CONCIERTO | SOVEREIGN INTELLIGENCE

Sovereign Perspectives · Purposeful Intelligence

The Importance of Being Correct

Why Rented AI Is Paid to Be Inefficient, and Owned AI Is Not

Series Sovereign Perspectives
Category The Economics of Owned Intelligence

Format Analytical
Published by Concierto



Version 1.0 · September 2026

Copyright © 2026 Concierto. All rights reserved.

Contents

Executive Summary

1. The Renter's Incentive
2. Correctness Is the Point
3. The Price of Getting It Wrong
4. Why the Cost of Inaccuracy Compounds
5. The Owner's Incentive
6. What Leaders Must Do
7. The Reframe

About This Series · Sources

Executive Summary

Every enterprise wants correctness from its AI. The dominant way of buying it is structured to reward the opposite.

Correctness means the right answer, the first time, on the task at hand, at the lowest cost. It is what an enterprise is actually trying to buy when it adopts AI. The trouble is that the prevailing pricing model is paid for the reverse. When intelligence is rented and billed by consumption, the supplier earns more the more the model is used, so inefficiency, extra iterations, verbosity, and time spent off the task are not defects in the arrangement; they are the arrangement's source of revenue. The unit cost of a fixed quality has fallen more than 280-fold in eighteen months, and yet enterprise bills keep climbing, because consumption is what pays, and consumption rewards everything correctness is not. This paper starts from that misalignment, defines correctness and why it matters, shows how the cost of getting it wrong compounds as agents multiply, and argues that correctness is cheap only when the party paying for it also owns it.

In Brief

- Consumption pricing pays the supplier for more tokens, more iterations, and more time on the tool, whether or not any of it serves the business, so efficiency runs against the supplier's revenue.
- Correctness, the right answer the first time, on task, cheaply, is what the enterprise is buying and almost nothing a consumption meter rewards.
- Getting it wrong is expensive: brute-force accuracy costs several times more, and every re-prompt and tangent is billed in full.
- The cost of inaccuracy is compounding, because growth in AI spend comes largely from existing customers using more, and agentic systems turn one task into dozens of billed model calls.
- Ownership flips the incentive: the owner profits from efficiency, first-time-right answers, and on-task use, so correctness becomes cheap rather than costly.

Correctness costs the enterprise money and costs the renter revenue. That is why the rented model is not built to deliver it.

Copyright © 2026 Concierto. All rights reserved.

1. The Renter's Incentive

Follow the money before reading the marketing, and the conflict at the center of rented AI is impossible to miss.

Under consumption pricing, the supplier's revenue is a direct function of use: more tokens generated, more iterations run, more context re-supplied, more time spent with the model. Everything the enterprise would call waste, the supplier's meter counts as income. A concise, correct first answer is the cheapest possible outcome for the buyer and the least profitable for the seller. A verbose answer that requires two follow-ups is more expensive for the buyer and more profitable for the seller. The two parties are not aligned around a shared goal with a price attached; they are pulled in opposite directions by the very structure of the bill.

Nothing in the arrangement rewards the things an enterprise most wants. There is no premium for getting the answer right the first time, because a right first answer ends the billable exchange. There is no reward for keeping a user on the business's task rather than an interesting tangent, because the tangent is billable too. There is no incentive to use a smaller, cheaper model where one would do, because the larger model bills more. And there is no reason to reduce the infrastructure a task consumes, because consumption is the product. The party that controls the model, its quality, its verbosity, the number of tries it takes, is paid more precisely when it controls those things poorly.

Backwards

The one commercial arrangement in which the seller profits from doing the job less efficiently is the one most enterprises have adopted by default.

