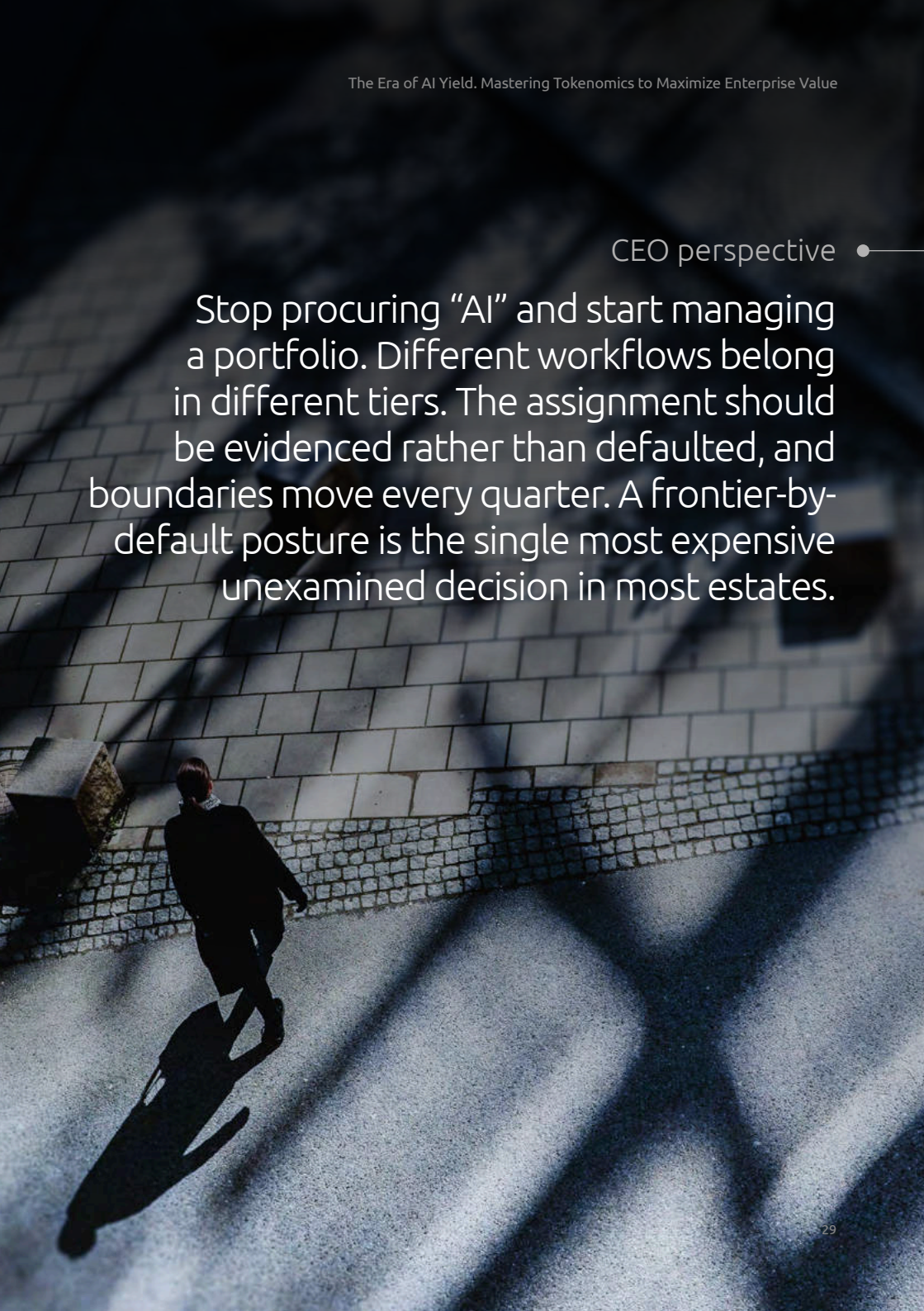
Two forces are pulling the tiers apart rather than together. The first is capability efficiency: the cost of intelligence at GPT-4's original capability level has fallen more than a thousandfold since March 2023, and high-intelligence inference is roughly 32 times cheaper than in September 2024 alone.

Effective blended enterprise rates fell 67% year on year, from $18.40 to $6.07 per million between Q1 2025 and Q1 2026. Capability that was frontier eighteen months ago is now commodity, and the migration is continuous.

The second is divergence in vendor business models, which determines whose cost curve you inherit. Anthropic earns roughly $211 per month per paying user against OpenAI's $42. This is a fivefold gap explained almost entirely by revenue architecture, with 70–75% of Anthropic revenue consumption-based against a subscription-majority model at OpenAI.

