CONCIERTO | SOVEREIGN INTELLIGENCE

Sovereign Perspectives · Economics of Owned Intelligence

Metering Work, Not Contexts and Tokens

Pricing Intelligence by Output, Not by Effort

Series Sovereign Perspectives
Category The Economics of Owned Intelligence

Format Position Paper
Published by Concierto



Version 1.0 · September 2026

Copyright © 2026 Concierto. All rights reserved.

Contents

Executive Summary

1. The Unit Is the Argument
2. What the Meter Charges, and What It Ignores
3. Why a Market Pays for Output, Not Effort
4. The Double Whammy: From the Meter to the Loss of Your Edge
5. Metering the Work
6. What Leaders Must Do
7. The Reframe

About This Series · Sources

Executive Summary

Before an enterprise argues about the price of AI, it should look at what it is being charged for, because the unit of the bill decides everything that follows.

Enterprises are billed for intelligence in units their suppliers chose: tokens, contexts, iterations. These units measure the model's effort, not the buyer's outcome, and the gap between the two is the whole of this argument. In every functioning part of a market economy, cost is tied to output, the result delivered, and not to the effort expended reaching it. A firm pays its people for work done, its suppliers for goods shipped, its partners for results produced. Token pricing is the rare and costly exception, because the model, not the buyer, decides how much effort each task consumes, and the buyer pays for that effort whether or not it produced anything of value. Worse, to be billed this way the enterprise must feed the model the very data and knowledge that are its competitive edge, so it pays twice: once in the meter, and again in the slow commoditization of its own business. This paper argues that the unit of AI cost should be the work an enterprise does, and that only an owner can meter it that way.

In Brief

- Enterprises are billed in the supplier's units, tokens and contexts, which measure the model's effort rather than the work the business gets done.
- The meter is blind to whether the answer was correct, whether it served the business, or whether any value was created; it counts effort and nothing else.
- A market economy prices output, not effort, because paying for effort rewards inefficiency and severs price from value; token pricing quietly reintroduces exactly that.
- The deeper cost is a double loss: to use the metered model the enterprise feeds it the data and IP that are its edge, commoditizing the very business the AI was meant to improve.
- Metering the work, authors, workloads, applications, and agentic events, restores a unit finance can forecast, ties cost to value, and keeps the enterprise's knowledge its own.

When every part of a business is rewarded for output, there is no reason AI alone should be paid for effort, and less reason to pay in the currency of your own IP.

Copyright © 2026 Concierto. All rights reserved.

1. The Unit Is the Argument

The moment an institution cannot perform its core function without AI, that AI has become infrastructure, whether or not anyone decided to treat it as such.

When an enterprise adopts AI, it inherits a vocabulary it did not choose. The bill arrives denominated in tokens, contexts, and iterations, terms the providers coined to describe the internal mechanics of a language model. To the supplier these units are natural, because they map directly onto the computation the model performs. To the buyer they are opaque, because they map onto nothing the business recognizes as value. No board sets a target for tokens consumed. No customer pays more because a supplier's model needed more context. The unit belongs to one side of the table and is imposed on the other.

That imposition matters far beyond the invoice, because the unit an organization is charged for becomes the thing it manages, measures, and optimizes. A company billed by the token learns to count tokens, to ration them, to fear them, and to reshape how its people work around a number that has no relationship to the value they produce. The unit quietly becomes the metric, the metric becomes the constraint, and the constraint reshapes the business. This is why the argument about AI pricing is not really about price at all. It is about which quantity a company will organize itself around, and whether that quantity is one the business or the supplier defines.

A token measures how hard the model worked. It says nothing about whether the enterprise's work was done. The buyer has been made to count the wrong thing.

2. What the Meter Charges, and What It Ignores

A consumption meter is precise about the one thing that does not matter, and silent about everything that does.

Consider what the meter actually records. It charges for every token the model generates, every time the context has to be supplied, and every iteration the task requires. What it does not record is whether any of that produced a correct answer, whether the answer served the business's purpose, whether a smaller and cheaper model could have done the same work, or whether the user was even on a task the company cared about. The meter is exhaustively precise about the model's effort and completely blind to the outcome. It cannot tell a brilliant answer from a wrong one, or essential work from an idle tangent, because it was never designed to; it was designed to count computation, and computation is what the supplier sells.

Copyright © 2026 Concierto. All rights reserved.