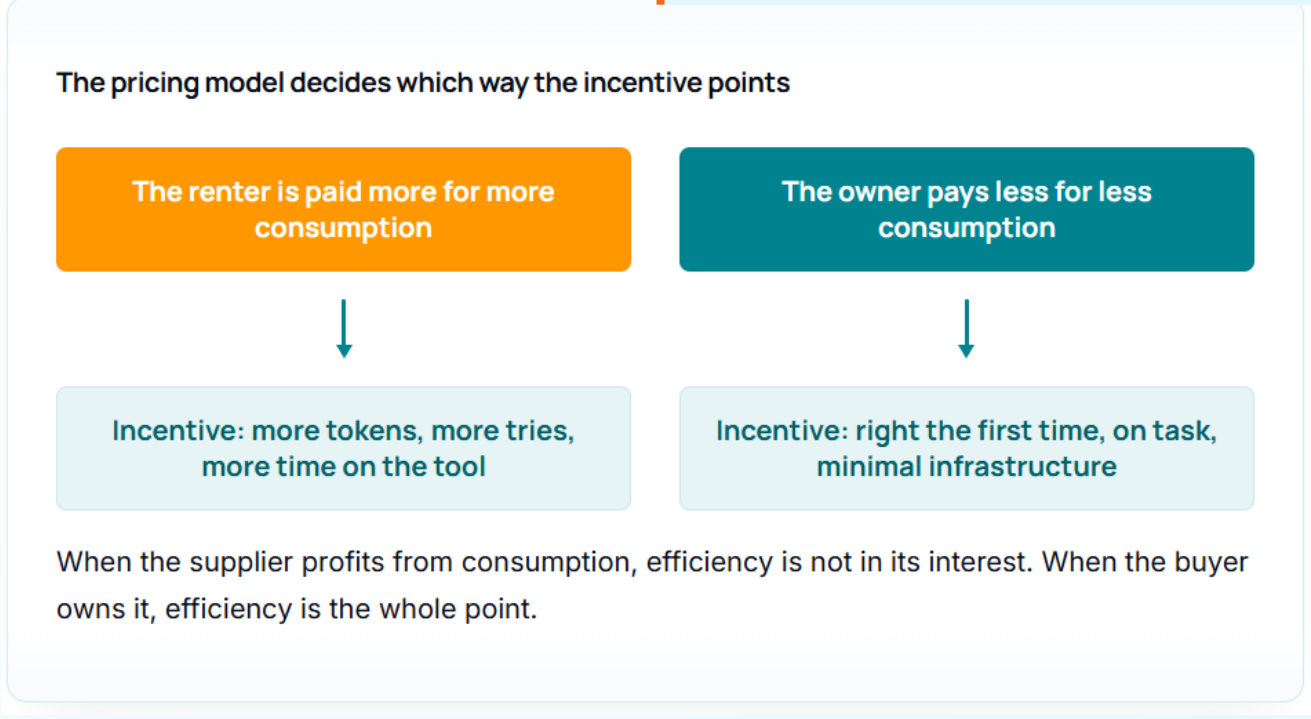


Figure 1. Parameters needed to match a leading 2022 model. Source: Stanford HAI, 2025 AI Index Report.

Copyright © 2026 Concierto. All rights reserved.

2. Correctness Is the Point

Set the pricing aside for a moment and be precise about what an enterprise is actually trying to buy, because the word carries more than it first appears.

Correctness is not a single property; it is four at once. It is accuracy, the answer is right. It is directness, the answer is reached on the first attempt rather than the third. It is relevance, the work stays on the business's task rather than wandering. And it is efficiency, all of the above achieved with the least computation, latency, and cost. An answer that is accurate but takes five tries is not fully correct in the sense that matters to a business, and neither is one that is fast but off-task, or right but ruinously expensive. Correctness is the whole of getting the right thing done, well, the first time, for as little as possible.

Why it matters is that correctness compounds, and its absence compounds faster. A correct first answer is cheaper, because it ends the exchange; it is faster, because there is no rework; it is more trusted, because users stop double-checking; and trust is what turns a promising tool into an adopted one. Incorrectness runs the same logic in reverse: a wrong or off-target answer triggers re-prompts, which cost money and time; it erodes trust, which slows adoption; and it forces human verification, which erases the productivity the AI was meant to deliver. The benefit of correctness, then, is not merely a good answer. It is a lower total cost of getting work done, a faster path to value, and an organization that actually comes to rely on its AI rather than warily supervising it.