The expansion coefficient tells the same story: Anthropic customers spend 3.75 times more over time while OpenAI revenue tracks its user count. Providers whose revenue is metered have a structural incentive to grow tokens per task; providers whose revenue is per seat have the opposite. Neither is aligned with yield.

**The Inference Cost Paradox** makes this counterintuitive. Falling unit prices should be good news for a cost line. However, they are not, because volume grows faster than price falls, essentially the Jevons paradox in an enterprise budget. Nvidia's CFO has described inference as already accounting for around 40% of data-center revenue, up from a training-dominated base a year earlier. Cheaper tokens do not produce a cheaper AI estate. They produce a bigger and less predictable one.



CEO perspective

Stop procuring “AI” and start managing a portfolio. Different workflows belong in different tiers. The assignment should be evidenced rather than defaulted, and boundaries move every quarter. A frontier-by-default posture is the single most expensive unexamined decision in most estates.

Conviction 07

# Our primer study shows cheap AI is often expensive

Every leadership team is having a version of the same conversation: AI is getting cheaper by the token, so how do we cut our token bill? That is the wrong conversation. The price of a token is observable, falling fast, and converging across vendors within a tier. Its value is none of those things. A token is a billing metric, not a unit of value, and confusing the two is a costly category error.

In his research, Capgemini Invent's Philippe Cordier has formalized the value of a token inside a task-based model of production, calibrated on a public agentic benchmark.[11]

## Four findings matter for capital allocation

1 **Price and value rank in opposite directions.** One of the most expensive tokens returns the most value per token; the single most expensive returns almost none; the cheapest returns the most negative value. Paying more does not buy value. Paying less can destroy it.

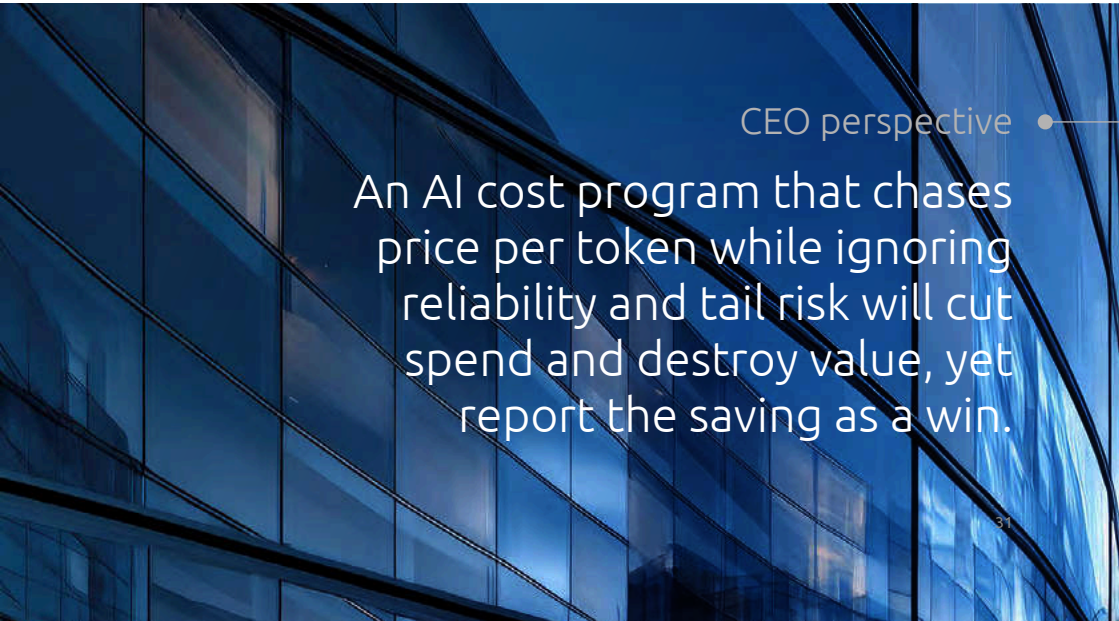
2 **Price explains essentially none of the variance in value.** A global variance decomposition attributes about **54% to model performance** and **19% to oversight intensity**, and approximately zero to token price. At a representative operating point, escaped-error cost is roughly **600 times**

[11] Colette Kress, CFO, NVIDIA, on inference share of data-centre revenue, cited in K. Bheemaiah, Capgemini Invent, "Inference Cost & Speed"

the token cost. Price re-enters only when an agent consumes on the order of a million tokens per task.