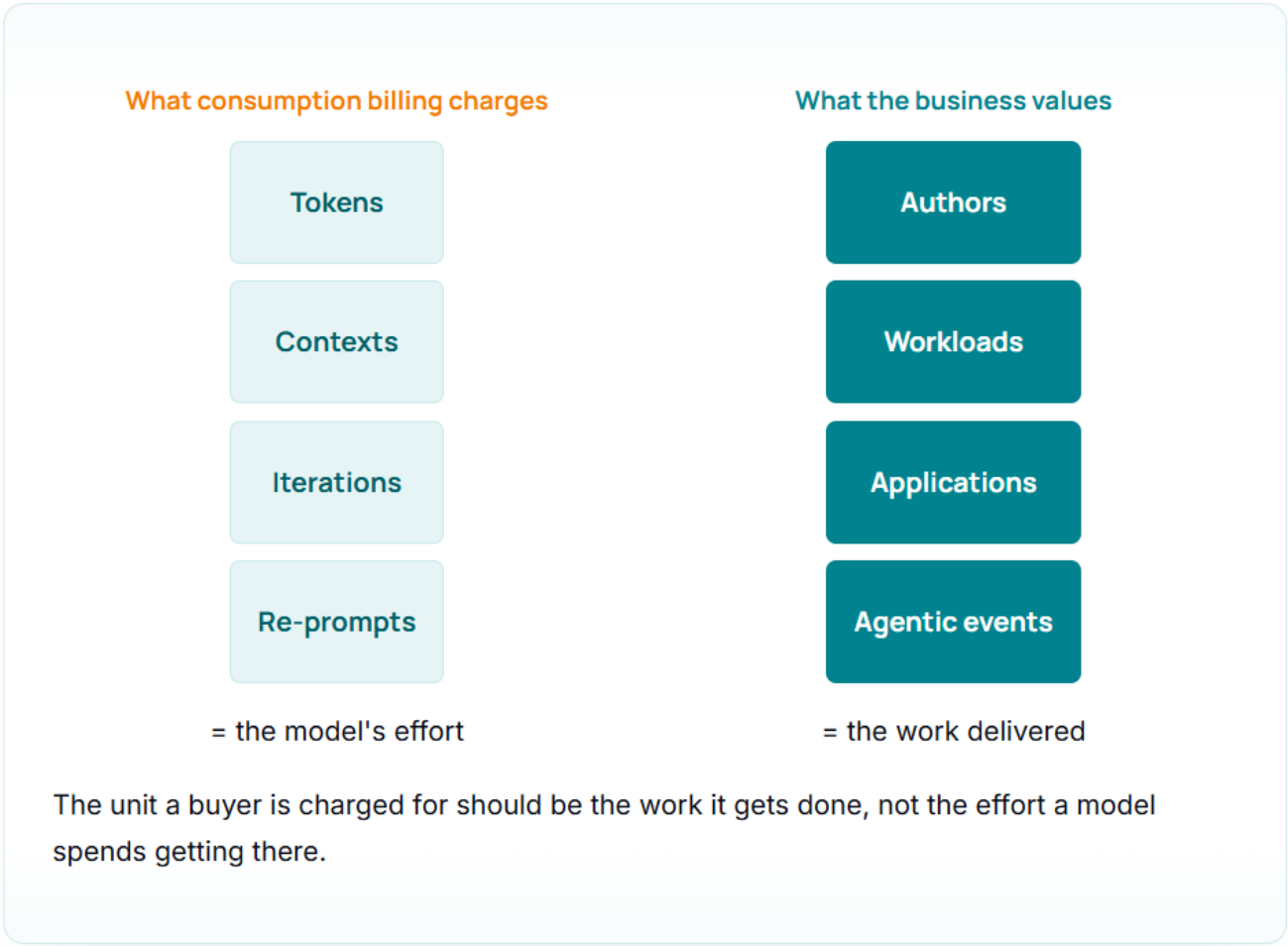


Figure 1. What consumption billing charges, and what the business actually values.

The deeper problem is who controls the quantity being metered. The model decides the quality of each answer, and therefore how many attempts a task takes; it decides how verbose to be, and therefore how many tokens it spends; it decides how much context it needs re-supplied. Every one of those decisions is billed to the buyer, and the buyer controls none of them. When a model returns a poor first answer and the user must re-prompt three times, the customer pays for all four exchanges. When a newer model spends ten times the tokens to reach the same result, the bill rises while the value stays flat. The enterprise is not paying for outcomes it defined; it is paying for the model's effort, including the model's inefficiency, on a meter held by the party that profits when that effort grows.