> **Correctness is a right answer, first time, on task, cheaply. Almost nothing about that goal is what a consumption meter is paid to deliver.**



3. The Price of Getting It Wrong

he cost of the misalignment is not abstract. It is measurable, and it is charged on every task, including the ones that never needed it.

There is a way to buy more accuracy from a given model, which is to let it think longer, and it is instructive precisely because it is expensive. On one hard task, more thinking lifted a score from 9 percent to 74 percent, but at roughly six times the cost and thirty times the latency. That premium is real accuracy, but it is the dearest accuracy on the menu, and under a consumption model it is charged on every task, including the great majority that a smaller, faster model would have answered correctly for a fraction of the price. The enterprise pays brute-force prices for routine work because the meter does not distinguish, and cannot, between the task that needed the deep model and the one that did not.

Copyright © 2026 Concierto. All rights reserved.



Figure 2. Accuracy bought by brute force, and its price. Source: Stanford HAI, 2025 AI Index Report.

Add to that the cost of every wrong first answer, every re-prompt, every clarification, every tangent, and the true price of incorrectness comes into view. None of that spend buys a better answer to the business's question; it buys more of the model's effort on the way to one. And because the meter profits from exactly this, there is no counterparty working to reduce it. The buyer absorbs the full cost of the model's distance from correctness, again and again, on a bill it cannot see coming.

~6x

The dearest accuracy on the menu is charged on every task, whether the task needed it or not, and nothing in the meter works to reduce it.



4. Why the Cost of Inaccuracy Compounds

Agents are where this compounds hardest. A single interaction pays the re-explanation tax once; an agentic workflow pays it on every step, because each call to an inefficient model re-establishes the context the last one failed to keep. Multiply that across a fleet of agents planning, calling tools, and checking one another and the cost of an imprecise model does not add, it multiplies. It is also why agents diverge most in price between the rented and the owned worlds: an agent runs heavy and around the clock, so on a metered model it costs close to a heavy engineer while on an owned platform it is a fraction of that. The least accurate model is most expensive exactly where the workload is heading.

What worsens is not the price of a token. It is the number of times an inefficient model has to try to understand your business before it responds properly, and that number is exploding.

A general model comes to every task knowing nothing of the business it serves. It does not know the company's products, its customers, its processes, or the way it makes decisions, so it guesses, and its first guess is often generic or wrong. The user corrects it, supplies more context, and asks again; the model tries once more; and each of those attempts is a billed round of the model learning, at the customer's expense, a business it was never taught. A model grounded in the business reaches a right answer in one pass because it already holds the context. A general model reaches it in three or five, because it is re-learning the context every time, and it retains none of it for the next request. The cost of inaccuracy is not one wrong answer. It is the tax of re-explanation, levied on every task, paid to a model that never remembers what it was told.

The growth in AI spending tells a specific story. Adoption has surged, from 55 percent of organizations to 78 percent in a single year, but the larger driver of spend is no longer new adopters signing up; it is existing customers using far more. That expansion is the point most buyers miss. Under consumption pricing, a customer's own success is what raises its bill, because success means more use, and more use, on an inefficient model, means more of the inaccuracy premium paid over and over. The enterprise is not billed once for the model's inefficiency; it is billed for it on every call, and it makes more calls every quarter.

Agentic systems turn that steady rise into a compounding one. A single human request handled by one model call is one chance to be inefficient. The same request handled by an agent that plans, retrieves, calls tools, checks its work, and hands off to other agents becomes dozens of model calls, each carrying the same inefficiency premium, each billed. As enterprises move from chat to agents, and from one agent to swarms of them, the number of billed calls per unit of real work rises by an order of magnitude, and the cost of the model being inefficient rises with it. A small inaccuracy tax on a single call is tolerable. The same tax levied on every step of an agentic workflow, running continuously across the enterprise, is not.