3 **Value has a reliability threshold.** Below a model reliability of roughly **0.89** on this workflow, no amount of oversight rescues the economics. More oversight lifts expected value but only moves break-even reliability from 0.95 to 0.89. Organizations must resist the temptation to supervise their way out of an unreliable model.

4 **Value has a tail.** Models that create value on average still carry a negative tail. Like any value stream, they are typically profitable, but lose heavily in their **worst 5%** of tasks, because one wrongful outcome often costs an order of magnitude more than a routine success earns. A decision rule based on the average waves these models through.

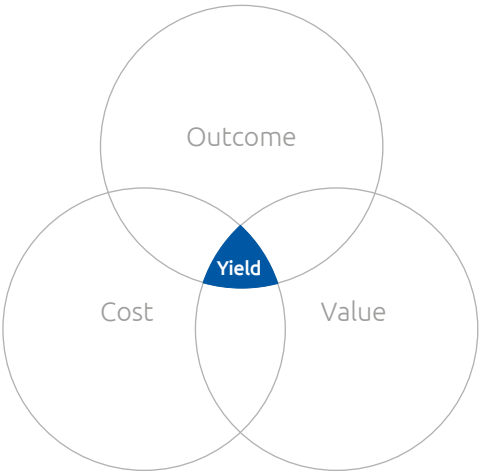


CEO perspective

An AI cost program that chases price per token while ignoring reliability and tail risk will cut spend and destroy value, yet report the saving as a win.

Conviction 08

# In the end, it is all about finding the right yield




At this point, the seven preceding points converge. Cost is knowable but not governable by the price list. Value is decisive but not always directly correlated with cost. Neither measure alone, tells you anything actionable. What matters is the ratio, at the level of the individual use case.

Extending traditional FinOps hits a hard limit here, because AI changes the unit being managed. A token bill depends on prompt length, retrieved context, output length, model choice, reasoning effort, tool calls, caching and agent loop count. None of which behaves like a seat count, and each of which is set anew by every task.

It resolves into three questions, in order.

## What did it cost?

The fully loaded Cost Per Verified Outcome: Key measures such as model usage, infrastructure, SaaS consumption, integration effort, human oversight and governance overhead, must be included. Not just the model bill.

## What outcome was produced?

A verified business result, not an activity metric, measured against the counterfactual of not using AI at all. This is the Effective FTE Run Cost.

## What value was created?

Revenue growth, cost reduction, risk reduction, productivity gain or customer impact, evidenced rather than assumed. It is also important to remember that value can be measured both quantitatively and qualitatively.

Two things follow that are easy to miss. Yield can be improved from either side, and the engineering side is currently the faster: the levers in Conviction 4 (Engineering Levers) and the architecture in Conviction 5 (UnTokenized Intelligence) move the denominator by 40 to 90% on qualifying workloads, in months. AI Yield is a distribution, not an average. A use case with a positive mean and a destructive

downside tail is not a success. It is an unpriced liability.

There is also one key consideration that should belong in any five-year model. The wider AI investment cycle is itself unproven: global generative AI revenue outside China is roughly 0.42% of US GDP against 9.4% for the IT sector, and cumulative hyperscaler and neocloud revenue has nearly but not fully covered cumulative infrastructure depreciation.

We foresee two consequences. As providers recover that depreciation, we expect upward rather than downward pressure on enterprise AI pricing. Late-stage capitalism has shown that a market recovering its capital base does not subsidize consumption indefinitely.

Further, the shift from subsidy to taxation is now real: Finland ended preferential electricity tax rates for data centers on 1 July 2026, a roughly 45-fold increase on the previous industrial rate. As AI displaces taxable labor income, we expect more governments to seek a replacement revenue base in the infrastructure that caused the displacement.

CEO perspective

AI Yield is the verified value an AI-assisted outcome creates, minus the full cost of achieving it safely, including the human oversight required to catch and correct errors before they escape. The metric worth tracking is not cost in isolation but yield per verified outcome, use case by use case.

# Part II — What AI Yield means for the CEO

For the CEO, AI yield is a capital allocation question, not a technology question, with none of the capital allocation controls, GAAP expenses every token, so the spend that will decide your margin never crosses a board or an approval gate.

## Consider the scenario in which token costs reach 10% of cost base

The following is a scenario, not a forecast. Its purpose is to establish the order of magnitude at which AI spend stops being a technology budget and becomes a margin question, and how much time each sector has before it gets there.