The supplier earns more precisely when its model is less efficient. Efficiency is not in the interest of the party that owns the meter, which is why the meter will never optimize for it.

Copyright © 2026 Concierto. All rights reserved.

3. Why a Market Pays for Output, Not Effort

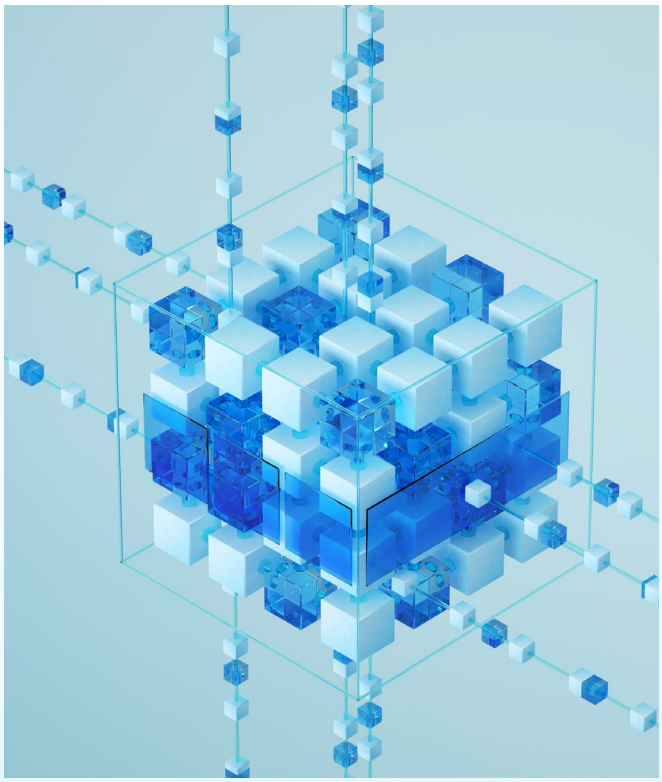
Paying for output rather than effort is not a preference. It is the mechanism that makes a market economy work at all.

Step back from AI and look at how a functioning economy prices almost everything, and the strangeness of token billing comes into focus. A manufacturer is paid for the goods it delivers, not for the hours its machines ran. A builder works to a fixed contract, a software vendor is paid per seat, a farmer per bushel, all paid for what they produce rather than for how hard they had to try. Even where labor is paid by the hour, competition drives the price toward the value the hour produces, and the worker who delivers more per hour earns more. Price, in a market, is tethered to value delivered, and that tether is the single most important discipline the market imposes.

The reason is not sentiment; it is incentive. When a supplier is paid for output, it profits by becoming more efficient, because efficiency lets it deliver the same value at lower cost and win against rivals who cannot. Effort-based pricing severs that tether and inverts the incentive. A supplier paid for effort profits by expending more effort, and has no reason to become more efficient, because efficiency would reduce the very thing it bills for. This is why economies abandoned effort-based pricing wherever they could: it rewards waste, hides value, and removes the pressure that drives improvement. Cost-plus arrangements are notorious for exactly this failure, and survive only where no better measure exists.

Token pricing takes that discredited model and installs it at the center of the fastest-growing cost line in the modern enterprise, at the precise moment that line is scaling. It pays a supplier for effort, on a technology whose effort the supplier controls, and presents it as the natural way to buy intelligence. An enterprise that accepts it has agreed, perhaps without noticing, to purchase its most strategic capability on the one basis a market economy learned long ago to avoid.

> **The rest of a business is paid for what it accomplishes. Token pricing is the one arrangement that pays the supplier for how hard its machine had to try, and rewards it for trying harder.**



Copyright © 2026 Concierto. All rights reserved.

4. The Double Whammy: From the Meter to the Loss of Your Edge

The meter is only the visible cost. The larger one is paid in the currency of the enterprise's own knowledge.

There is a second cost to the rented, metered model, and it is more damaging than the first because it is deferred and invisible. To get useful work from a general model, the enterprise must feed it the raw material of that work: its documents, its data, its processes, the accumulated knowledge that distinguishes it from its competitors. That knowledge is the enterprise's edge. Handed to a general model that improves from what it processes, the edge does not stay the enterprise's own. It is absorbed, generalized, and, in effect, commoditized, made available in diluted form to anyone who uses the same model, including the competitors the enterprise was trying to beat.