Every step

An inefficiency tax on one call is a nuisance. Levied on every step of an agent swarm, running continuously, it becomes the largest line in the AI budget.

Copyright © 2026 Concierto. All rights reserved.

5. The Owner's Incentive

Ownership does not soften the misalignment. It reverses it, and puts the party paying for correctness on the same side as the goal.

When an enterprise owns its intelligence, every efficiency becomes its own gain rather than a supplier's loss. A correct first answer saves the enterprise money; an on-task session saves it money; a smaller model that fits the task saves it money; and lower infrastructure use saves it money. For the first time the party that pays for correctness is the party that benefits from it, so the whole weight of the organization's self-interest points toward getting more right for less, which is exactly what correctness is. The incentive that ran backwards under rental runs forward under ownership.

Ownership also lets correctness compound in the enterprise's favor. A model the enterprise owns and grounds in its own data grows more accurate on the enterprise's tasks the more it is used, so accuracy rises while cost falls, the mirror image of the rented curve where consumption rises with success. Over time the owned model gets cheaper and more correct together, and the compounding that worked against the renter now works for the owner. The agentic future that made rented inaccuracy so dangerous becomes, under ownership, the very thing that makes an efficient owned model pay off, because the efficiency is captured on every one of those multiplying calls.

Aligned

Own the intelligence, and for the first time the party paying for correctness is the party that gains from it.

Rented and metered	Owned and aligned
Supplier profits from more consumption	Owner profits from less consumption
No reward for a correct first answer	Every efficiency is the owner's gain
Indifferent to off-task use	Rewarded for keeping work on task
Inefficiency compounds against the buyer as agents multiply calls	Efficiency compounds for the owner on every call
Accuracy fixed by the supplier's model	Accuracy rises as the owned model learns the domain

Copyright © 2026 Concierto. All rights reserved.

6. What Leaders Must Do

- **Read the incentive, not just the price.** Ask which party profits when the model is inefficient, and recognize that a rented meter is structured to reward consumption, not correctness.
- **Right-size the model to the task.** Reserve expensive reasoning for the tasks that need it and route the rest to cheaper, specialized capability, so brute-force prices are not charged on routine work.
- **Plan for the agentic multiplier.** Assume the number of model calls per unit of work will rise by an order of magnitude, and choose a cost model that does not compound an inefficiency premium across every step.
- **Own the correctness that compounds.** Ground a specialized model in your own data so accuracy rises and cost falls, and the gains stay with you rather than the supplier.

Own it

Correctness is expensive to rent because the renter is not paid to deliver it, and cheap to own because the owner is.