## Where the trajectory starts

The starting point is observed rather than modelled. The Atlanta Fed series allocates AI spend per employee of $1,358 in 2025, rising to $2,068 in 2026, employment-weighted.

At an implied US private wage bill of roughly $54,000 per employee, that is approximately 2.5% of payroll in 2025 rising to 3.7% in 2026. The sector detail is where divergence appears professional and business services at $3,470 per employee in 2026, up 74%, against manufacturing at roughly $900.

Two forward anchors bracket the projection. Gartner puts worldwide AI spending at $2.59 trillion in 2026, up 47% year on year, with enterprise AI budgets growing at roughly 44% a year. IDC projects AI reaching a quarter of enterprise IT spend within three years. We use 44% as the central case, with slow and fast scenarios at 20% and 70%, against payroll growth of 3%.

**The critical variable is not the growth rate.** It is how payroll-intensive the cost base is, because AI spend scales with headcount while the cost base does not.

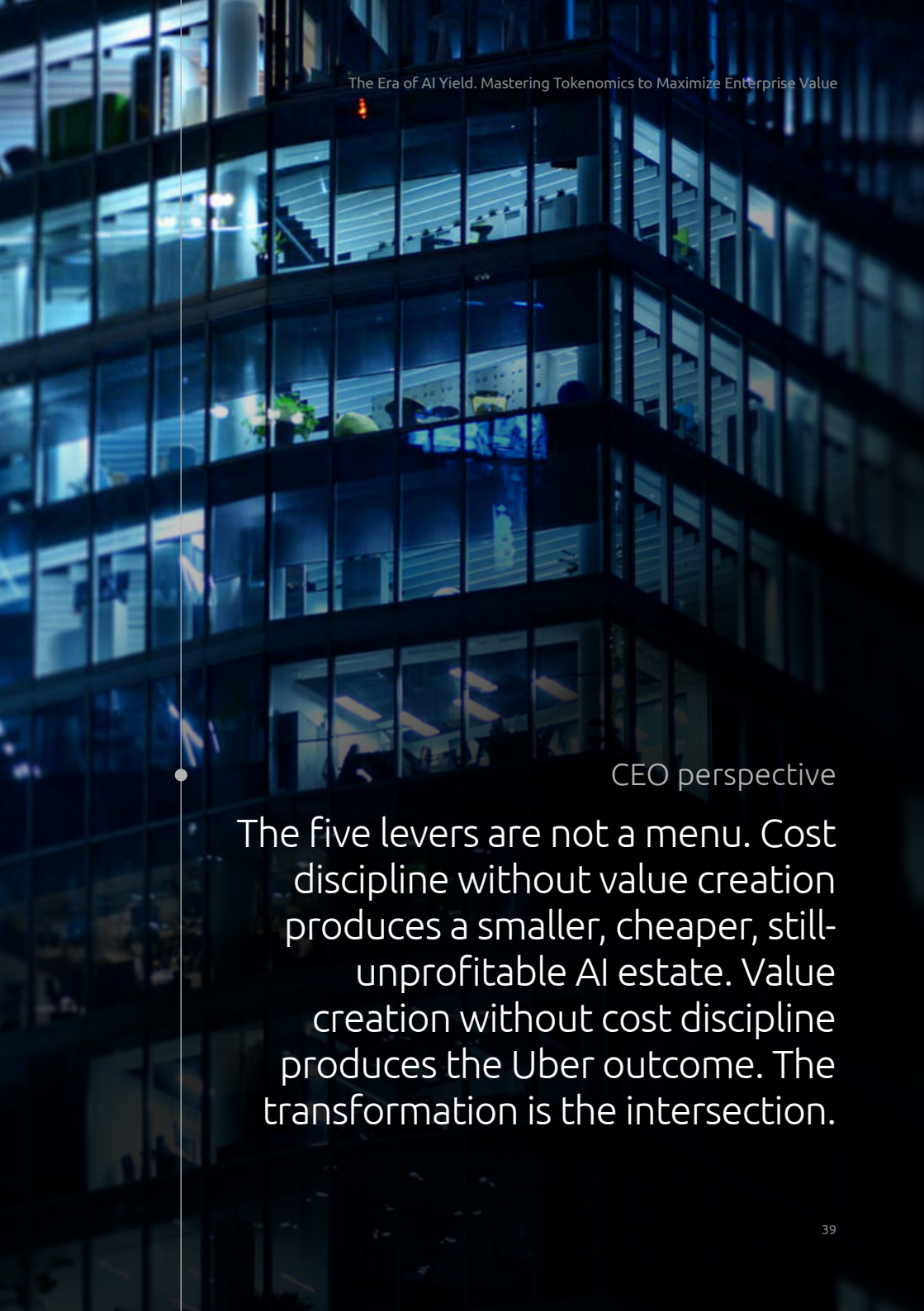
# Protecting margins requires E2E transformation

The common failure mode is treating rising AI cost first as a procurement problem, then a technology or FinOps problem, and only much later as a transformation problem. That sequence loses two years.