So the enterprise pays twice for the same transaction. It pays the meter now, in cash, for the model's effort. And it pays later, in eroded advantage, as the knowledge it supplied to get that work done makes the general model a little better at its business for everyone else. The immediate benefit is instant gratification: fast answers, no infrastructure to build, value today. The deferred cost is the slow commoditization of the very business the AI was supposed to improve. It is difficult to imagine a worse bargain for a company whose value lies in what it knows: to rent its intelligence by the token, and to pay for the privilege by giving away the knowledge that made it worth renting.

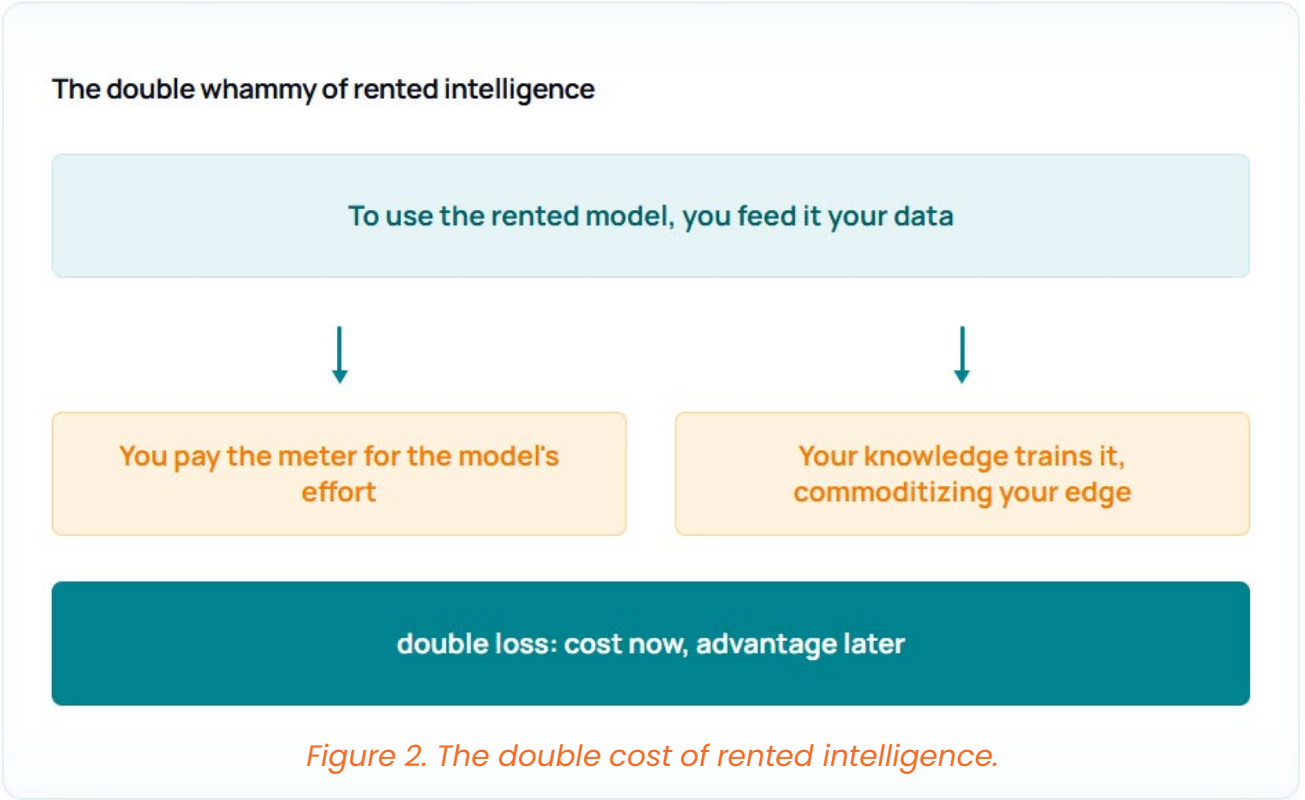


Figure 2. The double cost of rented intelligence.

Copyright © 2026 Concierto. All rights reserved.

Paying for instant gratification with the currency of your own IP is a double loss: cost today, and a commoditized business tomorrow.