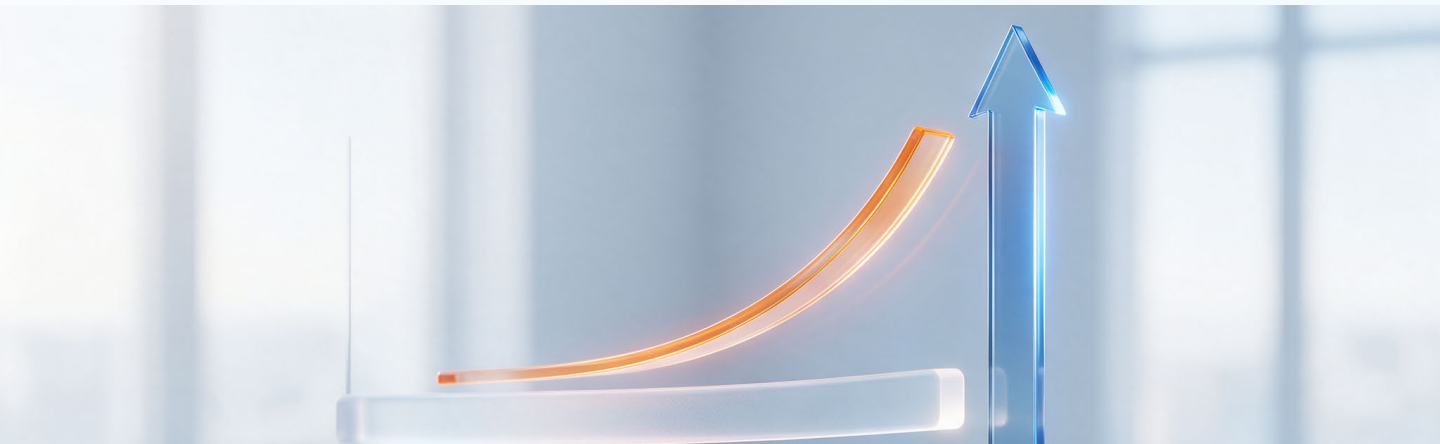
7. The Reframe

The market frames its cost as the price of capability, which sounds like a fair exchange. It is not, because the meter is paid for the opposite of what the buyer wants.

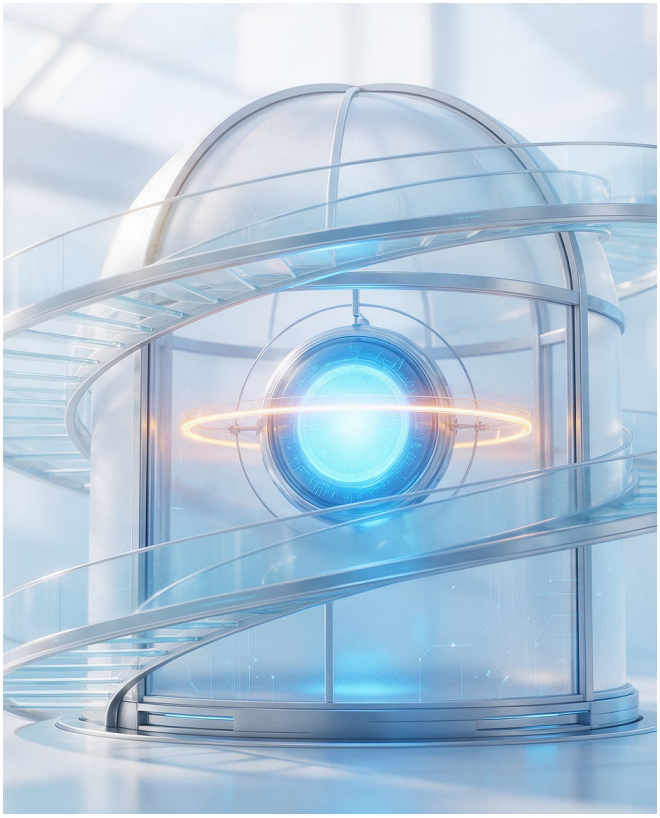
Correctness, the right answer the first time, on task, at the lowest cost, runs directly against a business model paid by consumption, and no negotiation over the token price changes that structural fact. As agents multiply the calls and existing customers expand their use, the cost of the model's inefficiency does not hold steady; it compounds, in the supplier's favor and against the buyer's. The only way to put correctness and cost on the same side of the ledger is to own the intelligence, so that every efficiency is the enterprise's gain rather than the supplier's loss, and every one of the multiplying calls is a saving rather than a charge.

Structural

You cannot negotiate your way to correctness with a supplier paid to be inefficient. You can only own your way there, before the agents multiply the bill.



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

Stanford Institute for Human-Centered AI (HAI), 2025 AI Index Report. Inference cost, reasoning-cost, and business-adoption data. hai.stanford.edu. Agentic call-volume growth described from the observed structure of agentic workflows.

reach@concierto.cloud

www.concierto.cloud

www.trianz.com

Copyright © 2026 Concierto. All rights reserved.