AI cost cannot be absorbed by cost discipline alone, for an arithmetic reason: at the growth rates shown in the scenario above, the increment outruns anything a sourcing program can feasibly recover.

Absorbing it requires movement on both sides of the ledger. This demands three value-creation levers and two cost-discipline levers.

- **Revenue innovation:** New AI-powered products, services, partnerships, channels and pricing models. The only lever that grows the numerator rather than shrinking the denominator.
- **Workforce redesign:** Human-AI collaboration, hybrid team design, targeted skill investment.
- **Operating model:** Reimagined workflows, the choice between shared and federated AI platforms, model gateways.
- **FinOps for AI:** Token budgets, model routing, consumption monitoring, cost-per-outcome tracking, and the cultural change that makes them stick.
- **Governance and risk:** Model risk, regulatory compliance including the EU AI Act, ethical guardrails.



CEO perspective

The five levers are not a menu. Cost discipline without value creation produces a smaller, cheaper, still-unprofitable AI estate. Value creation without cost discipline produces the Uber outcome. The transformation is the intersection.

# Monitor yield with the AI Economics Stress Test

AI Yield is the metric. Capgemini's **AI Economics Stress Test** applies it, use case by use case. It is a structured diagnostic scoring individual use cases across five dimensions.

- **Model economics:** The fully loaded cost of the tier serving the workflow, benchmarked against workflow accuracy, error rates and the cost of errors, not raw capability scores.
- **Routing economics:** Whether work reaches the cheapest capable tier, measured on observed rather than catalogue costs, and what the realized split is against the modelled one.
- **Human oversight economics:** Scored as its own dimension rather than netted into a blended cost line, because oversight and error cost are coupled rather than additive. Oversight converts an unreliable model's token from value-destroying to value-creating, and the optimal level is a derived economic quantity, not an engineering default. Netting it into a single number hides the lever most likely to be driving the result.
- **Volatility and risk economics:** Applied as a modifier after core scoring rather than embedded in threshold logic, because folding it into the base score obscures the CFO-level story.
- **Outcome economics:** The verified business result and the evidence supporting it.

Every stress test produces two artefacts at two reading depths: a CFO-readable resilience summary stating in plain terms whether a use case is economically sound and what would break it, and a technical layer for the CAIO, CIO and CTO showing scoring, assumptions and levers.

CFO explainability is the key design constraint that should govern every choice in the method.

Five questions belong to the board agenda, and the CEO should ask them before the board does:

- Which AI use cases can currently prove Cost Per Verified Outcome, and which are being scaled on faith?
- Who owns the number (budget, outcome and profit-and-loss line) for each material workflow?
- If AI spend doubled again next year in line with the market trajectory, could this organization still explain, line by line, what it bought?
- If AI were to do 30% of human work tomorrow, what value-based work would those people do?
- Which AI initiatives would we stop funding tomorrow if they had to be justified solely on verified outcomes?

# We have built a solution to track the AI Yield

The solution continuously measures the four relevant metrics continuously:

1. Cost Per Verified Outcome
2. The named owner of each material workflow
3. The attribution of every model call to a team and a feature
4. The value produced against the counterfactual of not using AI. This all occurs in near real time. The outcome is that the executive committee and the engineering team read the same numbers in the same week, rather than a quarter apart

Most CEOs cannot yet answer the third or fourth question with confidence, but that is a reason to build the measurement discipline now, while the cost base is still legible enough to instrument, not a reason to slow adoption.

[12] Epoch AI, Scaling constraints in frontier AI; METR, task-horizon measurements for frontier models

# Aligning your C-suite is of strategic importance

At the close of Nvidia's GPU Technology Conference, Jensen Huang offered listeners a thought experiment. Taking a software engineer as his example, he suggested that if by year-end that engineer's token consumption amounted to 1% of their salary, "there is reason to go mad." If they had not consumed at least half their salary in tokens, "there is reason to be deeply alarmed."[12]

This is a provocation, certainly. However, for any executive committee it poses a real strategic question:

- **How do you assure the productivity of a workforce that has become hybrid?** Run the experiment with your own team and each member will approach it through their own lens. The CFO's perspective is the one that ultimately releases or withholds the budget.