5. Metering the Work

There is a trap even in metering the right unit, and it is worth naming: the instinct to answer the token meter with the word unlimited. Unlimited is not a price; it is an entitlement, and an entitlement is close to irreversible. Promise infinite consumption at a fixed price and internal budgeting stops attaching any cost to use, the provider margin becomes hostage to the customer behaviour, and the day the terms must change the customer experiences it as a betrayal. The discipline is to sell deterministic capacity, not infinity: a fixed, predictable price for a defined, generous envelope, expandable on known terms. Determinism is the promise; abundance is not.

Priced this way the unit becomes legible. Human users are metered per seat at a flat annual rate; agents, which exercise the model far harder, are metered per agent at their own rate; the workload, not the token, is the thing paid for. The customer can forecast the bill, the provider keeps a healthy margin because it carries no per-token cost, and neither side is tied to a meter that rewards inefficiency.

The alternative is to bill for the work an enterprise produces, not the effort a model spends, and to keep the knowledge that produces it at home.

A better unit already exists, and it is the one the business itself recognizes as value. Meter the authors who build with AI, the workloads they run, the applications they stand up, and the agentic events those applications execute. Each is countable, forecastable, and explicable to a board, and each is tied to output rather than to the verbosity of a model. A cost expressed in these units tracks the value the enterprise creates, so a productive quarter registers as work delivered rather than as a bill that ran away. And because the meter no longer rewards consumption, the model's efficiency becomes the enterprise's gain instead of the supplier's loss; when the model gets better and cheaper, the saving flows to the owner.

Metering the work is only fully possible when the enterprise owns the intelligence, for two reasons. The effort-based meter belongs to the supplier and will not be surrendered by a business model built on it. And metering by value requires grounding the model in the enterprise's own knowledge without surrendering that knowledge, which an owned model does and a rented one cannot. Ownership closes both losses at once: it replaces the effort meter with a value meter, and it keeps the enterprise's data and IP inside the enterprise, so the model learns the business without commoditizing it.

Meter the work, and cost tracks value while knowledge stays home. Meter the tokens, and cost tracks the model's inefficiency while your edge leaks away.

Specific tier definitions and pricing are set during evaluation and confirmed by the finance office; no figures are published here.

Copyright © 2026 Concierto. All rights reserved.

6. What Leaders Must Do

- **Refuse the supplier's unit.** Insist that AI cost be tied to the work the business does, not to tokens and contexts the buyer cannot control and did not define.
- **Meter authors, workloads, applications, and agentic events.** Adopt units finance can forecast and that track value created rather than effort expended.
- **Protect the knowledge that is your edge.** Ensure the data and IP that ground the model stay inside the enterprise, so using AI improves the business without commoditizing it.
- **Make efficiency your gain.** Own the intelligence, so that when the model becomes more efficient the saving accrues to the enterprise rather than to the supplier.
- **Budget on the work you intend to do.** Plan against a unit tied to output, so a productive period reads as progress rather than as an overrun.

The buyer that lets the supplier choose the unit has already lost the negotiation. Choose the unit, own the meter, and keep the knowledge.

7. The Reframe

The pricing of AI looks like a commercial detail and is in fact the whole strategy, because the unit decides who profits from efficiency, whether a budget can be forecast, and who keeps the knowledge.

Priced by the token, AI rewards the model's effort and the supplier's inefficiency, on a meter the buyer does not hold, and it is paid for in part with the enterprise's own commoditized knowledge. Priced by the work, and owned, AI ties cost to value, makes efficiency the enterprise's gain, gives finance a number it can plan against, and keeps the data and IP that are the business's edge inside the business. The question is not how much a token should cost. It is why a company building its future on what it knows would agree to pay for intelligence in tokens at all, when doing so quietly sells the knowledge underneath.

Stop paying for the words a model consumes. Start paying for the work it does, own the meter that counts it, and keep the knowledge that makes the work worth doing.

Copyright © 2026 Concierto. All rights reserved.

About This Series



Sovereign Perspectives is a research program on the ownership and governance of intelligence, published by Concierto. Its papers examine the strategic, economic, and institutional questions that artificial intelligence raises for governments and other accountable institutions, and they are written for the leaders making those decisions. The program's aim is to advance the public argument, not to promote a product.

Sources

Structural analysis of AI pricing models. Illustrative figures on model efficiency draw on Stanford Institute for Human-Centered AI (HAI), 2025 AI Index Report. hai.stanford.edu.

CONCIERTO | SOVEREIGN INTELLIGENCE

reach@concierto.cloud

www.concierto.cloud

www.trianz.com

Copyright © 2026 Concierto. All rights reserved.