For leaders that work with the CEO, they must all face up to different challenges that ultimately align to the key question above:

- **The CFO carries the financial exposure.** Their problem is blind overspend: a cost line growing at 44% a year that nobody can attribute, forecast or defend. A study by Harness[13] puts numbers on it, 72% of organizations hit an unexpected AI bill in the past year, 52% have no clear owner of AI cost, and only 20% could explain a doubling within hours. No CFO signs off indefinitely on a line item with those properties. They will fund AI at scale once spend is attributable to an

[13] BlackRidge Research, "Finland Removes Data Center Power Tax Breaks" (May 2026); effective 1 July 2026

owner, forecastable within a tolerance, and defensible per outcome, and they will cap it, crudely, if it is not.

- **The CIO is at the center of the game.** They negotiate with the token supplier, directly or through a cloud provider, and are the guardian of the best purchase price. They can already see their budget being cannibalized, and they know that through successful engineering and disciplined FinOps, some of their peers have halved that budget at equivalent service.
- **The COO will insist on reworking processes before consuming any tokens at all**, imposing a one-for-one rule: every euro of token spend secured against an equivalent efficiency saving.
- **The CHRO will point out that seniority and grade remain the best predictors of expertise**, which is itself a good predictor of productivity, and will argue for decentralized allocation at team level.
- **Business leaders will read that prudence as a brake.** Their fear is being outrun by competitors. They are impatient to reinvent their products and will allocate budget first to the functions most exposed to competition.
- **The CEO will be concerned with identifying the domains of expertise that form the backbone of the firm** and drive its valuation, and with redefining the boundaries of the enterprise in the process.

Without a shared taxonomy, the committee is unlikely to reach agreement.

The key question is: where does AI best complement human expertise? Effective executive teams focus on

finding that answer. They encourage teams to experiment, learn from failures, improve, and test again, then come back together quickly to share what they have learned.

Three governance decisions follow and should be taken explicitly rather than by default.

1. **Allocation:** Central, federated to business units, or devolved to teams? Our view is that allocation should sit as close as possible to the person who can judge the marginal value of the next token. In most organizations, that is the team level.

2. **Token use caps:** Soft caps notify at 80% and let work continue, protecting productivity at the cost of open-ended exposure; hard caps block at the limit and require manager approval to reactivate, delivering cost certainty at the cost of friction. We recommend hard caps at team level with carry-forward for unused allowance, because during a period this volatile cost certainty is worth more than the friction.

3. **Recovery:** On billable engagements or transactions, AI usage must be coded and recharged, and where this is not designed in from the start it does not get retrofitted. It is important to note that recovery and attribution are where the CFO is won or lost. Blind overspend is not solved by a tighter budget, it is solved by making every euro attributable to a team, a workflow and an outcome before the spend happens. That is the difference between a CFO who caps AI and a CFO who funds it.

# Part III — What AI Yield means for the CAIO, CIO and CTO

For the leader accountable for delivering AI capability, AI Yield is only as credible as the measurement layer underneath it, and only as achievable as the architecture around it. Two decisions dominate everything else.

## The best strategy today is dual mode

Our recommendation is unambiguous: split workflows between frontier closed models and open-source self-hosted models, behind a routing layer that decides which is which on evidence rather than preference.

Neither strategy in isolation survives contact with a real workload mix. A frontier-only estate pays luxury prices for

commodity work. Most enterprise queries are routine extraction, classification, basic client interaction and report generation, none of which require frontier reasoning. An open-source-only estate fails on the minority of genuinely hard cases, and the failure is expensive precisely because those are the high-stakes ones.

The economics are decisive at scale. On the reference deployment in conviction five (UnTokenized Intelligence), a 50-50 routing split delivers 88–94% saving against cloud API rates on the locally routed half while preserving frontier capability for the half that needs it. AT&T's 90% run-cost reduction was achieved on exactly this pattern.

## Four design rules govern practical application

1. Route on classification and value, not on cost. An intelligent gateway classifies on intent, context, complexity and semantics. A FinOps model router that minimizes price is a different and inferior thing and will degrade quality in ways that do not appear in the cost report.

2. Gate admission to the local tier on measured accuracy. Build a golden dataset per workflow, sampled from production traces rather than written synthetically. Two to five hundred cases per workflow step is usually enough to separate a clear pass from a clear failure, and must include the tail of malformed inputs, unusual formats and long documents. Set the threshold per workflow, derived from the consequence of being wrong, not from a published benchmark. Record the frontier model's score on the same set: that is the real baseline.

3 Run candidates in shadow mode before switching. Send live traffic to both frontier and local candidate, serve only the frontier response, and compare. Shadow mode is the only honest test because it uses the real input distribution. Run it for at least one full business cycle. Set automated regression gates routing back to frontier on low confidence, schema-validation failure or an out-of-distribution signal. Then treat the fallback rate as a metric in its own right.

4 Instrument the realized split and re-run the economics when it moves. The routing split is an input to the financial model, not a property of the architecture. A split drifting from 50-50 to 30-70 changes the capital plan, and burst traffic beyond local capacity routes to frontier, which means the split (and the cost) moves under load.

Re-evaluate on every change: a new model version, quantization format, fine-tune or serving-engine upgrade all invalidate the previous result.

One caveat belongs here. Hosted open-weight endpoints serve the same weights as a local cluster at roughly 1.3 to 4 times less than self-hosting. Where there is no residency, audit or fine-tuning requirement, the correct answer is to own the routing policy rather than the hardware. Self-hosting is justified by sovereignty and control, and on unit cost only against frontier APIs. It should not be used against hosted open-weight endpoints serving the same models.

# Build your own portfolio of small language models (SLMs)

The second decision is what runs on the local tier. Our position is that the answer is not a shrunken generalist but a portfolio of domain-specific small models. This is where our value proposition sits.

## The market signal is clear

Small language model usage is projected to reach three times that of large models by 2027, growing at 279% year on year. Inference costs run roughly 500 times lower, training costs roughly 100 times lower, and total cost reduction reaches up to 75%. Regulated industries are projected to spend $66 billion by 2028 on this class of deployment, with roughly 80% of a sovereign shift expected by 2027 and the sovereign infrastructure market growing 4.5 times.

## The counterexample is instructive

Extreme-scale open-weight releases (for example, Kimi K3 at 2.8 trillion parameters with 104 billion activated per token), promise frontier capability with full sovereignty. But even with sparse activation the full parameter set must sit in high-

bandwidth memory to hold acceptable latency at production concurrency: roughly 5.6 TB at BF16, or a 1.56 TB checkpoint in the native MXFP4 release. Deployment guidance points to supernodes of 64 or more accelerators, putting silicon cost alone at $3–4 million. Owning a multi-trillion-parameter model locally is financially inefficient for most business tasks. The practical pattern is small models behind intelligent routing.

## What we have built

Capgemini Invent has fine-tuned and deployed a portfolio of vision and language models demonstrating the pattern at production scale. For example:

| Model | Domain | Base / trained parameters | Deployed footprint | Training |
|---|---|---|---|---|
| Solar-Smol-VLM | Drone-based solar panel inspection | 500M / 4.5M | 1.3 GB — RTX 5080 or Jetson Orin 8GB | 1,900 images, 14,203 records, ~15 hrs. |
| MedSmol-Derma | Skin disease detection | 500M / 9M | 1.3 GB — RTX 5080 or Jetson Orin 8GB | MMSkin, 6,578 images, ~12 hrs. |
| Legal-ToS-SLM v1/v2 | Terms of service compliance, 100+ clause classes | 1B / 406M | 1.54 GB — RTX 5080 | 13,898 records, ~4 hrs. |
| FF-Doc-3B | Freight forwarding document classification | 3B | 4–5 GB — A100/ H100 | Benchmarking in progress |

Two things are worth drawing out. The deployed footprint is between 1 and 5 GB, meaning these models run on a single consumer-class GPU or an edge device - not a cluster, and in several cases not even a data center. And the trained parameter counts are tiny relative to the base: 4.5 million trained parameters on a 500 million parameter base is a fine-tune measured in hours, not a training run measured in months.

## Does the quality hold?

On a drone image of a damaged solar panel, GPT-5 returns a competent generic description: cracks, burn marks, discoloration, reduced performance. However, Solar-Smol-VLM runs at three orders of magnitude smaller and can identify the module type, the crack pattern, the likely physical cause and the possible electrical fault. It is more useful because it was trained on the domain, i.e., specifically for this task, not because it is larger.

The dermatology comparison cuts the other way. On the same skin image, GPT-5, Gemma-3 and Qwen2.5 return plausible and mutually contradictory diagnoses. MedSmol-Derma returns a specific but incorrect diagnosis, together with an explicit statement that it is not qualified to give medical advice and that a doctor should confirm. The lesson here is about the workflow rather than the model. On a high-stakes clinical task none of these models clears the reliability threshold from Conviction 7, so the architectural answer is the accuracy gate and the oversight design.

## Our value proposition to help clients build the portfolio and the discipline around it:

- Qualifying workloads with a structured scorecard
- Building golden datasets and evaluation gates per workflow
- Fine-tuning domain-specific models on client data inside the client perimeter
- Engineering the serving stack that delivers the 5.8× capital efficiency multiplier
- Standing up the routing layer and the telemetry that closes the loop back to yield

Expect a minority of workloads to qualify on the first pass. Pilot on the two or three with the highest volume and the strongest residency requirement, at a single node, with the evaluation gate in place from day one. Decide on the second node from measured data.

# Conclusion

Most CEOs, CFOs, CIOs and CAIOs cannot yet answer, with evidence, whether their AI spend is buying anything measurable. That is not a failure of diligence. It is the predictable result of a cost structure that arrived faster than the instruments to govern it, in a market whose prices move monthly and whose value is uncorrelated with those prices.

The first wave of enterprise AI was defined by experimentation. The second is being defined by industrialization. The third will be defined by economics. Organizations that measure and optimize AI Yield will create more value from every workflow, every token and every employee. In the long run, AI advantage will not belong to those with the most intelligence. It will belong to those with the highest yield. The organizations that build this discipline now, while the cost base is still small enough to instrument and legible enough to reconstruct, will scale AI with confidence while their competitors are still guessing.

# References

Federal Reserve Bank of Atlanta, Policy Hub Macroblog, "How Much Are Firms Spending on AI (and What Will Happen to Headcounts)?", 6 May 2026 — Survey of Business Uncertainty, fielded March 2026, N = 855

Gartner, "Gartner Forecasts Worldwide AI Spending to Grow 47% in 2026," 19 May 2026; IDC, Worldwide AI and Generative AI Spending Guide

Capgemini Research Institute, "The multi-year AI advantage: Building the enterprise of tomorrow" (2026)

Stanford Institute for Human-Centered AI, AI Index Report 2026; Stanford Digital Economy Lab, "How Do AI Agents Spend Your Money?" (2026)

Artificial Analysis, State of AI Q1 2025 Highlights Report and 2026 intelligence-versus-cost index

FinOps Foundation, State of FinOps 2026 (1,192 practitioners, >$83bn annual cloud spend); Harness, 2026 State of AI in FinOps (conducted with 700 engineering leaders in five countries)

Exponential View, The State of the AI Economy, June 2026; BCG, AI Radar 2026 (2,360 executives, 16 markets, 640 CEOs)

Fortune, "After blowing through AI budget in a matter of months, Uber CTO says tokenmaxxing era is over" (August 2026); SemiAnalysis, "TokenBudgeting" (June 2026)

VentureBeat, AT&T agentic architecture and run-cost reduction (February 2026)

Colette Kress, CFO, NVIDIA, on inference share of data-centre revenue, cited in K. Bheemaiah, Capgemini Invent, "Inference Cost & Speed"

Epoch AI, Scaling constraints in frontier AI; METR, task-horizon measurements for frontier models

BlackRidge Research, "Finland Removes Data Center Power Tax Breaks" (May 2026); effective 1 July 2026

P. Cordier, "The Value of a Token," Capgemini Invent, https://hal.science/hal-05680441v1 (2026)

U. Gupta and S. Bhatnagar, "UnTokenized Intelligence," Capgemini Invent (August 2026)

Capgemini Invent, "TokenOps Governance and Delivery Briefing" and "Token Economics" offer material (internal, 2026)

E. Grass, Les Echos, columns on AI scaling laws, the end of the free lunch, and AI budget allocation (2026)

Bai et al. 2026 (arXiv:2604.22750); Chen et al. 2024 (FrugalGPT, TMLR); Liu et al. 2025 (arXiv:2511.17006); Semantic Router for vLLM (arXiv:2510.08731); Regmi et al. 2024 (arXiv:2411.05276); "Beyond Human-Readable" (arXiv:2604.07502); Trisciuoglio 2025; Snell et al. 2024

# Authors

Etienne Grass

Global Chief AI Officer
Capgemini Invent

Jez Back

Cloud & AI Economist and Global Offer Leader
Capgemini Invent

Philippe Cordier

Global Chief AI Scientist
Capgemini Invent

Ujjwal Gupta

Senior Director, Chief AI Enterprise Architect
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